

BIG DATA

Técnicas, herramientas y aplicaciones



María Pérez Marqués

 **Alfaomega**

libros


BIG DATA

Técnicas, herramientas y aplicaciones



María Pérez Marqués

 **Alfaomega**

 **libros R**

BIG DATA

Técnicas, herramientas y aplicaciones

Actualmente nos encontramos en la era de los grandes conjuntos de datos, procedentes de múltiples y variados orígenes, en formatos diversos y con una necesidad de procesamiento y análisis rápido y efectivo. Las técnicas de Big Data persiguen complementar el manejo de estos grandes volúmenes de datos con las técnicas de análisis de la información más avanzadas y efectivas para extraer de modo óptimo el conocimiento contenido en los datos.

Las herramientas de Big Data se basan en el paquete de código abierto llamado Hadoop para el análisis masivo de datos, que forma parte de prácticamente todo el software de Big Data. Por ejemplo, SAS incorpora Hadoop en sus aplicaciones (SAS Base, SAS Data Integration, SAS Visual Analytics, SAS Visual Statistics, etc.). IBM trabaja con Hadoop en su plataforma IBM InfoSphere BigInsights. Microsoft incluye Hadoop en su plataforma Windows Azure, SQL Server 2014, HDInsight y Polybase. Oracle incluye Hadoop en Oracle Big Data Appliance, Oracle Big Data Connectors y Oracle Loader for Hadoop.

Este libro presenta las posibilidades de trabajo que ofrecen las herramientas de Big Data para procesar y analizar grandes volúmenes de datos de una manera ordenada. Se describen a lo largo de los capítulos del libro las herramientas de Big Data que implementan SAS, IBM, Microsoft y Oracle, analizando a su vez, sus posibilidades para extraer el conocimiento contenido en los datos.

María Pérez Marqués es especialista en cálculo científico en el grado de Ingeniería Aeronáutica de la Universidad Politécnica de Madrid. Adicionalmente desempeña una parte importante de sus investigaciones en los campos de las bases de datos y los sistemas operativos. Colabora especialmente en las tareas de diseño en I+D+i (Investigación, desarrollo e Innovación). Asimismo participa en los programas de adaptación de los métodos computacionales a las técnicas de enseñanza. Es autora de varias publicaciones en los ámbitos científico y docente.

www.alfaomega.com.mx

ÁREA

Computación

SUBÁREA

Bases de Datos

ISBN 978-607-622-450-2



9 786076 224502

"Te acerca al conocimiento"



Alfaomega Grupo Editor

Big Data

Técnicas, herramientas y aplicaciones

Big Data

Técnicas, herramientas y aplicaciones

María Pérez Marqués

Diseño de la colección y pre-impresión:
Grupo RC

Diseño de la cubierta:
Cuadratín

Datos catalográficos

Pérez, María

Big Data. Técnicas, herramientas y aplicaciones

Primera Edición

Alfaomega Grupo Editor, S.A. de C. V., México

ISBN: 978-607-622-450-2 eISBN: 978-607-622-445-8

Formato: 17 x 23 cm

Páginas: 356

Big Data. Técnicas, herramientas y aplicaciones

María Pérez Marqués

ISBN: 978-84-943055-5-9 edición original publicada por RC Libros,
Madrid, España Derechos reservados © 2015 RC Libros

Primera edición: Alfaomega Grupo Editor, México, junio 2015

© 2015 Alfaomega Grupo Editor, S.A. de C.V.

Pitágoras 1139, Col. Del Valle, 03100, México D.F.

Miembro de la Cámara Nacional de la Industria Editorial Mexicana
Registro No. 2317

Pág. Web: <http://www.alfaomega.com.mx>

E-mail: atencionaldiente@alfaomega.com.mx

ISBN: 978-607-622-450-2 eISBN: 978-607-622-445-8

La transformación a libro electrónico del presente título fue realizada
por Sextil Online, S.A. de C.V./ Editorial Ink ® 2016.

+52 (55) 52 54 38 52

contacto@editorial-ink.com

www.editorial-ink.com

Derechos reservados:

Esta obra es propiedad intelectual de su autor y los derechos de publicación en lengua española han sido legalmente transferidos al editor. Prohibida su reproducción parcial o total por cualquier medio sin permiso por escrito del propietario de los derechos del copyright.

Nota importante:

La información contenida en esta obra tiene un fin exclusivamente didáctico y, por lo tanto, no está previsto su aprovechamiento a nivel profesional o industrial. Las indicaciones técnicas y programas incluidos, han sido elaborados con gran cuidado por el autor y reproducidos bajo estrictas normas de control. ALFAOMEGA GRUPO EDITOR, S.A. de C.V. no será jurídicamente responsable por: errores u omisiones; daños y perjuicios que se pudieran atribuir al uso de la información comprendida en este libro, ni por la utilización indebida que pudiera dársele.

Edición autorizada para venta en México y todo el continente americano.

Impreso en México. Printed in México.

Empresas del grupo:

México: Alfaomega Gmpo Editor, S.A. de C.V. - Pitágoras 1139, Col. Del Valle, México, D.F. -C.E 03100. Tel.: (52-55) 5575-5022 - Fax: (52-55) 5575-2420 / 2490. Sin costo: 01-800-020-4396 E-mail: atencionalcliente@alfaomega.com.mx

Colombia: Alfaomega Colombiana S.A. - Calle 62 No. 20-46, Barrio San Luis, Bogotá, Colombia,

Tels.: (57-1) 746 0102 / 210 0415 - E-mail: cliente@alfaomega.com.co

Chile: Alfaomega Grupo Editor, S.A. -Av. Providencia 1443. Oficina 24, Santiago, Chile Tel.: (56-2) 2235-4248 - Fax: (56-2) 2235-5786 - E-mail: agechile@alfaomega.cl

Argentina: Alfaomega Grupo Editor Argentino, S.A. - Paraguay 1307 RB. Of. 11, C.P 1057, Buenos Aires, Argentina. -Tel./Fax: (541114811-0887 v 48117183-E-mail: ventas@alfaome2aeditor.com.ar

INTRODUCCION

Ante el boom actual de la información, las organizaciones han tratado de abordar el problema de analizar grandes volúmenes de datos desde muchos ángulos diferentes. Las herramientas de BIG DATA utilizan tecnologías multinúcleo para ofrecer mayor capacidad de procesamiento a través de altas prestaciones, en base de datos y de análisis en memoria que ofrecen un mayor conocimiento más rápidamente de grandes volúmenes de datos y flujo de datos. Y todo ello independientemente de los formatos y las fuentes de los orígenes de datos. Con las herramientas de BIG DATA se puede procesar información online proveniente de múltiples orígenes como pueden ser las redes sociales o grandes bases de datos no estructuradas. También se pueden tratar los datos de múltiples fuentes y formatos, ya sean texto, datos, imágenes o mezcla de todo ello. Actualmente es posible implementar herramientas de BIG DATA en la forma que mejor se adapte a las necesidades de los usuarios.

El término Big Data suele aplicarse a la información que no puede ser procesada o analizada usando procesos o herramientas tradicionales. Las organizaciones de hoy en día se enfrentan cada vez más a menudo a retos Big Data. Las empresas tienen acceso a una gran cantidad de información, pero no saben cómo obtener valor añadido de la misma, ya que la información aparece en su forma más cruda o en un formato semi-estructurado o no estructurado. Una encuesta de IBM demostró que más de la mitad de los líderes empresariales de hoy en día se dan cuenta de que no tienen acceso a los conocimientos que necesitan para analizar sus datos. Las empresas se enfrentan a estos retos en un clima en el que tienen la capacidad de almacenar cualquier cosa, que están generando datos como nunca antes en la historia y, sin embargo, tienen un verdadero desafío con el análisis de la información.

Las técnicas de Big Data persiguen complementar el manejo de grandes volúmenes de datos con las técnicas de análisis de la información más avanzadas y efectivas para extraer de modo óptimo el conocimiento contenido en los datos.

La base que actualmente caracteriza a las herramientas de BIG DATA es el paquete de código abierto llamado Hadoop para el análisis masivo de datos. Hadoop también se incluye como parte de las herramientas de todo el software de BIG DATA, como SAS, IBM, MICROSOFT y ORACLE. Por ejemplo, SAS incorpora Hadoop en sus aplicaciones (SAS Base SAS Data Integration, Sas Enterprise Guide, SAS Enterprise Miner, ...). También SAS permite trabajar en memoria a través de Hadoop (SAS Visual Analytics y SAS Visual Statistics). IBM trabaja con Hadoop en su plataforma IBM InfoSphere BigInsights (BigInsights). Microsoft incluye Hadoop en SQL Server 2014, Windows Server 2012, HDInsight and Polybase. Oracle incluye Hadoop en Oracle Big Data Appliance, Oracle Big Data Connectors y Oracle Loader for Hadoop.

Este libro presenta las posibilidades de trabajo que ofrecen las herramientas de BIG DATA para procesar y analizar grandes volúmenes de datos de una manera ordenada. A su vez, estas herramientas también permiten extraer el conocimiento contenido en los datos.

CAPITULO 1

CONCEPTOS DE BIG DATA

DEFINICIÓN, NECESIDAD Y CARACTERÍSTICAS DE BIG DATA

El término “Big data” suele aplicarse a conjuntos de datos que superan la capacidad del software habitual para ser capturados, gestionados y procesados en un tiempo razonable y por los medios habituales de procesamiento de la información. Este término suele referirse a los siguientes tipos de datos:

Datos de la empresa tradicional: incluye información de los clientes en sistemas de CRM, datos transaccionales ERP, las transacciones de tienda web, los datos contables, etcétera.

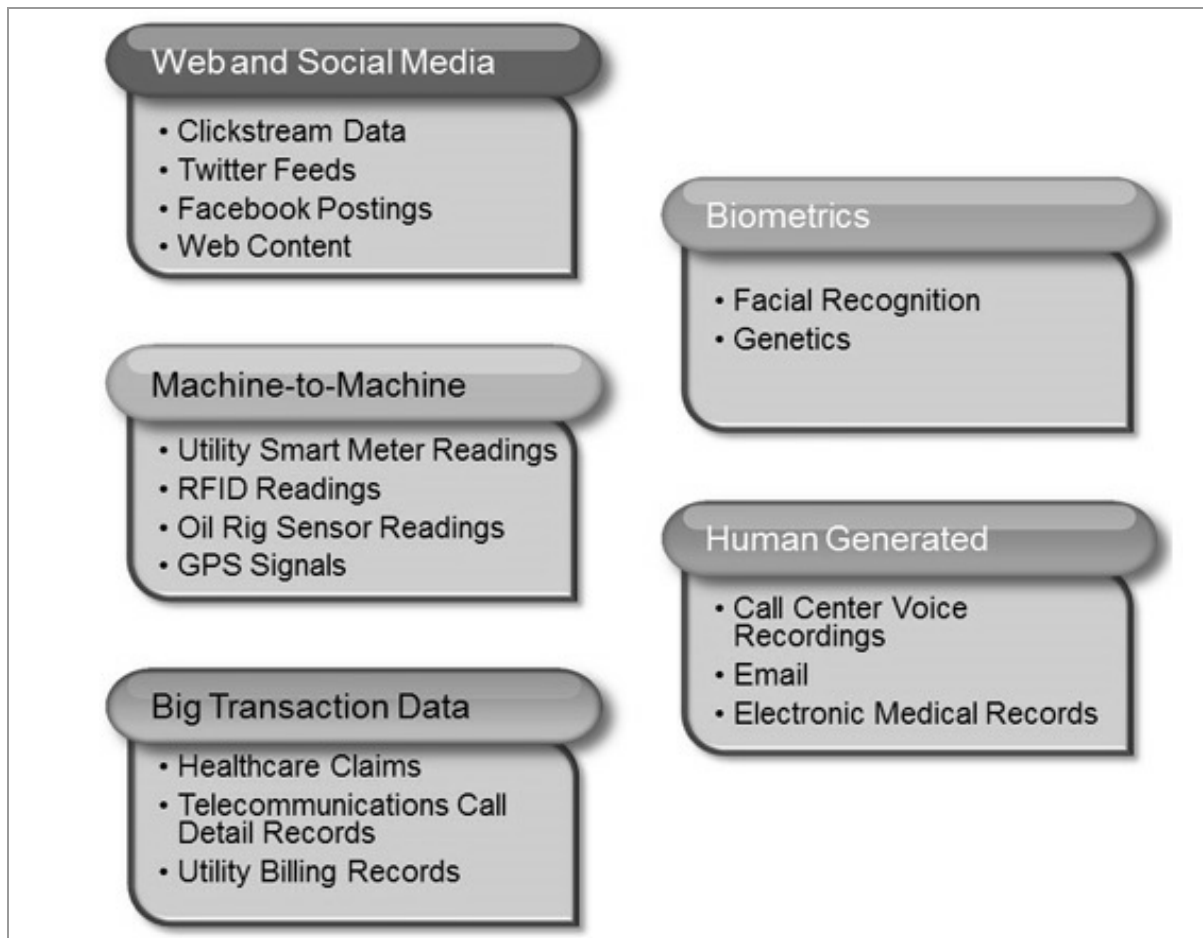
Machine-generated /sensor data: incluye registros de detalles de llamadas (“Call Detail Records, CDR”), los weblogs, los medidores inteligentes, los sensores de fabricación, registros de equipos, datos de sistemas comerciales, etc.

Datos de medios sociales: Incluye datos sobre blogs, Twitter, plataformas de Social Media como Facebook, etc.

Grandes bases de datos: con información multidimensional, relacional y no relacional.

Grandes conjuntos de datos no estructurados con mezcla de fuentes de origen y tipos de datos: numéricos, textuales, gráficos, etc.

El esquema siguiente amplía un poco más los tipos de datos a tener en cuenta en el tratamiento con técnicas de Big Data.



1. - *Web and Social Media*: incluye contenido web e información que es obtenida de las redes sociales como Facebook, Twitter, LinkedIn, etc., blogs.

2. - *Machine-to-Machine (M2M)*: M2M se refiere a las tecnologías que permiten conectarse a otros dispositivos. M2M utiliza dispositivos como sensores o medidores que capturan algún evento en particular (velocidad, temperatura, presión, variables meteorológicas, variables químicas como la salinidad, etc.), los cuales transmiten a través de redes alámbricas, inalámbricas o híbridas a otras aplicaciones que traducen estos eventos en información significativa.

3. - *Big Transaction Data*: incluye registros de facturación, en telecomunicaciones registros detallados de las llamadas (CDR), etc. Estos datos transaccionales están disponibles en formatos tanto semiestructurados como no estructurados.

4. - *Biometrics*: información biométrica en la que se incluye huellas digitales, escaneo de la retina, reconocimiento facial, genética, etc. En el área de seguridad e inteligencia, los datos biométricos han sido

información importante para las agencias de investigación.

5.- *Human Generated*: las personas generamos diversas cantidades de datos como la información que guarda un call center al establecer una llamada telefónica, notas de voz, correos electrónicos, documentos electrónicos, estudios médicos, etc.

Dentro del sector de tecnologías de la información y la comunicación, Big Data es una referencia a los sistemas que manipulan grandes conjuntos de datos. Las dificultades más habituales en estos casos se centran en la captura, almacenamiento, búsqueda, compartición, análisis y visualización.

Además del gran volumen de información, existe en una gran variedad de datos que pueden ser representados de diversas maneras en todo el mundo, por ejemplo de dispositivos móviles, audio, video, sistemas GPS, incontables sensores digitales en equipos industriales, automóviles, medidores eléctricos, veletas, anemómetros, etc., los cuales pueden medir y comunicar el posicionamiento, movimiento, vibración, temperatura, humedad y hasta los cambios químicos que sufre el aire, de tal forma que las aplicaciones que analizan estos datos requieren que la velocidad de respuesta sea lo demasiado rápida para lograr obtener la información correcta en el momento preciso. Estas son las características principales de las aplicaciones típicas de Big Data.

Dado el gran avance que existe día a día en las tecnologías de información, las organizaciones se han tenido que enfrentar a nuevos desafíos que les permitan analizar, descubrir y entender más allá de lo que sus herramientas tradicionales reportan sobre su información. La necesidad del Big Data surge al mismo tiempo que el gran crecimiento durante los últimos años de las aplicaciones disponibles en internet (geo-referenciamiento, redes sociales, etc.) que han sido parte importante en las decisiones de negocio de las empresas.

El concepto de Big Data se aplica para toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales. Hay cuatro características clave que definen la información relativa al Big Data:

- *Volumen*. Los datos relativos al Big data se producen en cantidades mucho más grandes que los datos tradicionales. Por ejemplo, un solo motor a reacción puede generar 10 TB de datos en 30 minutos. Con más de 25000 vuelos de aerolíneas por día, el volumen diario de solo

esta única fuente de datos se ejecuta en petabytes. Los medidores inteligentes y equipos industriales pesados como las refinerías de petróleo y plataformas de perforación generan volúmenes de datos similares, lo que agrava el problema.

- **Velocidad.** Los flujos de datos de medios sociales, aunque no es tan masivo como los datos generados por máquinas, producen una gran afluencia de opiniones y valiosas relaciones para la gestión de clientes. Incluso a 140 caracteres por tweet, la alta velocidad (o frecuencia) de los datos de Twitter proporciona grandes volúmenes de información (más de 8 TB por día).

- **Variedad.** Los formatos de datos tradicionales tienden a ser relativamente bien definidos por un esquema de datos. En contraste, los formatos de datos no tradicionales exhiben un ritmo vertiginoso del cambio. A medida que se añaden nuevos servicios, nuevos sensores desplegados, o nuevas campañas de marketing, se necesitan nuevos tipos de datos para capturar la información resultante.

- **Valor.** El valor económico de los diferentes datos varía significativamente. Por lo general hay buena información embebida en un gran conjunto más amplio de datos no tradicionales: El desafío esencial es identificar la información valiosa, transformarla y extraer los datos para su análisis. A partir de los datos convenientemente extraídos y transformados se analiza el conocimiento contenido en los mismos.

APLICACIONES TÍPICAS DE BIG DATA

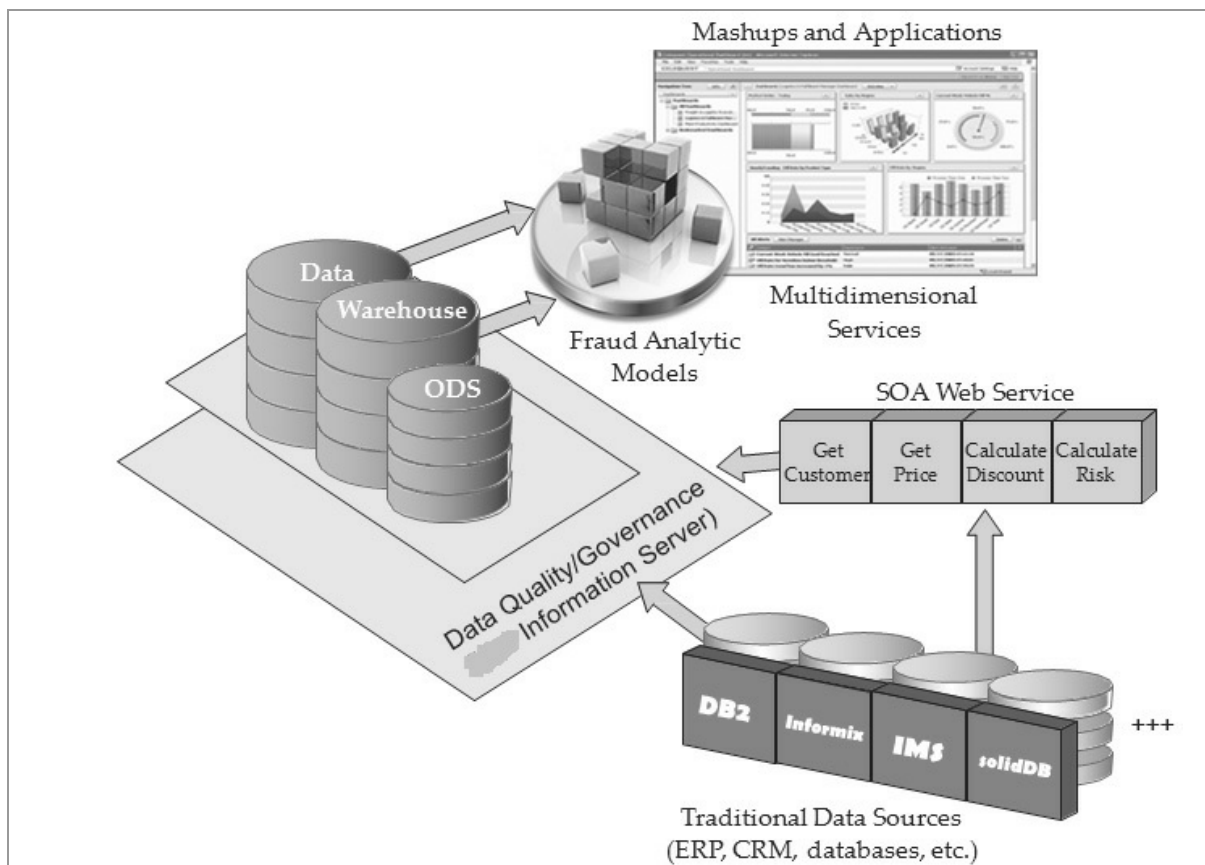
Existe una gran variedad de aplicaciones de las técnicas de Big Data. Siempre que sea necesario extraer el conocimiento inmerso en grandes volúmenes de datos estructurados, semiestructurados o no estructurados, tienen cabida las aplicaciones de Big Data. Pero estas técnicas no solo se aplican en la fase de análisis de la información, sino también en su propia recogida, transformación y puesta a disposición para los analistas. En los párrafos siguientes se citan algunos de los campos donde las técnicas de Big Data tienen más aplicación.

Patrones de detección del fraude

La detección de fraude es un problema típico en los servicios financieros verticales, pero se encuentra en cualquier tipo de transacciones (subastas en línea, juego online, reclamaciones de seguros, fraude fiscal, etc.). Prácticamente en cualquier lugar donde haya transacciones financieras está involucrado el fraude. Este tipo de transacciones presenta un potencial para el abuso y está omnipresente el fantasma del fraude. Una plataforma Big Data puede aportar la oportunidad de hacer más de lo que se ha hecho hasta ahora para identificar y paliar el fraude.

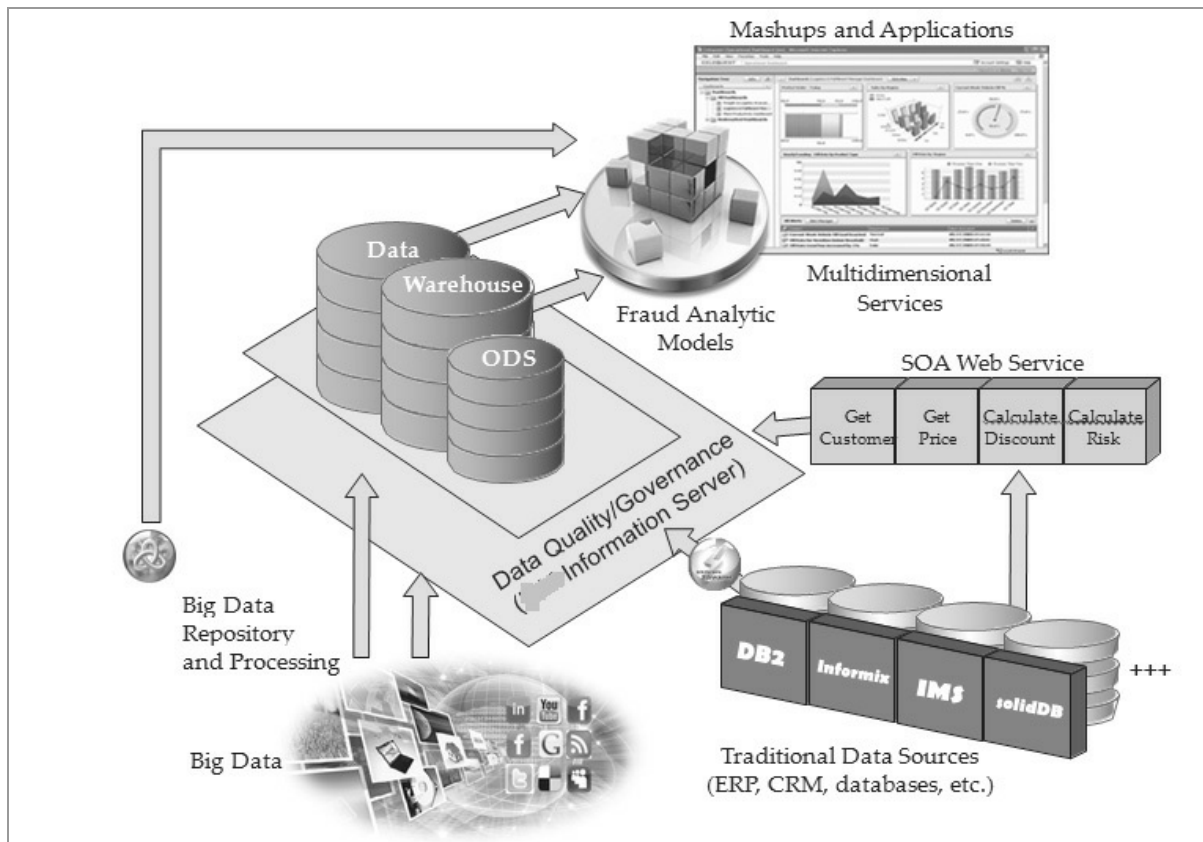
Varios desafíos en el patrón de detección de fraude son directamente atribuibles exclusivamente utilizando las tecnologías convencionales. El tema más común y recurrente que se observa en todos los patrones de Big Data es el relativo a los límites de almacenamiento de datos. Asimismo, son de gran importancia los recursos computacionales disponibles para procesar información relativa al fraude. Sin las tecnologías Big Data, estos factores limitan la información que puede ser analizada. Es más, entornos altamente dinámicos tienen patrones de fraude cíclico que van y vienen en horas, días o semanas. Si los datos utilizados para identificar o impulsar nuevos modelos de detección de fraude no está disponibles con inmediatez, el descubrimiento de los patrones de fraude puede llegar tarde cuando ya se haya ejecutado el daño.

Tradicionalmente, en casos de fraude, se utilizan muestras y modelos para identificar a los clientes que caracterizan a un determinado tipo de perfil. El problema con esta aproximación es que aunque funciona, está perfilando un segmento y no el microtratamiento a nivel transacción o persona individual. Sencillamente, hacer una previsión basada en un segmento es bueno, pero tomar una decisión basándose en los datos reales de una transacción individual es obviamente mejor. Para hacer esto, necesitamos trabajar con un conjunto mayor de datos que en el caso de la aproximación convencional tradicional. Se estima que a través de las herramientas tradicionales solo se está utilizando un 20 por ciento de la información disponible que podría ser útil para el modelado del fraude. El enfoque tradicional se muestra en la figura siguiente:



Es posible utilizar herramientas de Big Data para proveer un repositorio elástico y rentable para utilizar el 80 por ciento restante de la información y transformarla en útil para modelar el fraude. Posteriormente esta información alimentará la elaboración del modelo de fraude. En la figura siguiente se presenta el esquema. Se trata de un moderno sistema de detección de fraude típico de una plataforma de Big Data de bajo costo para modelado de exploración y descubrimiento. Se observa cómo pueden

aprovecharse los datos mediante sistemas tradicionales directamente o a través de la integración en protocolos de calidad y gestión de datos existentes.



Patrones de Social Media

Tal vez el patrón de uso de Big Data más comentado son los medios de comunicación social y el sentimiento del cliente. Puede utilizar Big Data para averiguar lo que los clientes opinan sobre uno mismo (y tal vez lo que están diciendo acerca de la competencia). Además, se puede utilizar este resultado recién encontrado para averiguar cómo esta información repercute en las decisiones y la forma en que su empresa se comporta. Más específicamente, puede determinar qué factores están impactando a las ventas, la efectividad o la receptividad de sus campañas de marketing, la exactitud de su marketing (producto, precio, promoción y colocación) y así sucesivamente.

Aunque los accesos básicos a las redes sociales pueden aportar la tendencia de las opiniones, no pueden responder lo que en definitiva es una cuestión más importante: “¿por qué dice la gente lo que están diciendo

y comportándose de la manera que se están comportando?”. La necesidad de este tipo de respuesta obliga a enriquecer el acceso a los medios de comunicación social con información adicional y en forma diferente que es probable que residen en múltiples sistemas empresariales. En pocas palabras, es necesaria la analítica de los medios de comunicación social utilizando también los repositorios de datos tradicionales (SAP, DB2, Teradata, Oracle, SAS, etc.). No obstante, es necesario mirar más allá de solo los datos. Hay que observar la interacción de las personas con sus comportamientos, tendencias financieras, transacciones reales y así sucesivamente. Ventas, promociones, programas de fidelización, acciones de mercado e incluso variables tales como el clima pueden ser conductores por los que podemos detectar el comportamiento de los consumidores para poder modelizarlo. Llegar a la base de por qué sus clientes están comportando de una determinada manera requiere tipos de información en forma dinámica y rentable, especialmente durante las fases de exploración inicial del proyecto.

Es un hecho que el análisis de los tweets es un indicador revelador sobre el impacto potencial del sentimiento del cliente sobre los productos. Este tipo de registros es muy elocuente, no solo por el volumen y la velocidad de su crecimiento, sino también porque el sentimiento está siendo expresado para cualquier producto o servicio. Además, todo el mundo es capaz de expresar la reacción y sentimiento en segundos y sin filtros ni trabas geográficas.

Patrones de modelado y gestión de riesgo

El modelado para la gestión de riesgos es otro patrón de aplicación y uso común del Big Data. La crisis financiera de 2008, la crisis de las hipotecas “subprime” asociadas y sus secuelas han hecho del modelado de riesgos y su gestión, un área clave de interés para las instituciones financieras. Como se sabe por los mercados financieros de hoy, una carencia de entender el riesgo puede tener efectos devastadores de creación de riqueza. Además, conocidas las normas reguladoras que afectan a las instituciones financieras en todo el mundo, es necesario asegurarse rápidamente de que los niveles de riesgo caen dentro de límites aceptables.

Como fue el caso en el patrón de detección de fraude, las empresas utilizan entre el 15 y 20 por ciento de los datos estructurados disponibles

en sus modelos de riesgo. No es que no se reconozca que hay un montón de datos que están potencialmente subutilizados, sino que no saben dónde puede encontrarse la información relevante en el resto de los datos. Además, puede ser demasiado caro en la infraestructura actual de muchas empresas analizar a muchos clientes para investigar.

También es típico analizar lo que pasa al final de una jornada bursátil en una firma financiera. Es esencial conseguir una instantánea de sus posiciones a la clausura de la jornada. Instantáneamente, las empresas pueden derivar e identificar su posición financiera usando sus modelos en poco tiempo e informar a los reguladores para el control de riesgos internos.

Dos problemas iniciales se asocian a este patrón de uso de modelado y gestión de riesgo: “¿cuántos datos se van a utilizar para el modelo?” y “¿cuál es la velocidad de los datos?”. Desafortunadamente, la respuesta a la segunda pregunta es a menudo difícil. Finalmente, se persigue considerar la tendencia de servicios financieros para mover el modelo de riesgo y ajustar las posiciones del día a día. Este desafío no puede ser resuelto con los sistemas tradicionales. Otra característica de los mercados financieros de hoy es que hay enormes volúmenes de comercio. Si mezclamos los picos de volumen con los requisitos para construir el mejor modelo y gestionar el riesgo adecuadamente con ejecución diaria, tenemos un problema de Big Data delante de nosotros.

Big Data y el sector de la energía

El sector de la energía ofrece muchos retos de casos de uso de Big Data. El problema principal consiste en cómo hacer frente a los grandes volúmenes de datos de los sensores de las instalaciones remotas. Muchas empresas están utilizando solo una fracción de los datos, ya que carecen de la infraestructura necesaria para almacenar o analizar la escala de los datos disponibles.

Tomemos por ejemplo una plataforma de perforación de petróleo típico que puede tener de 20000 a 40000 sensores a bordo. Todos estos sensores están fluyendo los datos sobre la calidad de la plataforma petrolera y otras variables. No todos los sensores están en acción en todo momento, pero algunos están reportando muchas veces por segundo. Se

necesita tener una pista sobre qué porcentaje de esos sensores se utilizan activamente, aunque conocer todo el problema sea imposible.

De manera similar los clientes no están utilizando toda la información de datos que están disponibles para ellos en su proceso de toma de decisiones. Por supuesto, cuando se trata de datos de energía, tasas de recaudación o variables

similares, lo que realmente nos preguntamos es si hemos hecho todo lo posible para la captura y el aprovechamiento de la información que se está recopilando.

Con la idea de la ganancia, la seguridad y la eficiencia en mente, las empresas deben estar constantemente en busca de señales y ser capaces de relacionar esas señales con sus resultados potenciales o probables. Si se descarta el 90 por ciento de los datos de los sensores, no es posible que se puedan comprender o modelar las correlaciones existentes.

Big Data en el Call Center

El reto de la eficiencia del centro de llamadas es similar al caso de la detección del patrón de fraude. Al igual que la dinámica apropiada en información de fraude es crítica para los modelos de fraude robustos, en un centro de llamadas si no se gestiona bien la relación entre el tiempo de la resolución de la llamada y la gestión posterior de los patrones de descontento, la información recogida va a perder su valor. Es vital poder aplicar un patrón de respuesta óptima dinámicamente de modo que los desfases de tiempo de respuesta no resulten nocivos. Esta gestión exige el uso de herramientas de Big Data.

CAPÍTULO 2

COMPONENTES DE UNA PLATAFORMA DE BIG DATA

PLATAFORMA DE CÓDIGO ABIERTO HADOOP

Hadoop es una infraestructura digital de desarrollo creada en código abierto bajo licencia Apache. Se trata de un proyecto construido y utilizado por una gran variedad de programadores que usan Java. Doug Cutting inició su desarrollo cuando estaba en Yahoo! inspirándose en tecnologías liberadas por Google, concretamente MapReduce y Google File System (GFS), con el fin de utilizarla como base para un motor de búsqueda distribuido. Tras dedicarse a tiempo completo a su desarrollo y convertir a Yahoo! en el principal contribuidor del proyecto, Cutting abandonó Yahoo! para unirse a Cloudera, una compañía cuya oferta de productos gira íntegramente en torno a Hadoop.

La importancia de Hadoop radica básicamente en que permite desarrollar tareas muy intensivas de computación masiva, dividiéndolas en pequeñas piezas y distribuyéndolas en un conjunto de máquinas todo lo grande que se quiera. El análisis se realiza en petabytes de datos, en entornos distribuidos formados por muchas máquinas sencillas. Se trata de una propuesta de valor muy razonable en los tiempos hiperconectados que vivimos, y que utilizan hasta la saciedad empresas como Google, Yahoo!, Tuenti, Twitter, eBay o Facebook. Pero no son las únicas: el uso de Hadoop se está popularizando a gran velocidad en todo tipo de empresas.

Además, es un caso interesante, porque su licencia libre está haciendo que sea adoptado por un gran número de competidores, incluyendo Oracle, Dell, NetApp, EMC, etc. Este hecho está llevando a una aceleración tanto de su difusión como de sus prestaciones. Si estás en el mundo de la

tecnología corporativa o preparando tu desarrollo profesional dentro del mismo, Hadoop es una de las áreas que, en función de su potencial, deberías definitivamente considerar: más tarde o más temprano, te encontrarás con el elefante.

La plataforma de código abierto Hadoop ostenta el liderazgo en la actualidad como herramienta para analizar grandes cantidades de datos.

Hadoop está inspirado en el proyecto de Google File System (GFS) y en el paradigma de programación *MapReduce*, el cual consiste en dividir en dos tareas (*mapper* - *reducer*) la manipulación de los datos distribuidos a nodos de un cluster logrando un alto paralelismo en el procesamiento. Hadoop está compuesto de tres piezas: *Hadoop Distributed File System* (HDFS), *Hadoop MapReduce* y *Hadoop Common*.

Hadoop Distributed File System (HDFS)

El tema de la computación de alto rendimiento (HPC o *High Performance Computing*) lleva ya años dando vueltas, y hay soluciones ya maduras y establecidas (tanto gestores de colas como Condor, Oracle Grid Engine, Torque en la parte de cluster como Globus o Glite en la parte de grid computing).

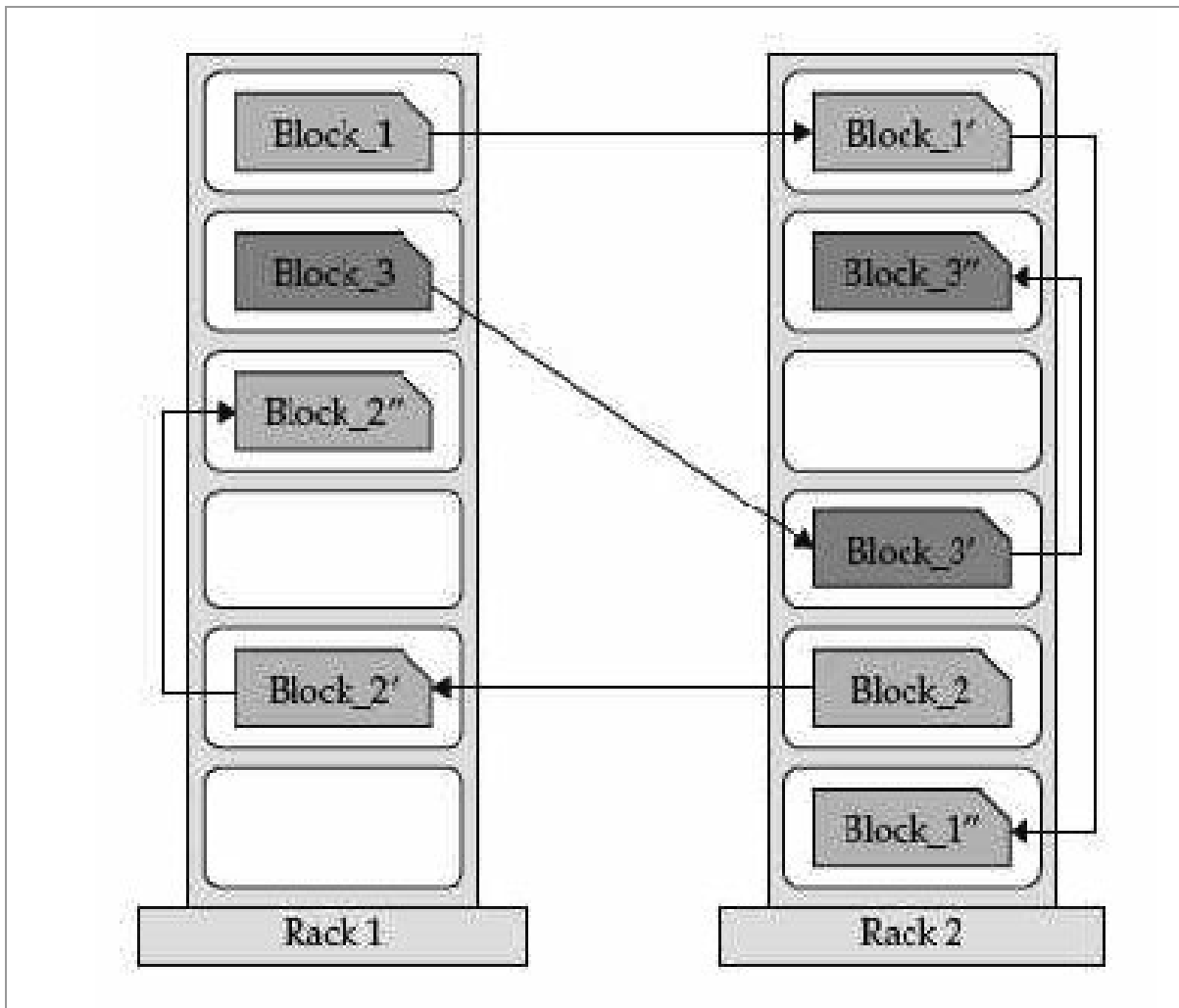
Lo que realmente aporta Hadoop es una capacidad de gestionar grandes cantidades de datos. Los cluster tradicionales están orientados a tener que dar mucha potencia de cálculo gestionando relativamente poco espacio en disco, pero ¿qué pasa cuando la base de datos tiene 100 Tb o 1 Pb? En estos casos se necesita algo más potente como Hadoop.

El HDFS (Hadoop Distributed File System) es quizás el componente principal de Hadoop, ya que permite crear sistemas de ficheros empleando servidores “commodity” ofreciendo redundancia, capacidad y rendimiento (solo para ficheros muy grandes, ojo). Y lo mejor de todo es que estos servidores commodity son los que hacen la computación, permitiendo el paradigma de “llevar los datos a la computación”, uno de los factores principales del rendimiento de Hadoop.

Los datos en el cluster de Hadoop son divididos en pequeñas piezas llamadas *bloques* y distribuidas a través del cluster. De esta manera, las funciones *map* y *reduce* pueden ser ejecutadas en pequeños subconjuntos y esto provee de la escalabilidad necesaria para el procesamiento de grandes

volúmenes.

La siguiente figura ejemplifica cómo los bloques de datos son escritos hacia HDFS. Observe que cada bloque es almacenado tres veces y al menos un bloque se almacena en un diferente Rack para lograr redundancia.



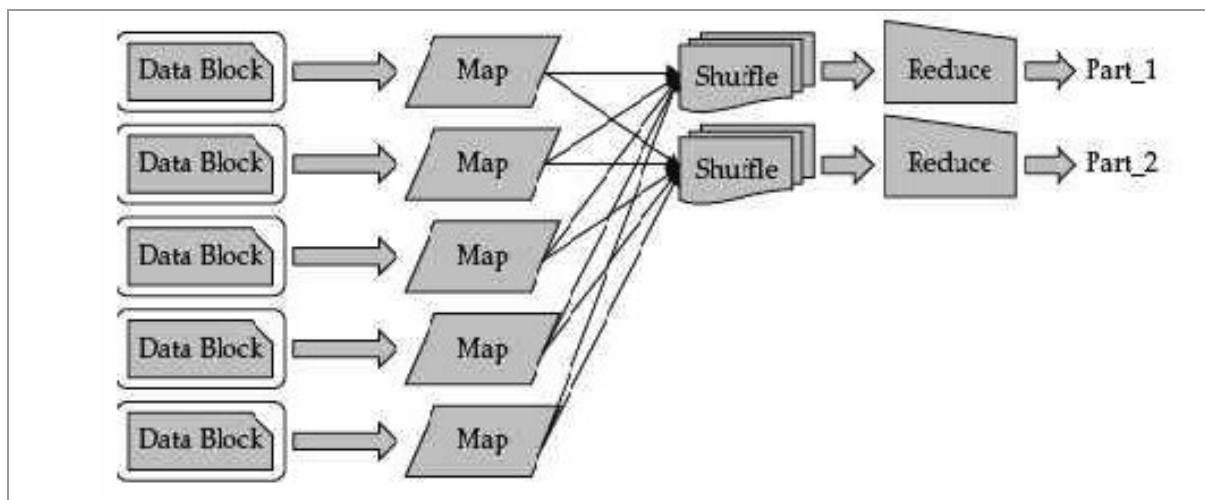
Hadoop MapReduce

Map/Reduce es un distribuidor de tareas que encaja perfectamente con HDFS y que permite de forma bastante sencilla el repartir trozos de tareas entre el cluster con una curva de aprendizaje relativamente sencilla (si lo que se va a analizar no son ficheros de texto cuesta más, pero es posible trabajar con vídeo o imágenes).

MapReduce es el núcleo de Hadoop. El término MapReduce en realidad se refiere a dos procesos separados que Hadoop ejecuta. El primer

proceso es *map*, el cual toma un conjunto de datos y lo convierte en otro conjunto, donde los elementos individuales son separados en tupias (pares de clave/valor). El proceso *reduce* obtiene la salida de *map* como datos de entrada y combina las tupias en un conjunto más pequeño de las mismas. Una fase intermedia es la denominada *Shuffle* la cual obtiene las tupias del proceso *map* y determina qué nodo procesará estos datos dirigiendo la salida a una tarea *reduce* en específico.

La siguiente figura ejemplifica un flujo de datos en un proceso sencillo de MapReduce.



Hadoop Common

Hadoop Common Components son un conjunto de librerías que soportan varios subproyectos de Hadoop.

APLICACIONES DE DESARROLLO EN HADOOP

Como se deduce de la sección anterior, la plataforma Hadoop puede ser una poderosa herramienta para manipular grandes conjuntos de datos. Sin embargo, el núcleo Hadoop - MapReduce - APIs se basa principalmente en Java, que requiere de programadores capacitados. Además, es aún más complejo para los programadores desarrollar y mantener aplicaciones MapReduce para aplicaciones empresariales que requieren un procesamiento largo y canalizado.

Si se lleva bastante tiempo de programación, se observará que la historia tiene una forma de repetirse. Por ejemplo, a menudo citamos XML como “La venganza de IMS” debido a su naturaleza jerárquica y sistema de recuperación. En el área del desarrollo del lenguaje Computer, tal como ensamblador dio lugar a lenguajes estructurados de programación y luego al desarrollo de lenguajes 3GL y 4GL, así también funciona el mundo de los lenguajes de programación de aplicación Hadoop. Para abstraerse de la complejidad del modelo de programación Hadoop, han surgido varios lenguajes de programación de aplicaciones que se ejecutan sobre Hadoop. A continuación se citan y describen varios de entre los más populares.

Además de todo esto, se ha creado un verdadero ecosistema encima de Hadoop con cosas como HIVE (Datawarehousing), HBase (BD NoSQL), Pig (Framework) o Mahout (Machine Learning/Datamining) que hace que en algunos casos el desarrollador ni siquiera tenga que pegarse con “lo complicado” de Hadoop.

Por lo tanto, además de los tres componentes principales de Hadoop, existen otros proyectos relacionados, los cuales se definen a continuación:

Avro

Es un proyecto de Apache que provee servicios de serialización. Cuando se guardan datos en un archivo, el esquema que define ese archivo es guardado dentro del mismo; de este modo es más sencillo para cualquier aplicación leerlo posteriormente puesto que el esquema está definido

dentro del archivo.

Cassandra

Cassandra es una base de datos no relacional distribuida y basada en un modelo de almacenamiento de <clave-valor>, desarrollada en Java. Permite grandes volúmenes de datos en forma distribuida. Twitter es una de las empresas que utiliza Cassandra dentro de su plataforma.

Chukwa

Diseñado para la colección y análisis a gran escala de “logs”. Incluye un toolkit para desplegar los resultados del análisis y monitoreo.

Flume

Tal como su nombre lo indica, su tarea principal es dirigir los datos de una fuente hacia alguna otra localidad, en este caso hacia el ambiente de Hadoop. Existen tres entidades principales: sources, decorators y sinks. Un source es básicamente cualquier fuente de datos, sink es el destino de una operación en específico y un decorator es una operación dentro del flujo de datos que transforma esa información de alguna manera, como por ejemplo comprimir o descomprimir los datos o alguna otra operación en particular sobre los mismos.

HBase (NoSQL)

Es una base de datos columnar (column-oriented database) que se ejecuta en HDFS. HBase no soporta SQL, ya que no es una base de datos relacional. Cada tabla contiene filas y columnas como una base de datos relacional. HBase permite que muchos atributos sean agrupados llamándolos familias de columnas, de tal manera que los elementos de una familia de columnas son almacenados en un solo conjunto. Eso es distinto a las bases de datos relacionales orientadas a filas, donde todas las columnas de una fila dada son almacenadas en conjunto. Facebook utiliza HBase en su plataforma desde noviembre de 2010.

Hive

Es una infraestructura de data warehouse que facilita administrar grandes conjuntos de datos que se encuentran almacenados en un ambiente distribuido. Hive tiene definido un lenguaje similar a SQL llamado Hive Query Language (HQL), estas sentencias HQL son separadas por un servicio de Hive y son enviadas a procesos MapReduce ejecutados en el cluster de Hadoop.

El siguiente es un ejemplo en HQL para crear una tabla, cargar datos y obtener información de la tabla utilizando Hive:

```
CREATE TABLE Tweets (from_user STRING, userid BIGINT, tweettext
STRING, retweets INT)
COMMENT 'This is the Twitter feed table'
STORED AS SEQUENCEFILE;
LOAD DATA INPATH 'hdfs://node/tweetdata' INTO TABLE TWEETS;
SELECT from_user, SUM(retweets)
FROM TWEETS
GROUP BY from_user;
```

Jaql

Fue donado por IBM a la comunidad de software libre. Query Language for JavaScript Object Notation (JSON) es un lenguaje funcional y declarativo que permite la explotación de datos en formato JSON diseñado para procesar grandes volúmenes de información. Para explotar el paralelismo, Jaql reescribe los queries de alto nivel (cuando es necesario) en queries de “bajo nivel” para distribuirlos como procesos MapReduce.

Internamente el motor de Jaql transforma el query en procesos *map* y *reduce* para reducir el tiempo de desarrollo asociado en analizar los datos en Fladoop. Jaql posee una infraestructura flexible para administrar y analizar datos semiestructurados como XML, archivos CSV, archivos planos, datos relacionales, etc.

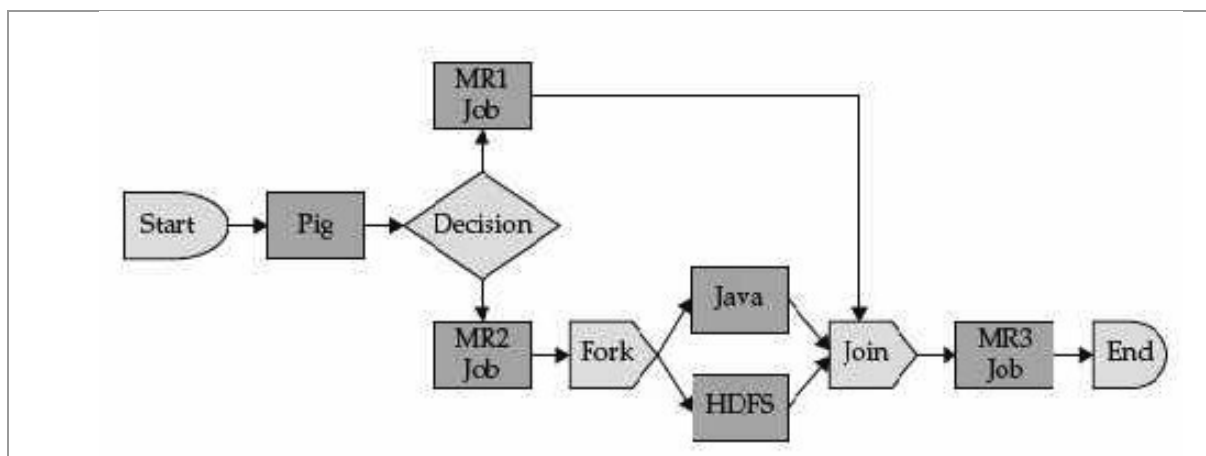
Lucene

Es un proyecto de Apache bastante popular para realizar búsquedas sobre textos. Lucene provee librerías para indexación y búsqueda de texto. Ha sido principalmente utilizado en la implementación de motores de búsqueda (aunque hay que considerar que no tiene funciones de “crawling” ni análisis de documentos HTML ya incorporadas). El concepto a nivel de arquitectura de Lucene es simple, básicamente los documentos (document) son divididos en campos de texto (fields) y se genera un índice sobre estos campos de texto. La indexación es el componente clave de Lucene, lo que le permite realizar búsquedas rápidamente independientemente del formato del archivo, ya sean PDFs, documentos HTML, etc.

Oozie

Existen varios procesos que son ejecutados en distintos momentos, los cuales necesitan ser orquestados para satisfacer las necesidades de tan complejo análisis de información. Oozie es un proyecto de código abierto que simplifica los flujos de trabajo y la coordinación entre cada uno de los procesos. Permite que el usuario pueda definir acciones y las dependencias entre dichas acciones.

Un flujo de trabajo en Oozie es definido mediante un grafo acíclico llamado *Directed Acyclical Graph* (DAG), y es acíclico puesto que no permite ciclos en el grafo; es decir, solo hay un punto de entrada y de salida y todas las tareas y dependencias parten del punto inicial al punto final sin puntos de retorno. Un ejemplo de un flujo de trabajo en Oozie se representa de la siguiente manera:



Pig

Inicialmente desarrollado por Yahoo para permitir a los usuarios de Hadoop enfocarse más en analizar todos los conjuntos de datos y dedicar menos tiempo en construir los programas MapReduce. Tal como su nombre indica, al igual que cualquier cerdo que come cualquier cosa, el lenguaje PigLatin fue diseñado para manejar cualquier tipo de dato y Pig es el ambiente de ejecución donde estos programas son ejecutados, de manera muy similar a la relación entre la máquina virtual de Java (JVM) y una aplicación Java.

ZooKeeper

ZooKeeper es otro proyecto de código abierto de Apache que provee de una infraestructura centralizada y de servicios que pueden ser utilizados por aplicaciones para asegurarse de que los procesos a través de un cluster sean señalizados o sincronizados.

Internamente en ZooKeeper una aplicación puede crear un archivo que se persiste en memoria en los servidores ZooKeeper llamado znode. Este archivo znode puede ser actualizado por cualquier nodo en el cluster, y cualquier nodo puede registrar que sea informado de los cambios ocurridos en ese znode; es decir, un servidor puede ser configurado para “vigilar” un znode en particular. De este modo, las aplicaciones pueden sincronizar sus procesos a través de un cluster distribuido actualizando su estatus en cada znode, el cual informará al resto del cluster sobre el estatus correspondiente de algún nodo en específico.

Como podrá observar, más allá de Hadoop, una plataforma de Big Data consiste de todo un ecosistema de proyectos que en conjunto permiten simplificar, administrar, coordinar y analizar grandes volúmenes de información.

HADOOP STREAMING

En adición a Java, se pueden escribir funciones map y reduce en otros lenguajes e invocarlos usando un API conocido como Hadoop Streaming. Streaming sigue en el concepto de UNIX streaming, donde la entrada se lee desde la entrada estándar, y la salida se escribe en la salida estándar. Estos flujos de datos representan la interfaz entre Hadoop y sus aplicaciones.

La interfaz de Streaming se presta mejor a cortas y simples aplicaciones que típicamente se desarrollan utilizando un lenguaje de Scripts como Python o Ruby. Una razón importante para esto es la naturaleza basada en texto del flujo de datos, donde cada línea de texto representa un solo registro.

SITUAR DATOS EN HADOOP

Uno de los retos con HDFS es que no es un sistema de archivos compatibles con POSIX. Esto significa que todas las cosas que usted está acostumbrado cuando se trata de interactuar con un sistema de archivos típico (copiar, crear, mover, borrar o acceder a un archivo y mucho más) no se aplican automáticamente en HDFS. Para hacer cualquier cosa con un archivo en HDFS, debe utilizar las interfaces HDFS o APIs directamente. Sin embargo, otra ventaja es usar el sistema de archivos GPFS-SNC. Con GPFS-SNC, se interactúa con grandes archivos de la misma manera que lo haría cualquier otro sistema de archivos y, por lo tanto, tareas de manipulación de archivos con Hadoop ejecutándose sobre GPFS-SNC se reducen considerablemente.

COPIA DE DATOS BÁSICA

Se deben utilizar comandos específicos para mover archivos en HDFS a través de APIs o utilizando el shell de comandos. La forma más común de mover archivos de un sistema de archivos local en HDFS es mediante el comando `copyFromLocal`. Para obtener archivos de HDFS al sistema de archivos local, normalmente se utilizará el comando `copyToLocal`. Aquí se muestra un ejemplo de cada uno de estos comandos:

```
HDFS dfs - copyFromLocal /user/dir/file
hdfs://si.ni.com/dir/hdfsfile hdfs dfs - copyToLocal
hdfs://si.ni.com/dir/hdfsfile /user/dir/file
```

Estos comandos se ejecutan a través del programa shell HDFS, que simplifica una aplicación Java. La máscara utiliza la API de Java para obtener datos que entran y salen en HDFS. Estas API se pueden llamar desde cualquier aplicación Java.

El problema con este método es que se debe disponer de desarrolladores Java con la finalidad de escribir los programas leyendo y escribiendo datos de HDFS. Otros métodos están disponibles (como C++ APIs, o mediante el marco de segunda mano para los servicios de comunicación en varios idiomas), pero estos son meramente contenedores para la base API de Java. Si usted necesita acceder a archivos HDFS desde sus aplicaciones Java, utilizaría los métodos en el paquete `org.apache.hadoop.fs`. Esto permite incorporar a leer y escribir las operaciones directamente y desde HDFS, desde dentro de sus aplicaciones de MapReduce. Sin embargo, HDFS está diseñado para escritura y lectura secuencial. Esto significa que cuando escribe datos en un archivo HDFS, puede escribir solamente al final del archivo (se denomina un APPEND en el mundo de base de datos). Aquí reside otra ventaja de utilizar GPFS-SNC como la columna vertebral sistema de archivo para su cluster de Hadoop.

BIG DATA Y EL CAMPO DE LA INVESTIGACIÓN

Los científicos e investigadores han analizado datos desde ya hace mucho tiempo, lo que ahora representa el gran reto es la escala en la que estos son generados.

Esta explosión de “grandes datos” está transformando la manera en que se conduce una investigación adquiriendo habilidades en el uso de Big Data para resolver problemas complejos relacionados con el descubrimiento científico, investigación ambiental y biomédica, educación, salud, seguridad nacional, entre otros.

Con la capacidad de generar toda esta información valiosa de diferentes sistemas, las empresas y los gobiernos están lidiando con el problema de analizar los datos para dos propósitos importantes: ser capaces de detectar y responder a los acontecimientos actuales de una manera oportuna, y para poder utilizar las predicciones del aprendizaje histórico. Esta situación requiere del análisis tanto de datos en movimiento (datos actuales) como de datos en reposo (datos históricos), que son representados a diferentes y enormes volúmenes, variedades y velocidades.

La naturaleza de la información hoy es diferente a la información en el pasado. Debido a la abundancia de sensores, micrófonos, cámaras, escáneres médicos, imágenes, etc., en nuestras vidas, los datos generados a partir de estos elementos serán dentro de poco el segmento más grande de toda la información disponible.

El uso de Big Data ha ayudado a los investigadores a descubrir cosas que les podrían haber llevado años en descubrir por sí mismos sin el uso de estas herramientas, debido a la velocidad del análisis, es posible que el analista de datos pueda cambiar sus ideas basándose en el resultado obtenido y retrabajar el procedimiento una y otra vez hasta encontrar el verdadero valor al que se está tratando de llegar.

Implementar una solución alrededor de Big Data Implica de la Integración de diversos componentes y proyectos que en conjunto forman el ecosistema necesario para analizar grandes cantidades de datos.

Sin una plataforma de Big Data se necesitaría desarrollar adicionalmente código que permita administrar cada uno de esos componentes como por ejemplo: manejo de eventos, conectividad, alta disponibilidad, seguridad, optimización y desempeño, depuración, monitoreo, administración de las aplicaciones, SQL y scripts personalizados.

CAPÍTULO 3

BIG DATA CON HERRAMIENTAS DE IBM

IBM POWER SYSTEMS

Ante el crecimiento exponencial de Big Data y de analítica de negocio, las organizaciones buscan cómo utilizar y convertir volúmenes masivos de datos en bruto en conocimiento inteligente. Hoy en día, los negocios pueden analizar, supervisar y predecir resultados más rápido que nunca antes y aquellos que lo hacen consiguen un rendimiento superior al de sus competidores.

Power Systems es una herramienta de vanguardia en el suministro de herramientas que obtienen información más rápida a partir de la analítica de información estructurada y de Big Data no estructurada, como vídeo, imágenes y contenido procedente de dispositivos móviles, redes sociales y sensores. Para extraer información y tomar mejores decisiones, las compañías precisan una herramienta, software de sistemas abierto y una plataforma flexible y segura como Power Systems para dar soporte a la continuada carga de datos, ejecutar múltiples consultas simultáneas y ofrecer analítica en tiempo real soportada por un ancho de banda de E/S masivo.

Power System incorpora hardware, software y herramientas adecuadas para el trabajo en Big Data.

Hardware

Power System incorpora servidores escalables, rentables, fáciles de desplegar y eficientes energéticamente como servidores de aplicaciones, servidores de consolidación o servidores autónomos para cargas de trabajo

Linux, UNIX e IBM. Disponibles con hasta 24 cores, estos servidores de 1 y 2 sockets ofrecen mayor rentabilidad y seguridad en cloud para empresas que necesitan opciones de despliegue inferiores o escalables para aplicaciones centradas en datos.

Los sistemas Power Systems basados solo en Linux que ejecutan Ubuntu, SUSE o Red Hat Linux tienen un precio competitivo en relación con otras alternativas, además de ofrecer un rendimiento superior y un mayor retorno de la inversión para aplicaciones de gran intensidad de cálculo y datos. Con ambas opciones de virtualización, PowerVM y PowerKVM, disponibles, estos sistemas ofrecen la flexibilidad necesaria para integrar de forma rápida herramientas de tecnología innovadora, evitar la dependencia de un solo proveedor y acelerar los resultados de negocio.

En cuanto al enfoque empresarial, los servidores Power Enterprise están diseñados para trabajo con datos masivos y ofrecen a las empresas lo último en disponibilidad, seguridad y rendimiento. Esta clase de sistema, capaz de ejecutar sistemas de relación junto con sistemas de registro, ejecuta AIX (UNIX), IBM i y Linux, proporciona hasta 256 núcleos de procesador POWER7 con hasta 8 TB de memoria, e incluye la flexibilidad para activar y desactivar procesos y memoria, según lo demande la carga de trabajo de la aplicación.

Las herramientas informáticas de alto rendimiento Power Systems configuradas en clusters AIX y Linux muy escalables, ofrecen un rendimiento excelente para demandar análisis y cargas de trabajo de grandes volúmenes de datos, como la química computacional, la creación de modelos para la determinación de depósitos de petróleo, la previsión meteorológica, la creación de modelos climáticos o los servicios financieros.

Por otra parte, IBM PureFlex System ofrece recursos de cálculo, almacenamiento y red en un entorno eficaz y fácil de gestionar. Los componentes de IBM Flex System ofrecen un ecosistema abierto de tecnologías avanzadas de red, almacenamiento y virtualización con flexibilidad para adaptarse a grandes cargas de trabajo.

Sistemas operativos

Los servidores Power ofrecen flexibilidad y opción de sistemas operativos para trabajar con las mejores aplicaciones de acuerdo con las necesidades empresariales. Los sistemas operativos AIX, IBM i y Linux, junto con PowerVM, maximizan las ventajas de Power Systems en el trabajo habitual.

Los servidores con tecnología Power ejecutan distribuciones Linux estándar del sector Red Hat, SUSE y Ubuntu y su precio supone una alternativa más escalable para las opciones de x86. Ahora también dan soporte a PowerKVM (US), la opción de virtualización abierta para los servidores Linux de escalabilidad horizontal de Power.

El sistema operativo AIX aprovecha décadas de innovación en tecnología de IBM y está diseñado para ofrecer el nivel más elevado de rendimiento, seguridad y fiabilidad de cualquier sistema operativo UNIX. Este sistema operativo tiene características muy importantes como compatibilidad binaria con versiones anteriores de IBM AIX, escalabilidad vertical que permite a su infraestructura de TI un mayor crecimiento y capacidad, funcionalidades de cluster Incluidas para simplificar una alta disponibilidad (HA), mejoras en las prestaciones de virtualización para ofrecer una mayor flexibilidad para soportar cargas de trabajo variables. Además, utiliza la virtualización y tecnología IBM POWER para permitir un rendimiento y una eficiencia espectaculares y está disponible en tres versiones para obtener incluso una mayor flexibilidad y capacidad.

IBM i es un entorno operativo integrado con una reputación arraigada por su excepcional seguridad y resistencia. IBM I (anteriormente conocido como i5/OS), que se ejecuta en el servidor IBM Power Systems, ofrece una arquitectura resistente a virus y altamente escalable con un amplio reconocimiento como entorno de excepcional solidez empresarial. La ejecución de aplicaciones basadas en i ha ayudado a las empresas a centrarse durante años en la Innovación y en ofrecer un nuevo valor a sus negocios y no solo en gestionar las operaciones del centro de datos. Contar con una Infraestructura más dinámica (DI) solo consiste en seleccionar los sistemas y el software adecuados para que las empresas avancen con agilidad y rapidez. Gracias a IBM i podrá implementar reconocidas herramientas en una plataforma en la que puede confiar. Al elegir la plataforma Power más actual, las aplicaciones IBM I logran un rendimiento de primer nivel además de la flexibilidad de las Infraestructuras dinámicas, con la posibilidad de reducir los costes

operativos mensuales.

IBM i integra una combinación fiable de funciones de bases de datos relacionales, seguridad, servicios web, redes y gestión de almacenamiento. Ofrece unos fundamentos amplios y altamente estables de middleware y base de datos para la implementación eficaz de aplicaciones de procesamiento empresarial, con soporte de más de 5000 herramientas procedentes de más de 2500 proveedores de software independientes (ISV). Las herramientas i se ofrecen a través de una amplia red mundial especializada de Business Partners de IBM, respaldada por la infraestructura de soporte y los fiables servicios de IBM.

La virtualización y la gestión de cargas de trabajo también se integran en i para permitirle ejecutar varias aplicaciones y componentes de forma conjunta en el mismo sistema, aumentando así la utilización del sistema y ofreciendo una mejor rentabilidad de las inversiones en TI.

Software System

Software System es el nombre que adopta el software de IBM para Power Systems. Este software habilita la virtualización, alta disponibilidad, flexibilidad, seguridad y conformidad en Power Systems. Este software se enfoca esencialmente hacia los siguientes campos:

Inteligencia de rendimiento de virtualización - PowerVP (US)

Virtualización sin límites - PowerVM

Resiliencia sin tiempos de inactividad - PowerHA (US)

Seguridad y conformidad - PowerSC (US)

Centro de virtualización - PowerVC (US)

Gestión energética (US)

Herramientas

Las herramientas de IBM e IBM Business Partner aprovechan las ventajas clave de Power Systems para proporcionar a su empresa nuevas funcionalidades y nuevas ventajas competitivas. Entre las herramientas más importantes destacan:

Analytics Cloud computing Colaboración Business Intelligence

ANALYTICS CON POWER SYSTEM

Analítica y Big Data forman parte de la estrategia de negocio y permiten obtener información de valor de forma rápida. Con el espectacular crecimiento de datos estructurados y no estructurados procedentes de múltiples fuentes, las empresas buscan extraer información de valor de sus datos. Los sistemas optimizados para Big Data y aplicaciones analíticas de gran intensidad de cálculo ayudan a aportar información de valor en el punto de impacto. IBM aporta las herramientas analíticas que se comentan brevemente en los párrafos siguientes:

IBM Solution for Hadoop Power Systems Edition

Se trata de una plataforma de Big Data integrada con una gran densidad de almacenamiento, optimizada para simplificar y acelerar la analítica de Big Data sin estructurar. Incluye todos los componentes necesarios para las aplicaciones Big Data, incluyendo servidores, redes, almacenamiento, sistema operativo, software de gestión, software compatible Hadoop y las bibliotecas de tiempo de ejecución. Esta plataforma aporta una aplicación de configuración optimizada. La configuración del cluster está cuidadosamente diseñada para optimizar el rendimiento de las aplicaciones y reducir el costo. El grupo está integrado con el Administrador de IBM Platform Cluster - Edición estándar (PCM SE) e IBM InfoSphere BigInsights Enterprise Edition (EE BigInsights) que tiene IBM Platform e IBM Sistema General Parallel File (GPFS™) como componentes integrados. Esta configuración optimizada permite a los usuarios mostrar resultados más rápidamente.

La plataforma aporta tecnología avanzada para el rendimiento y la robustez. Los componentes de hardware y software en esta infraestructura se pueden personalizar para permitir el mejor rendimiento o la mejor relación precio / rendimiento. La elección de una infraestructura que puede escalar para manejar las demandas de información y análisis es vital para el tratamiento continuo de grandes volúmenes de datos con alto rendimiento y alta fiabilidad.

IBM Solution for Analytics Power Systems Edition

Se trata de una herramienta flexible integrada para obtener información de forma más rápida con opciones para el despliegue de aplicaciones de analítica predictiva y business intelligence con aceleración de data warehouse en memoria.

El Big Data se ha convertido en recurso novedoso para la ventaja competitiva. Las organizaciones necesitan la capacidad de analizar de forma rápida y consistente de datos de múltiples fuentes para tomar decisiones que afectan el crecimiento y rentabilidad del negocio. Muchas de estas decisiones deben tomarse en tiempo real basado en la información más actualizada, en los momentos críticos que pueden afectar a las relaciones de los clientes o la eficiencia operativa.

Analytics proporciona una imagen más viva de la organización y de las fuerzas que lo afectan. Ayuda a actuar más rápidamente, reduciendo el tiempo entre obtener una visión de los datos y la adopción de medidas para la mejora de los resultados empresariales. La toma de decisiones tradicionalmente depende de los informes periódicos generados por los analistas de datos profesionales, lo que potencia la importancia de la analítica en todos los niveles de la organización. Se busca tomar mejores decisiones basadas en la información de hoy en comparación con el informe de ayer. El análisis se convierte en una parte fundamental de las operaciones del negocio, siendo la velocidad y disponibilidad un factor muy importante.

IBM Solution for Analytics Power Systems Edition ofrece un alto rendimiento y una herramienta fiable para la inteligencia empresarial y análisis predictivo. Se trata de una herramienta integrada flexible que ofrece opciones para pre-cargar y configurar uno o más aplicaciones analíticas de IBM, con una aceleración de almacenamiento de datos en un servidor basado en el procesador POWER8. Construido con el primer procesador diseñado para cargas de trabajo de datos, diseño de Power Systems combina la potencia de cálculo, el ancho de banda de memoria y E / S de manera que sean más fáciles de consumir y gestionar, a partir de una fuerte capacidad de recuperación, disponibilidad y seguridad.

La herramienta puede incluir una o más de las siguientes opciones:

- IBM Cognos Business Intelligence

- IBM SPSS Modeler, Collaboration and Decisión Support, and Analytical Decisión Management (ADM)
- IBM DB2 Advanced Workgroup Edition or DB2 Advanced Enterprise Edition
- IBM InfoSphere Information Server (IIS) DataStage

La opción *Cognos Business Intelligence* está diseñada para ayudar a los usuarios de negocios, ejecutivos y analistas en una organización a comprender el negocio y tomar decisiones más inteligentes. La herramienta ofrece una gama completa de capacidades de BI, incluyendo informes, análisis, cuadros de mando, cuadros de mandos de BI móvil, y otras opciones.

Las capacidades de análisis predictivo de IBM Solution for Analytics Power Systems Edition se centran en las opciones de IBM SPSS. SPSS combina el análisis predictivo con reglas de negocio que permiten a las organizaciones implementar las acciones de impacto óptimo. *SPSS Modeler*, *SPSS Collaboration and Deployment Services* y *Analytical Decisión Management* juntos contribuyen a optimizar las decisiones en las organizaciones.

Una o ambas de estas aplicaciones de análisis se puede utilizar en conjunción con DB2 apoyado con capacidades de aceleración BLU que se incluyen con *IBM DB2 Advanced Workgroup Edition* or *DB2 Advanced Enterprise Edition* para un mejor rendimiento. BLU Acceleration ofrece aceleración dinámica en memoria de almacenamiento de datos, que permite a las organizaciones utilizar tanto el almacenamiento de datos basado en columnas o en filas y de forma simultánea.

IBM InfoSphere DataStage es también una opción de pre-carga para proporcionar capacidades de extracción, transformación y carga (ETL) de un almacén de datos existente en el almacén de Aceleración BLU.

IBM BLU Acceleration Solution Power Systems Edition

Se trata de la herramienta para el tratamiento In-Memory de IBM. Es una herramienta dinámica en memoria por columnas para acelerar la obtención de información procedente de grandes volúmenes de datos para

aplicaciones de data warehouse y analítica.

Permite consultas analíticas más rápidas y menos complejas y la presentación de informes utilizando tecnologías de columnas dinámicas en memoria para la inteligencia empresarial, análisis predictivo y aplicaciones de almacenamiento de datos.

IBM AIX Solution Editions para Cognos y SPSS

Hoy en día es un hecho que el volumen y la velocidad de los datos que están disponibles simplemente son demasiados para que los usuarios de negocios puedan recopilarlos y procesarlos. A menudo es demasiado difícil y lleva mucho tiempo encontrar la información adecuada, organizarla y comparar las posibles opciones cuando se trata de tomar decisiones.

Para hacer frente a esto, las empresas están recurriendo a la analítica para ayudar a derivar percepciones de esta enorme cantidad de datos para impulsar la competitividad, el crecimiento de los ingresos y la eficiencia operativa. Como la toma de decisiones basada en los datos se convierte en omnipresente en todas las organizaciones, se hace misión crítica para el éxito del negocio, y requiere una infraestructura de alta disponibilidad y capacidad de recuperación. La velocidad y el acceso también son importantes, ya que muchas de estas decisiones son sensibles al tiempo que requiere un análisis rápido de los datos más actuales, entregados directamente al usuario de negocios en el punto de impacto. Para tratar esta problemática IBM ofrece aplicaciones preconfiguradas de análisis como AIX Solution Edition para Cognos y AIX Solution Edition for Analytical Decision Management (SPSS).

IBM AIX Solution Edition para Cognos ofrece capacidades de Cognos Business Intelligence en la mejor plataforma de la industria para las cargas de trabajo analíticas de cálculo intensivo, Power Systems. Esta herramienta está diseñada para satisfacer a todo tipo de usuarios y estilos de trabajo, lo que permite que todos en la organización puedan explorar libremente los datos, analizar los hechos clave, colaborar y actuar con confianza. Para una implementación más rápida, IBM AIX Solution Edition para Cognos viene con el sistema operativo AIX, PowerVM virtualization, WebSphere Application Server, DB2 Enterprise y Cognos

Business Intelligence preinstalado.

La arquitectura de POWER 7 permite el análisis de grandes cantidades de datos, y la entrega de resultados a los usuarios en toda la organización de forma rápida, segura y rentable. AIX Solution Edition para Cognos ofrece un sistema que crece y se adapta con el negocio. Las organizaciones pueden comenzar por la aplicación de las capacidades más críticas, como cuadros de mando, informes clave y añadir funcionalidades como móvil y colaboración realizando el análisis en un solo servidor. Esto puede reducir la proliferación de servidores, así como los costos de licencias y gestión de software asociados.

Los informes mensuales o diarios de acceso en tiempo real por los usuarios en toda la organización y el rendimiento de análisis se vuelven cada vez más importantes. Los cubos dinámicos de Cognos permiten situar en memoria los datos y el conocimiento global con el fin de lograr un análisis interactivo de alto rendimiento y presentación de informes sobre terabytes de datos de almacén. Para maximizar el rendimiento, Cognos utiliza tablas de resumen y bases de datos en memoria, reutilizando los resultados de la caché del procesador, que funciona de manera más eficaz cuando todos los procesadores están en un solo sistema.

Por su parte, *AIX Solution Edition for Analytical Decision Management (SPSS)* permite el análisis predictivo y ayuda a las organizaciones a utilizar sus datos para tomar mejores decisiones por lo que les permite llegar a conclusiones fiables, basados en datos sobre las condiciones actuales y acontecimientos futuros. Mediante la implementación de análisis predictivo, las organizaciones están abordando sus problemas de negocio de forma proactiva para obtener los mejores resultados.

Con AIX Solution Edition for Analytical Decision Management (SPSS), los análisis predictivos y las capacidades de gestión de decisiones de IBM SPSS Modeler, Collaboration & Decision Support y Analytical Decision Management junto con Power Systems constituyen una de las mejores plataformas de la industria para el análisis intensivo de grandes volúmenes de datos. De esta forma se pueden determinar las mejores acciones en tiempo real. Al combinar e integrar el análisis predictivo y las técnicas de optimización en los sistemas de la organización, AIX Solution Edition for Analytical Decision Management ofrece análisis optimizado de la información sin importar su tamaño y fuente de origen.

IBM PureData System for Operational Analytics (US)

Esta herramienta permite el despliegue, optimización y gestión de cargas de trabajo de datos intensivos para analíticas operacionales con un sistema experto integrado.

Este sistema PureData permite a los departamentos implementar, optimizar y gestionar las cargas de trabajo intensivas de datos para análisis de operaciones. Se ofrece un valor excepcional de dos maneras:

- Incorporando conocimientos de análisis de operaciones, sobre la base de años de experiencia de IBM y las mejores prácticas de miles de compromisos con los clientes. De esta forma se proporciona una herramienta completa que puede ofrecer un valor fuera añadido.
- Proporcionando integración del diseño de software, servidores, almacenamiento y redes para obtener resultados en los sistemas optimizados de fábrica y también eficiencia y alto rendimiento.

Big Data Solution with InfoSphere BigInsights and Streams

Es posible analizar datos a escala con Apache Hadoop, InfoSphere BigInsights e InfoSphere Streams en Power Systems. Se destacan las siguientes capacidades:

- Aumentar la velocidad y precisión del análisis utilizando IBM InfoSphere BigInsights y InfoSphere Streams sobre Linux en Power Systems para analizar los datos estructurados y no estructurados, según sea necesario.
- Tomar decisiones de negocio en el punto de Impacto mediante el análisis de datos en reposo con IBM InfoSphere BigInsights y datos en movimiento con IBM InfoSphere Streams optimizados en Power Systems.
- Obtener alta calidad y servicios de análisis de grandes cantidades de Información más eficientes construidos sobre una arquitectura POWER® fiable y segura.

Los datos pueden explotarse provenientes de una variedad de fuentes,

tales como sensores de clima, sitios de medios sociales, fotos digitales, vídeos en línea, registros de transacciones, registros de datos de llamadas de teléfonos móviles, etc.

Las organizaciones de hoy necesitan aprovechar la explosión de datos para crear ventajas competitivas. Aprovechando datos tradicionales estructurados solo en el modo por lotes ya no es suficiente. Las organizaciones deben encontrar una manera de utilizar más datos para tomar mejores decisiones. Las herramientas diseñadas para grandes volúmenes de datos se extienden más allá de fuentes transaccionales tradicionales de datos para generar una visión a través de fuentes no estructuradas, semiestructuradas y streaming más rápido y con confianza.

Los datos seguirán creciendo a tasas astronómicas. Los líderes tienen que encontrar puntos de vista críticos y conducir las decisiones oportunas para alcanzar los objetivos de negocio.

El análisis de Big Data Solutions en IBM Power Systems puede ayudar a las empresas a obtener nuevos conocimientos con ofertas escalables de gran alcance. La combinación del software de IBM InfoSphere BigInsights para analizar los datos en reposo y software InfoSphere Streams para analizar los datos en movimiento proporciona el conjunto más robusto de capacidades analíticas para grandes volúmenes de datos.

IBM i para Business Intelligence (US)

Se trata de una herramienta que aumenta la generación de valor con facilidad para implementar una herramienta empaquetada que convierta la información en conocimiento procesable.

IBM i para Business Intelligence es una herramienta empaquetada que es fácil de ordenar y fácil de implementar. Se trata de una opción perfecta para la mejora de la capacidad para analizar los datos convirtiéndolos en información que puede ayudar a transformar su negocio. Es posible obtener resultados inmediatos con la información operativa y la transición a un almacén de datos robusto, basado en sus demandas de negocio.

Esta herramienta combina las ventajas de Power Systems, IBM i, DB2 para i, DB2 Web Query, y el software de transporte de datos para ofrecer una plataforma integrada que contiene los datos extraídos y transportados provenientes de sus sistemas de producción. IBM i para Business

Intelligence proporciona todo lo necesario para la puesta en marcha de su entorno de informes. Se trata de una plataforma de crecimiento ampliable para aumentar sus capacidades de análisis de datos mediante la transformación de los datos de origen mediante el aprovechamiento de las herramientas y servicios de ETL adicional para construir datos almacén.

IBM DB2 Web Query for i

Es una herramienta segura que todos los responsables de la toma de decisiones en la organización pueden encontrar, analizar y compartir la información necesaria para mejorar y acelerar su trabajo. Presenta las siguientes capacidades:

- Desplegar de modo simple la inteligencia empresarial y análisis.
- Modernizar consultas e informes de bases de datos heredadas.
- Mejorar el rendimiento con la optimización de consultas avanzadas
Reducir costes mediante la utilización adecuada de IBM i.
- Gestionar el riesgo mediante el aprovechamiento de la seguridad de IBM i.
- Acelerar el tiempo de desempeño con soluciones empaquetadas.

Nuestro planeta se está volviendo más inteligente, y con este cambio se produce una explosión de información. Para aprovechar ideas de esa información, un alto porcentaje de las organizaciones tienen planes que incluyen inteligencia empresarial y análisis para incrementar la competitividad de las empresas. Hoy en día, el análisis ha evolucionado desde una iniciativa de negocio a un imperativo de negocios, y desde las organizaciones individuales hasta las industrias enteras. Las empresas que adoptan la analítica están ganando una ventaja competitiva y superan a sus pares.

Business Intelligence (BI) es un término amplio en relación con las aplicaciones diseñadas para analizar los datos con fines de comprensión y que actúan sobre las métricas clave que impulsan la rentabilidad de una empresa. La clave para el análisis de los datos es que se está proporcionando acceso rápido y fácil a ellos al mismo tiempo que se tratan en los formatos o herramientas que mejor se adapten a las necesidades del usuario final. En el centro de cualquier solución de inteligencia

empresarial están presentes el análisis, las consultas del usuario final y las herramientas de informes que proporcionan un acceso intuitivo a los datos.

CLOUD COMPUTING EN POWER SYSTEMS

La infraestructura de TI es importante. Las expectativas en el tiempo de respuesta están disminuyendo, las demandas crecen y las organizaciones deben ser capaces de ofrecer una cantidad cada vez mayor de información a empleados, clientes y socios a través de navegadores, aplicaciones y servicios distribuidos en diversos dispositivos y canales. Una solución cloud eficiente y eficaz necesita contar con una base abierta igual de eficiente y eficaz.

IBM Power Systems facilita la transformación de la organización y un rendimiento óptimo del negocio mediante información de valor basada en datos, distribuida mediante una plataforma abierta y flexible. Power ofrece simplicidad y diversas opciones a sus clientes para gestionar y prestar servicios críticos de TI en las propias instalaciones o mediante entornos virtualizados de forma eficiente, segura y rentable.

Las herramientas cloud de IBM Power Systems permiten crear una base de virtualización sólida y segura que permita un despliegue eficiente en la nube.

Es conveniente empezar con una solución cloud de entrada, fácil de desplegar y utilizar, basada en estándares abiertos, que ofrece las características y la flexibilidad necesarias tanto para ahora como para el futuro. A continuación se tratará de descubrir más sobre las prestaciones avanzadas de cloud y saber si son necesarias.

Las características principales de las soluciones de cloud IBM Power Systems son las siguientes:

- La virtualización Integrada en la plataforma IBM Power Systems, no anclada, garantiza la utilización y la gestión óptimas de recursos, seguridad y calidad de servicio empresarial (QoS).
- Las soluciones de cloud y gestión de la virtualización abiertas y ampliables ofrecen una base escalable, adaptable y sólida para las clouds públicas y privadas.
- Optimización para gestionar la distribución de servicios de cloud de

Big Data y análisis a nivel empresarial con un rendimiento y una economía excepcionales.

Dado que el mundo va evolucionando y la TI desempeña un papel cada vez más importante, las empresas y los gobiernos buscan la manera de responder de una forma más rápida a la cambiante demanda empresarial transformando el modo en el que ofrecen sus prestaciones de TI a través de una mayor eficiencia operativa. La cloud computing puede simplificar la distribución de servicios y mejorar la economía de TI.

Los modelos de implementación de Infraestructuras de TI pueden suponer un reto para las organizaciones de TI que desean cambiar a métodos más ágiles como:

- Afrontar retos, tales como crecimiento rápido de los datos, cumplimiento normativo, Integridad de la Información y preocupaciones relacionadas con la seguridad, a la vez que trata de controlar el aumento constante de los costes de TI.
- Ser Inflexibles ante cambios rápidos e Imprevistos de los mercados, demandas de servicios y expectativas de las partes Interesadas.
- Incluir islas estáticas de recursos informáticos que dan como resultado deficiencias y activos infrutilizados.

Entre las soluciones de clud en Power Systems tenemos las siguientes:

Virtualization Foundation Solutions

Basado en el procesador POWER, los servidores Power Systems presentan la arquitectura para lograr el máximo rendimiento y eficiencia, tanto para el sistema como para sus máquinas virtuales. La asignación de recursos basada en la carga de trabajo inteligente con conmutación hilo procesador dinámico y expansión de memoria lógica ofrecen un rendimiento óptimo para los servicios críticos de la nube.

PowerVM hace un uso eficiente de los recursos del sistema e impone un impacto insignificante en el rendimiento porque PowerVM está integrado directamente en el firmware de todos los sistemas de energía, en lugar de x86 basados en productos de virtualización que son típicamente de software add-ons de terceros.

PowerKVM presenta una opción para los clientes que buscan un bajo costo, la solución de virtualización de código abierto para Linux en Power Systems se basa en la arquitectura POWER8. PowerKVM proporciona soporte para Red Hat Enterprise Linux, SUSE Linux Enterprise Server y Ubuntu y está optimizado para la plataforma Power Systems con soporte para el procesador y la memoria compartida para utilizaciones superiores, además de dinámica y migración VM.

La gestión avanzada de virtualización está disponible por PowerVC (Centro de virtualización). Construido sobre OpenStack, permite la infraestructura cloud Power Systems para conectarse en una amplia gama de soluciones de gestión. También permite la máxima utilización de los sistemas de energía con sus algoritmos de equivalencia universitaria para máquinas virtuales con interfaz fácil de usar que ayuda a los administradores.

Power Systems ofrece Capacity on Demand, que permite la activación instantánea y seguimiento de la utilización de memoria adicional y recursos informáticos para manejar picos temporales en el uso del sistema y optimizar así el trabajo en la nube.

Con las tecnologías de gestión de plataformas de Power Systems, las empresas tienen las herramientas necesarias para implementar de forma automática, optimizar y mantener estos sistemas al máximo la eficacia, la eficiencia energética y el control de costes.

IBM SmartCloud Entry for Power Systems

Construida sobre OpenStack, es una herramienta modular, altamente flexible y fácil de usar, diseñada para ofrecer servicios en la nube para las nubes privadas o públicas, a través de una selección de la infraestructura heterogénea incluyendo Power Systems y x86. Permite reducir las tareas administrativas manuales, mejorar la productividad y reducir los errores con la gestión de procesos automatizada y aprovisionamiento de Imágenes estandarizadas. Proporciona administración de la nube simplificada a través de una interfaz intuitiva para la gestión de proyectos y usuarios, así como el seguimiento de las cargas de trabajo heterogéneas y recursos de la nube. Permiten mantener la supervisión y el desempeño óptimo de su nube con dosificación de la carga de trabajo.

IBM Power Systems Solution Edition for Cloud

IBM Power Systems Solution Edition for Cloud es un sistema configurado de fábrica personalizable, con la red pre-construidos, cómputo, almacenamiento y software, así que usted puede conseguir su nube en funcionamiento y agregando valor rápidamente. Los clientes pueden elegir entre una variedad de configuraciones de sistema para adaptarse a su negocio y las necesidades de carga de trabajo: desde la entrada a la empresa, la solución de edición para la nube que sea fácil de ordenar el tamaño justo y la escala para cualquier nube basado PowerVM y ayuda a los clientes a lograr implementaciones de servicios cloud con las cualidades del sistema de la empresa de servicio, la más alta escala, y la economía de TI superior.

IBM Power Systems Solution Edition for Scale Out Cloud

Construida con la última tecnología de procesador POWER8 de IBM, esta herramienta permite el escalamiento en la nube de forma inteligente, con menos requisitos de hardware, energía y enfriamiento y una mejor economía que aprovechan más de dos veces el ancho de banda de las generaciones anteriores. Esta herramienta también es una ventaja para aquellos servicios en la nube que exigen potencia informática excepcional y ancho de banda de memoria, como Big Data y Analytics.

Herramientas avanzadas en la nube

IBM ofrece soluciones para permitir capacidades de nube avanzadas tales como el aprovisionamiento rápido y escalable y la optimización para los proveedores de servicios, flujo de trabajo robusto, gestión del ciclo de vida de la virtualización y otros servicios con SmartCloud Provisioning y SmartCloud Orquestation. Estas soluciones, construidas en soluciones de código abierto, incluyendo OpenStack y Chef, están diseñadas para reducir el riesgo asociado con la integración de software y acelerar la entrega de capacidades de nube de computación avanzada.

IBM SPSS MODELER

IBM SPSS Modeler es una herramienta integrada de Big Data, business intelligence y minería de datos que incluye diversas fuentes de datos (ASCII, XLS, ODBC, etc.), una interfaz visual basada en procesos/flujos de datos (*streams*), distintas herramientas de minería de datos (correlación, reglas de asociación, regresión, segmentación, clasificación, redes neuronales, reglas y árboles de decisión, etc.), manipulación de datos (*pick & mix*, muestreo, combinación y separación, etc.), combinación de modelos, visualización de datos, exportación de modelos a distintos lenguajes (C, SPSS, SAS, etc.), exportación de datos integrada a otros programas (XLS) y generación de informes.

El entorno de IBM SPSS Modeler está basado en nodos que se van disponiendo y conectando para formar un flujo, o *stream*, traducido por Modeler también como “ruta”. Los *streams* pueden almacenarse en ficheros separados o en proyectos que engloban a varios de ellos que se pueden cargar, guardar, modificar, reejecutar o reorganizar utilizando las opciones del menú *Archivo* de la pantalla de entrada de IBM SPSS Modeler (Figura 3-1) y que son independientes de las fuentes de datos.

En la Figura 3-2 se muestra la pantalla de carga de la ruta *druglearn.str* obtenida al hacer clic en la opción *Abrir* del menú *Archivo*. En la Figura 3-3 se muestra la ruta *druglearn.str* con seis nodos interconectados ya cargada en Modeler.



Figura 3-1

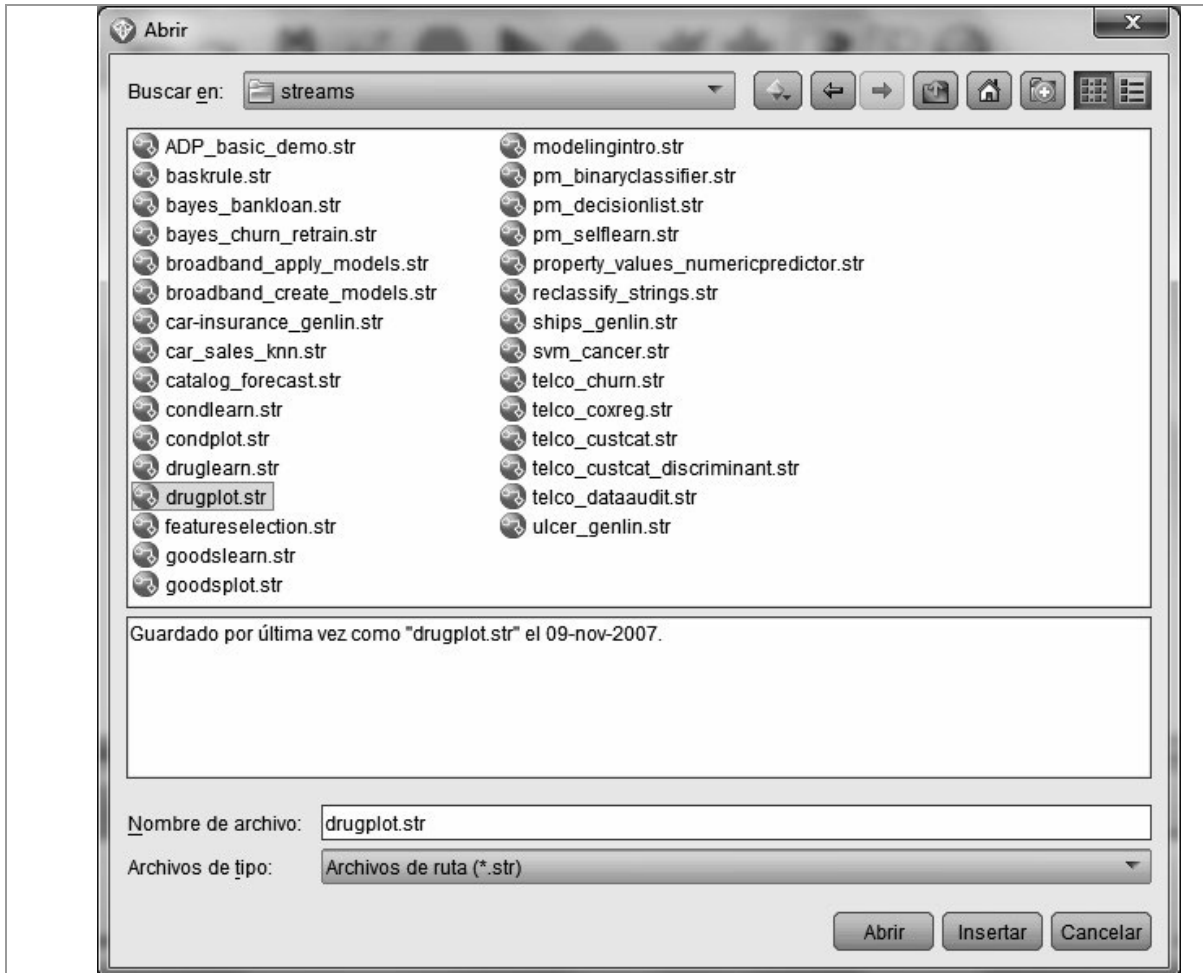


Figura 3-2

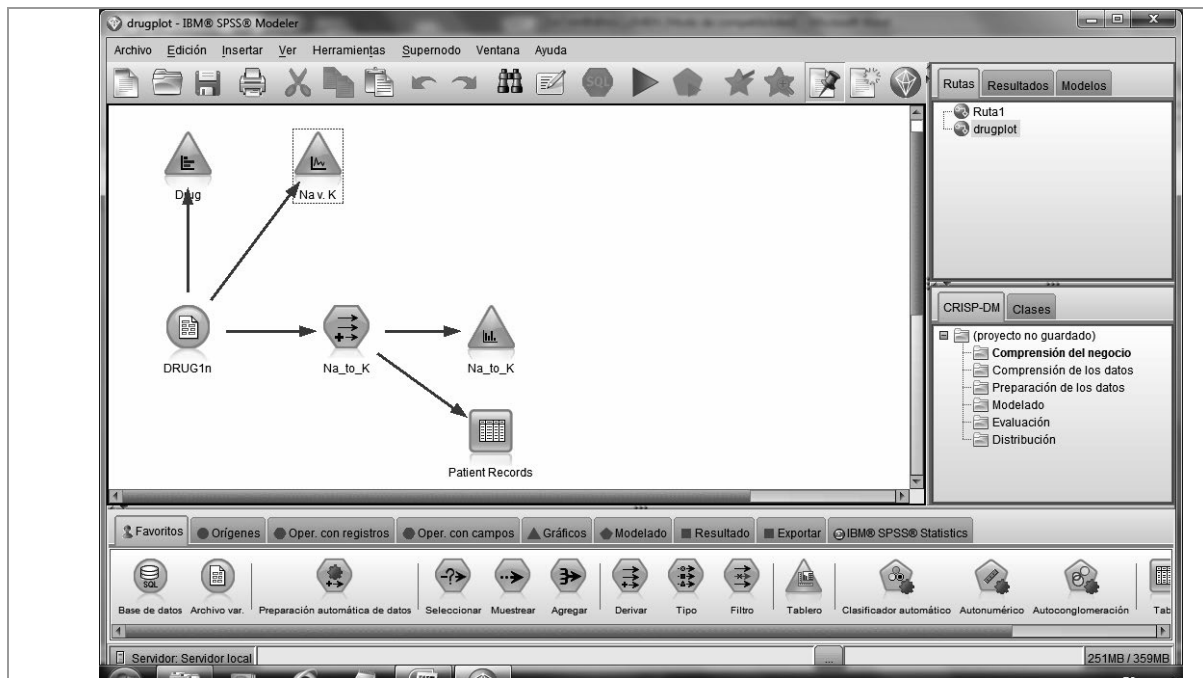


Figura 3-3

Como se puede ver en la parte inferior de la Figura 3-3 (Figura 3-4), Modeler presenta varias paletas que clasifican los nodos en seis categorías:

- *Favoritos*: nodos más utilizados.
- *Orígenes*: nodos para obtener los datos de trabajo (fuentes de datos).
- *Oper. con registros*: operadores para modificar o combinar registros (filas) de distintas fuentes. Es decir, selecciones y combinaciones.
- *Oper. con campos*: operadores para modificar o combinar campos (columnas).
- *Gráficos*: gráficas.
- *Modelado*: tipos de modelos/patronos que puede generar Modeler.
- *Resultado*: presentación de tablas, análisis de modelos, estadísticas, exportación de datos.
- *Exportar*, exportación de información a otros formatos y aplicaciones.
- *IBM SPSS Statistics*: conexión con otros procedimientos de IBM SPSS Statistics.



Figura 3-4

En la parte superior derecha de la pantalla se encuentran las paletas *Rutas*, *Resultados* y *Modelos generados* (Figura 3-5) que muestra las rutas, resultados y modelos que actualmente se están elaborando.



Figura 3-5

Para ejecutar una ruta se hace clic con el botón derecho del ratón sobre el nodo objetivo y se hace clic en *Ejecutar* en el menú emergente resultante (Figura 3-6).

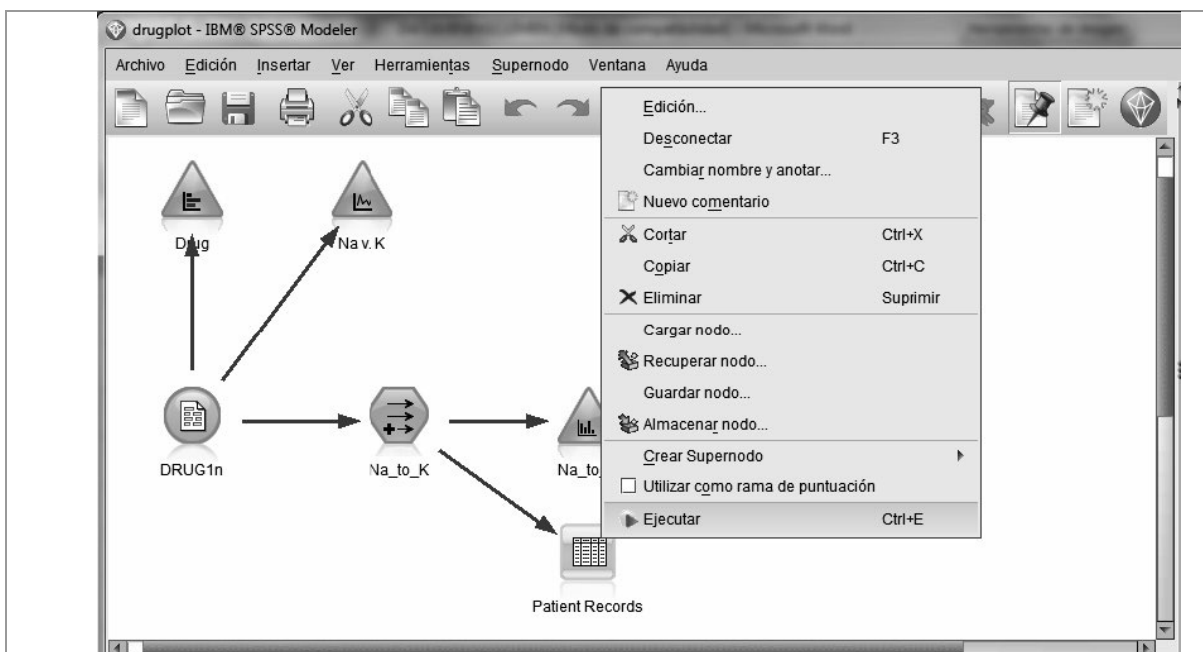


Figura 3-6

Usando el ratón

Algunas de las operaciones en Modeler se ven facilitadas con un ratón de tres botones. El tercer botón suele utilizarse a menudo para realizar conexiones entre los diferentes nodos de una ruta. Si el ratón no tiene el tercer botón, puede emularse su efecto presionando simultáneamente los dos botones.

El clic simple con los botones izquierdo y derecho del ratón permite seleccionar opciones de menús o abrir menús contextuales.

El doble clic con el botón izquierdo del ratón permite situar nodos en una ruta y editar nodos existentes.

El clic simple con el tercer botón del ratón (equivalente al clic simultáneo de los dos botones cuando no existe el tercero) seguido de arrastre, permite conectar nodos en una ruta. El doble clic en el tercer botón permite desconectar nodos.

Ayuda en Modeler

La opción Ayuda del menú de Modeler (Figura 3-7) permite varios caminos para acceder a su contenido. La subopción Temas de ayuda permite acceder a toda la ayuda de Modeler por capítulos (Figura 3-8). La subopción Ejemplos de aplicaciones da acceso a un tutorial Interactivo sencillo con ejemplos de aplicaciones (Figura 3-9).

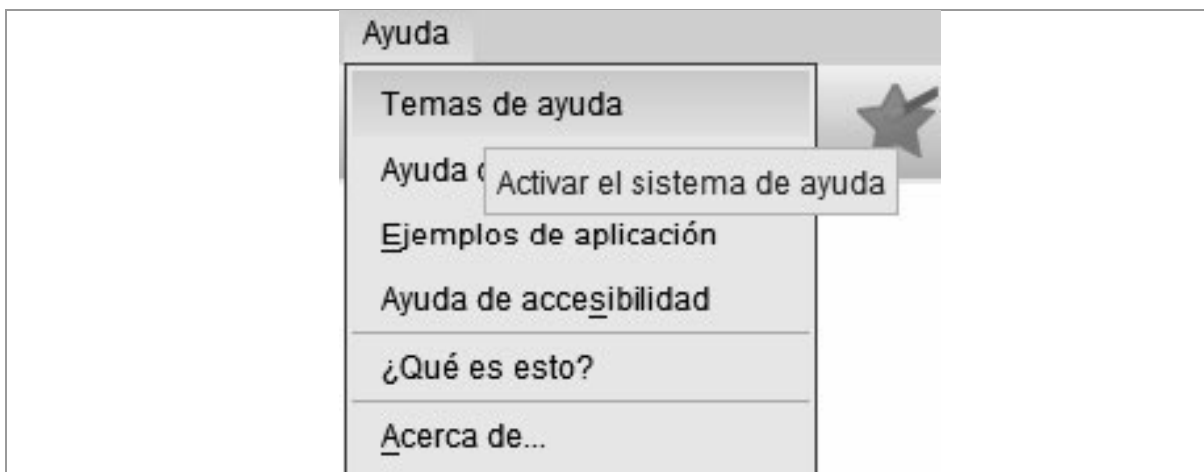


Figura 3-7

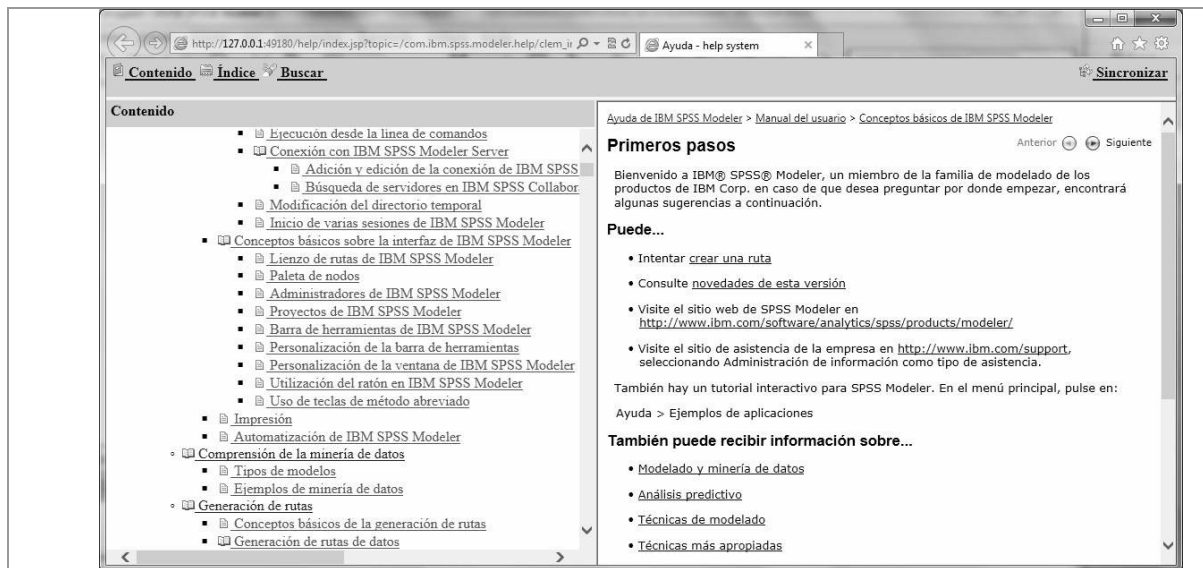


Figura 3-8

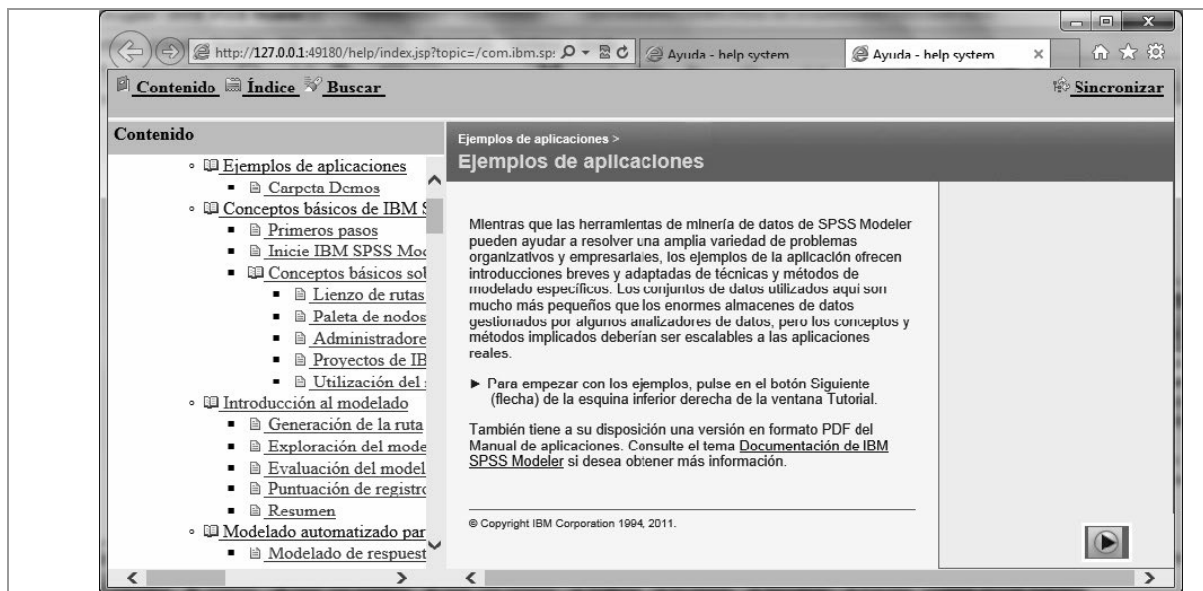


Figura 3-9

El menú Herramientas de Modeler

En el menú de Modeler aparece la opción *Herramientas* (Figuras 3-10 a 3-13) que nos va a permitir configurar el Inicio de sesión del servidor (Figura 3-14), las bases de datos de trabajo (Figura 3-15), el repositorio (Figura 3-16), la contraseña, las opciones del sistema (Figura 3-17), usuarios (Figura 3-18) y complementos (Figura 3-19), las propiedades de ruta (Figura 3-20), los parámetros de sesión (Figura 3-21), paletas (Figura 3-22), etc.

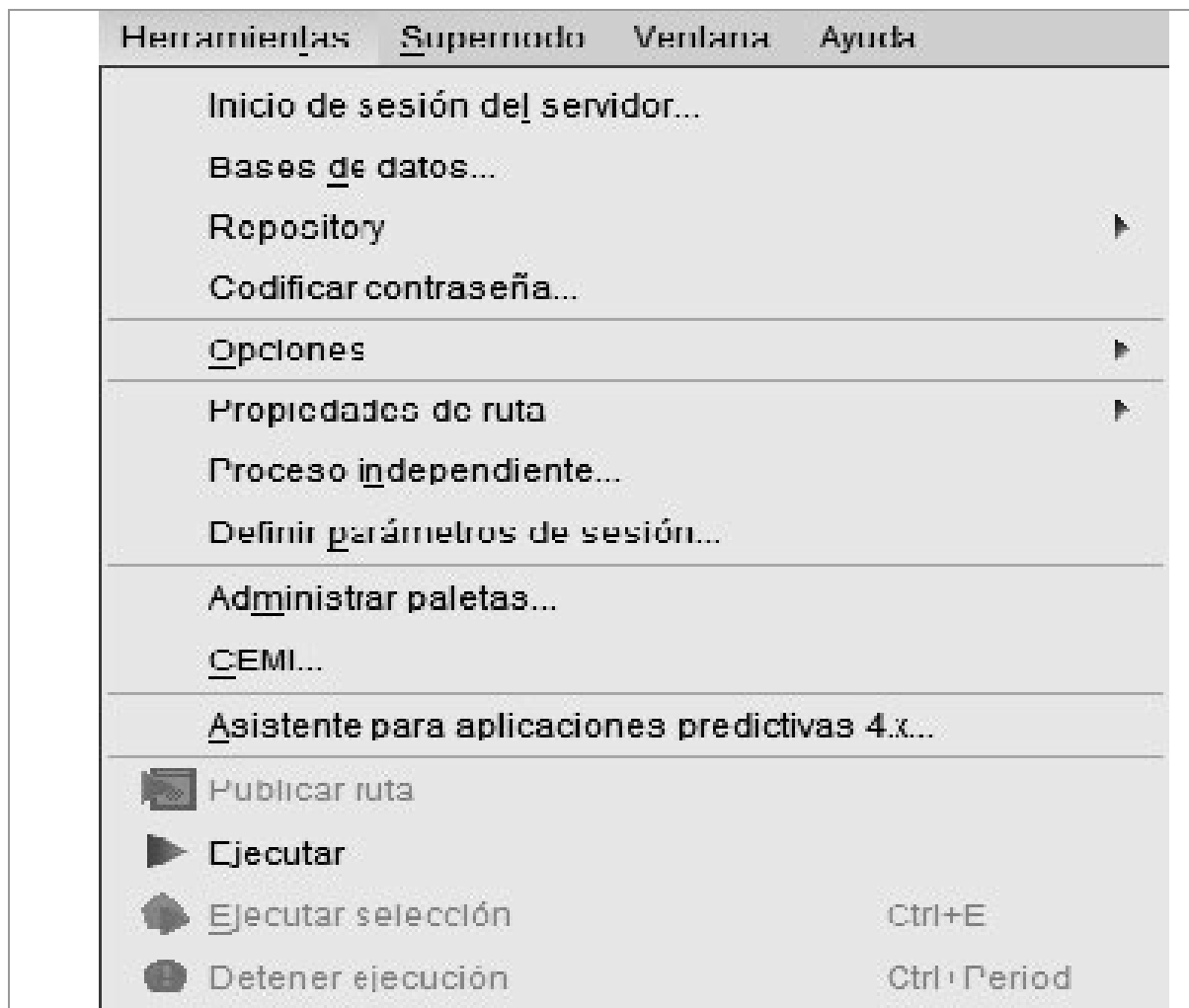


Figura 3-10

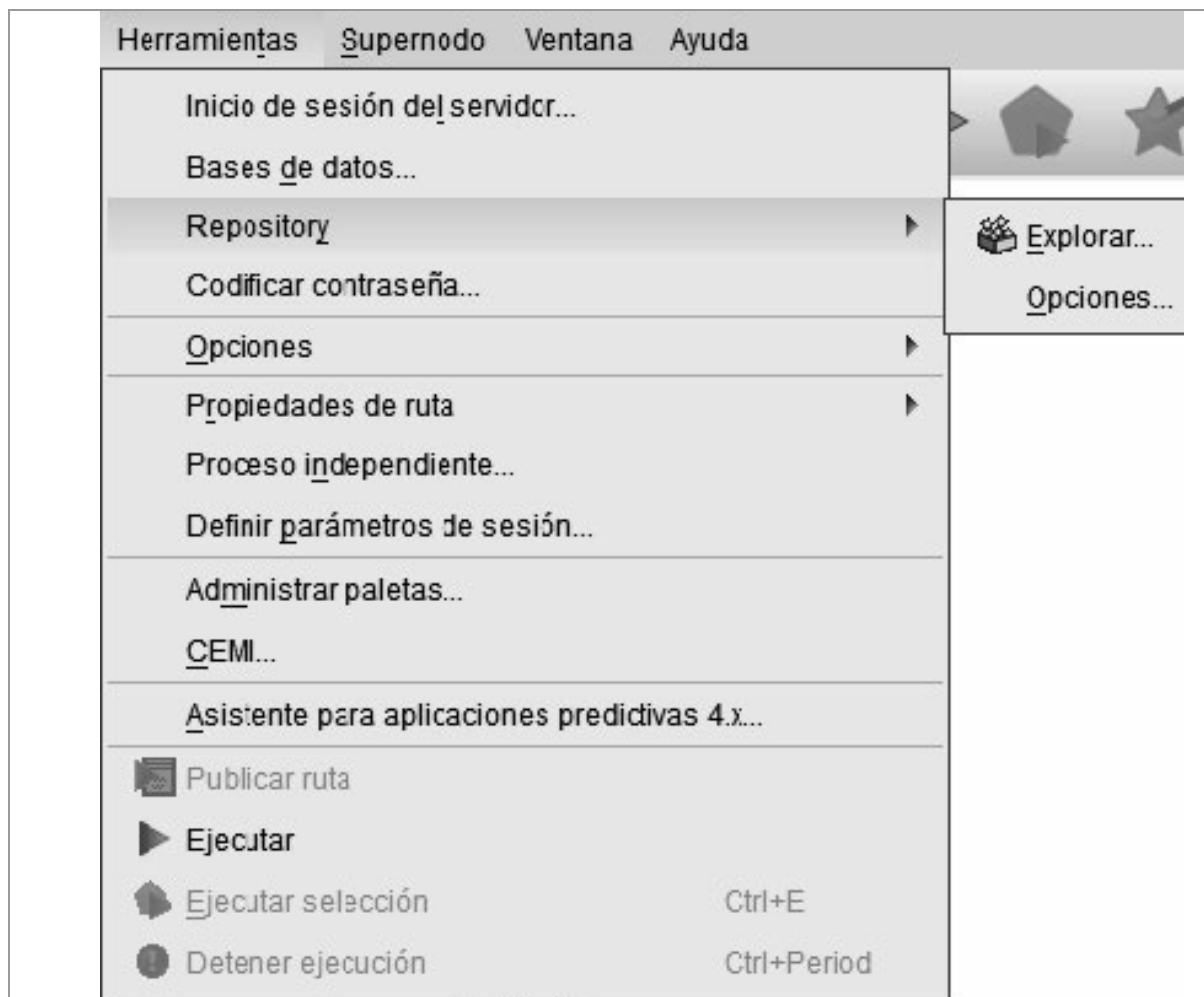


Figura 3-11

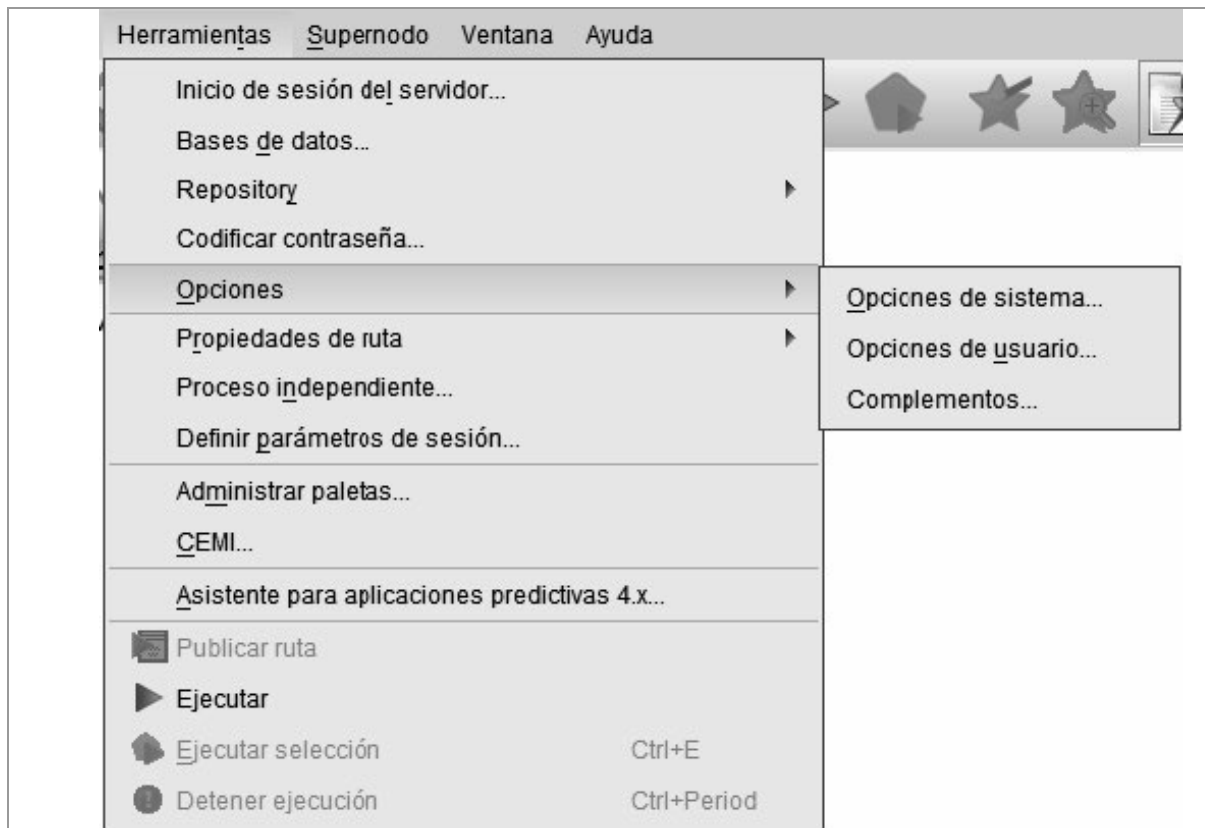


Figura 3-12

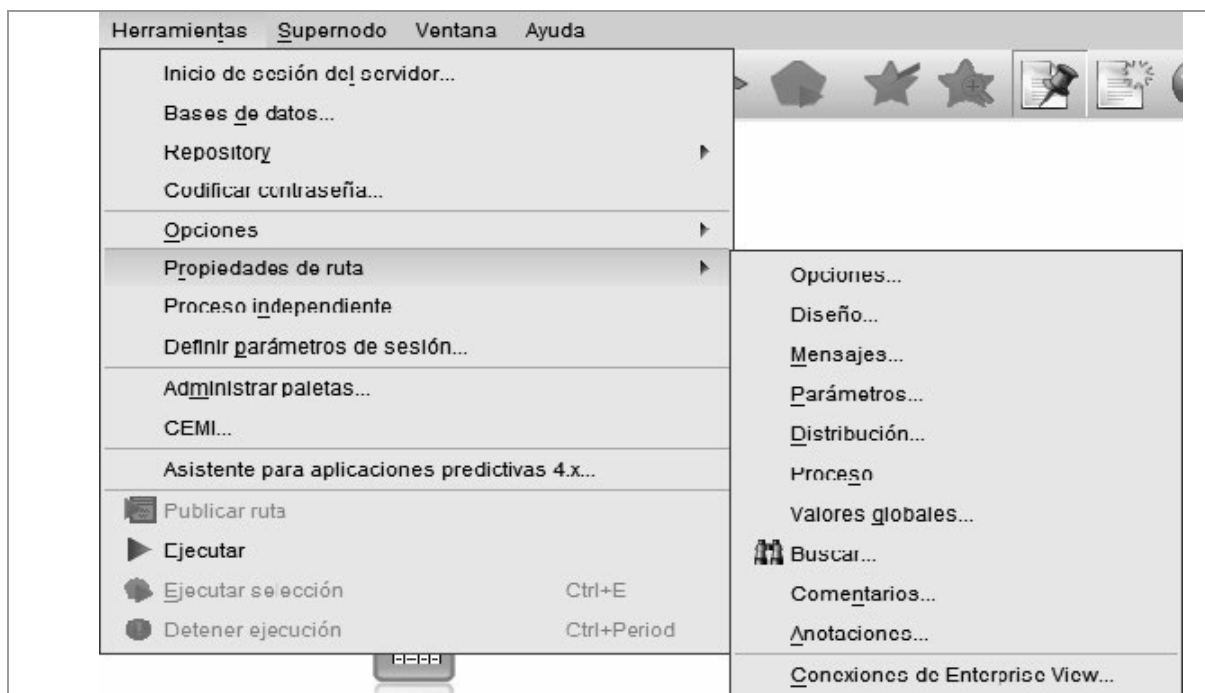


Figura 3-13

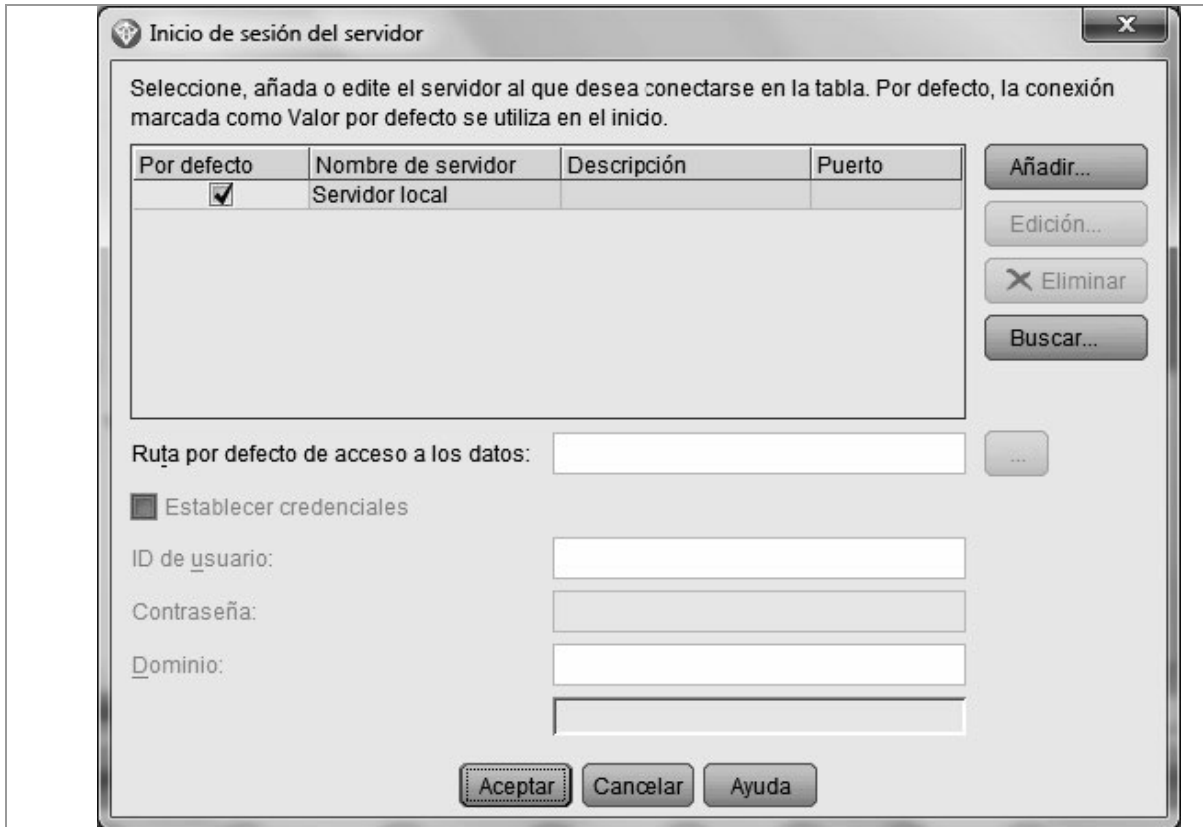


Figura 3-14

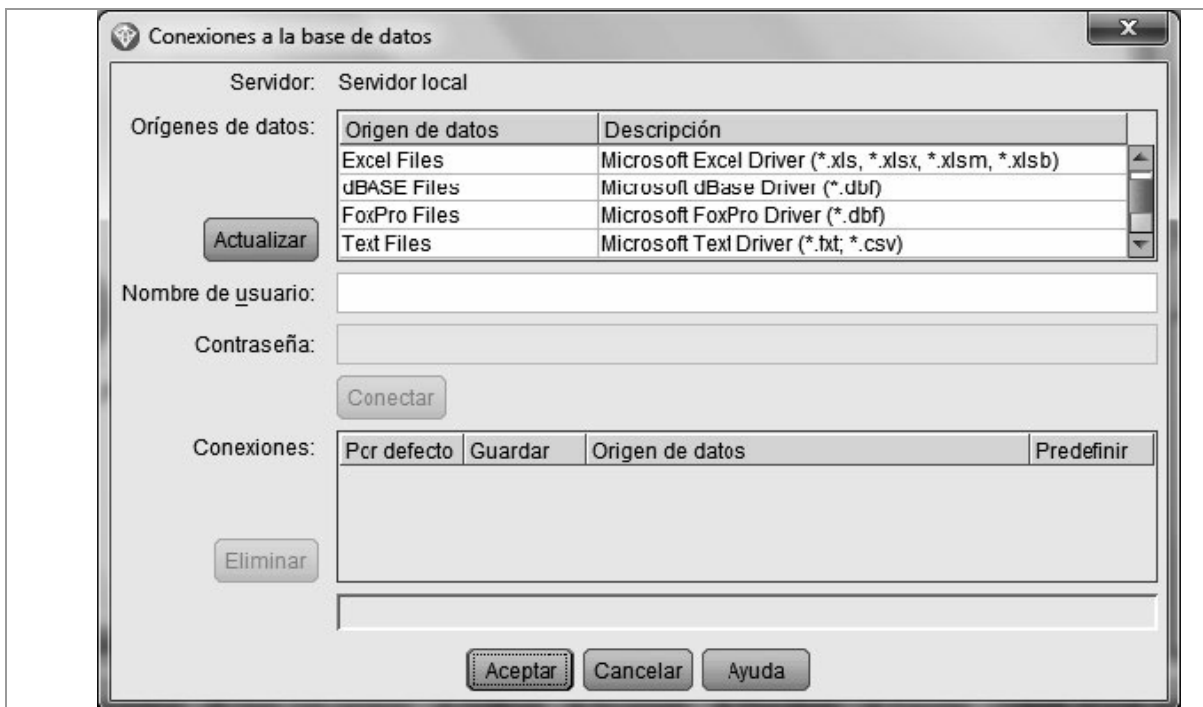


Figura 3-15

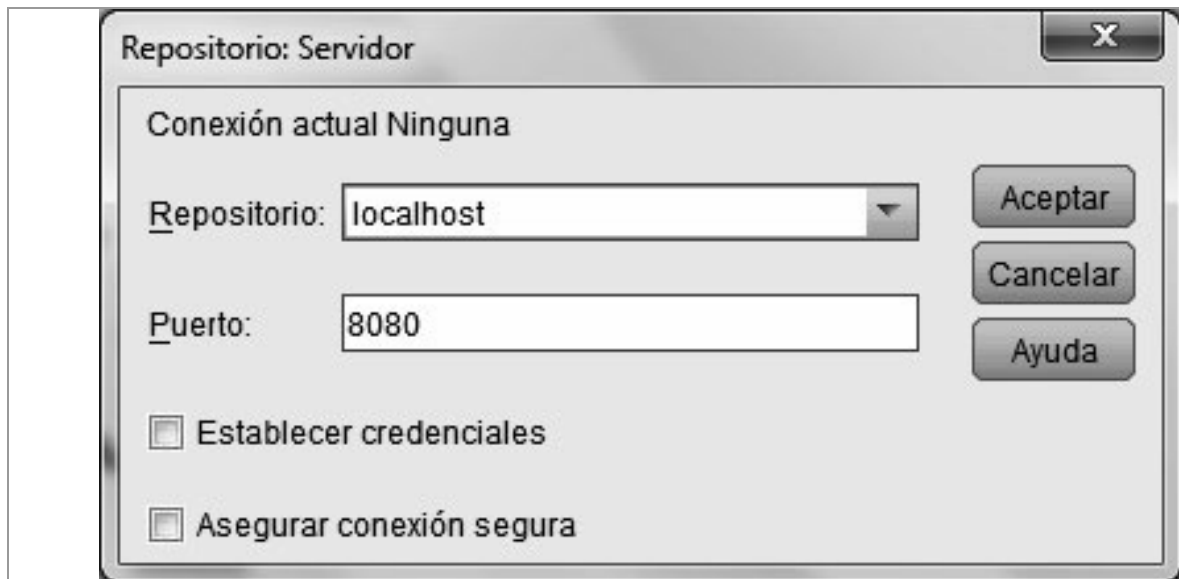


Figura 3-16

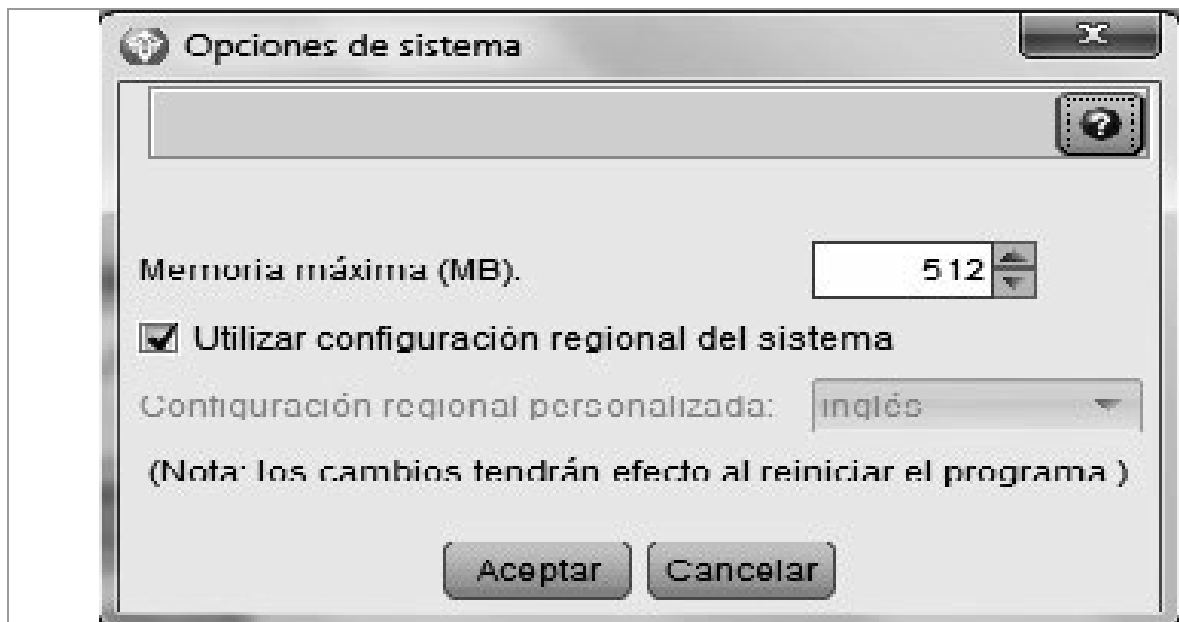


Figura 3-17



Figura 3-18



Figura 3-19

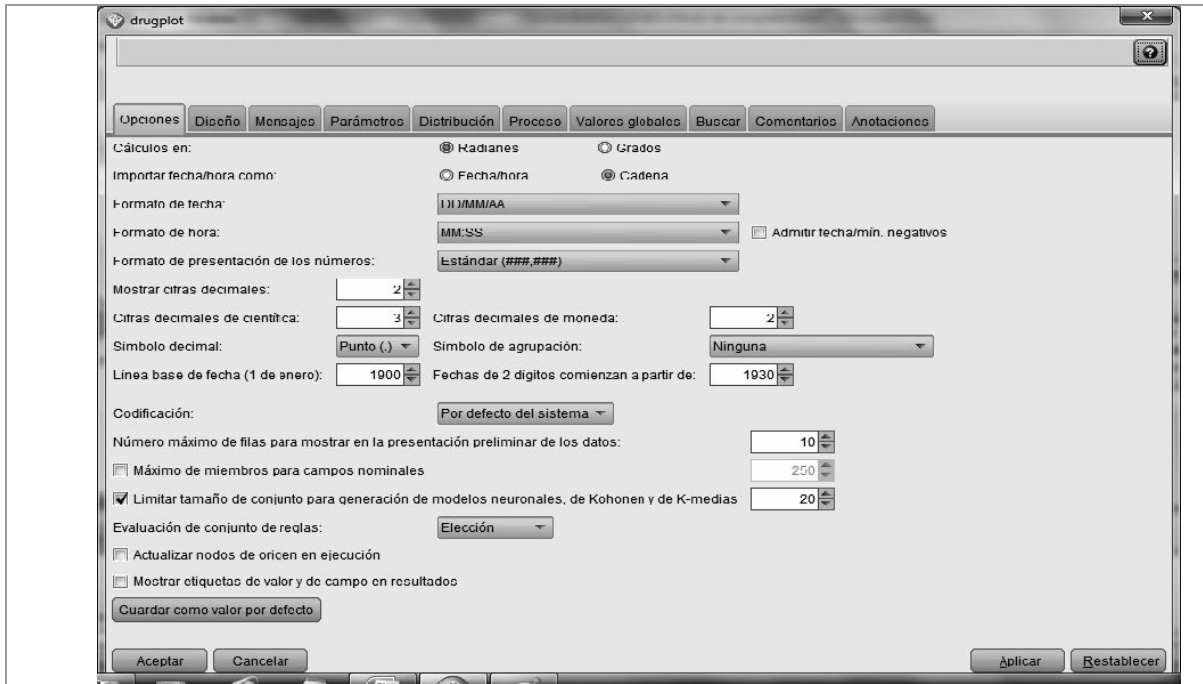


Figura 3-20

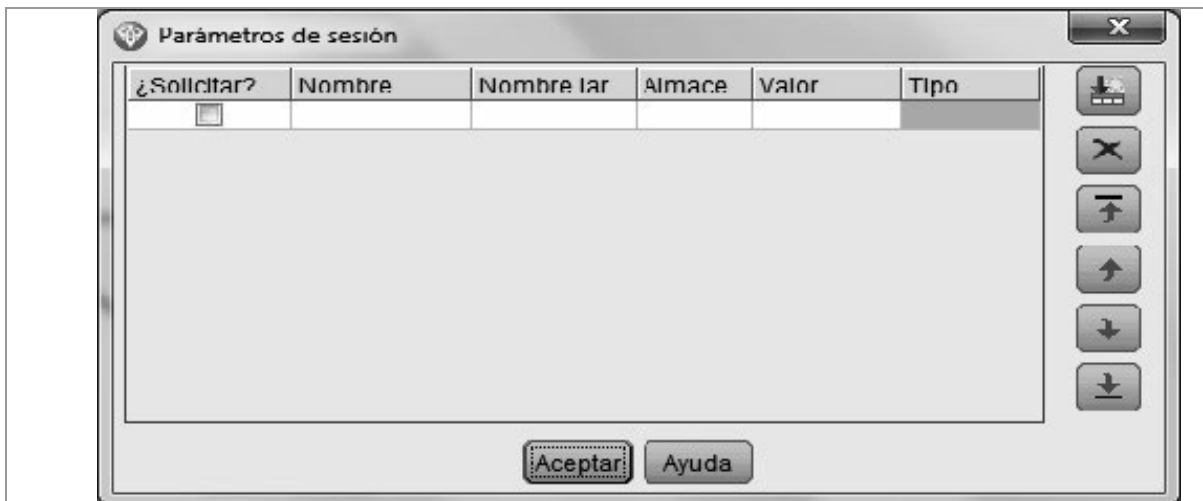


Figura 3-21

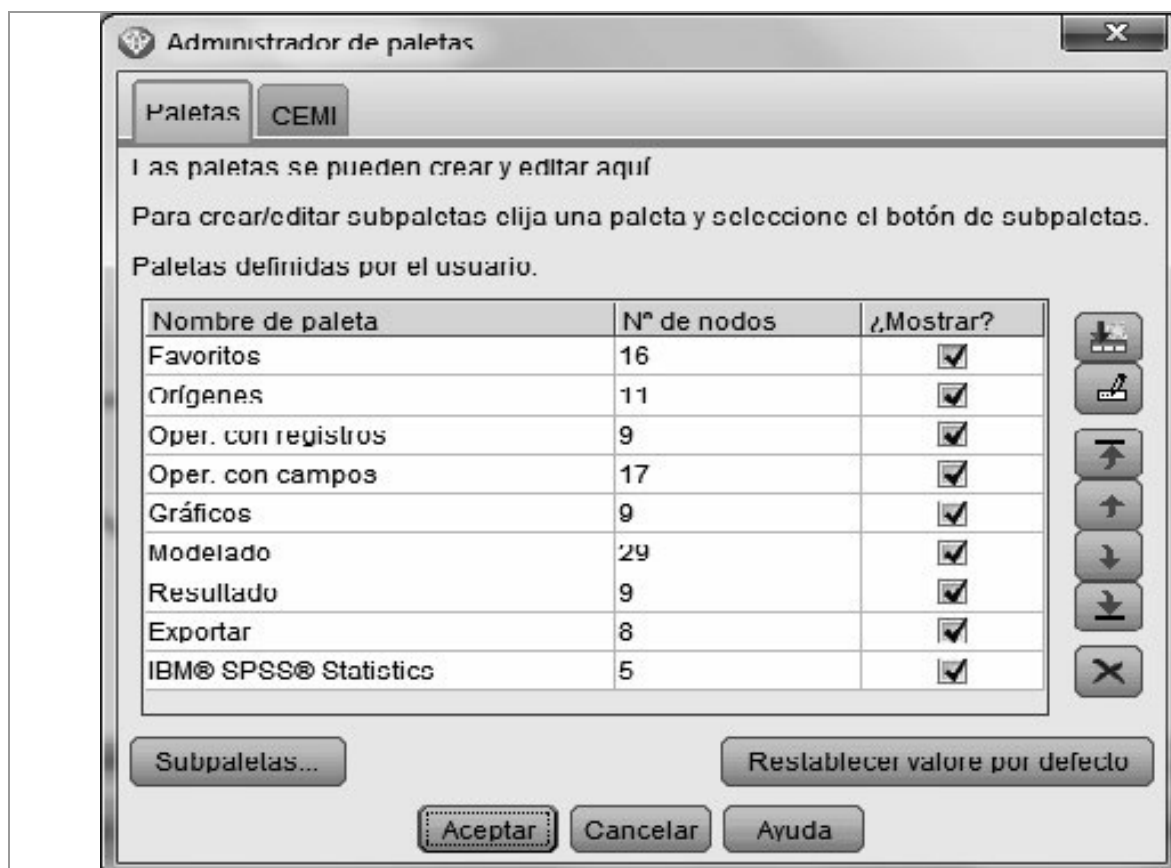
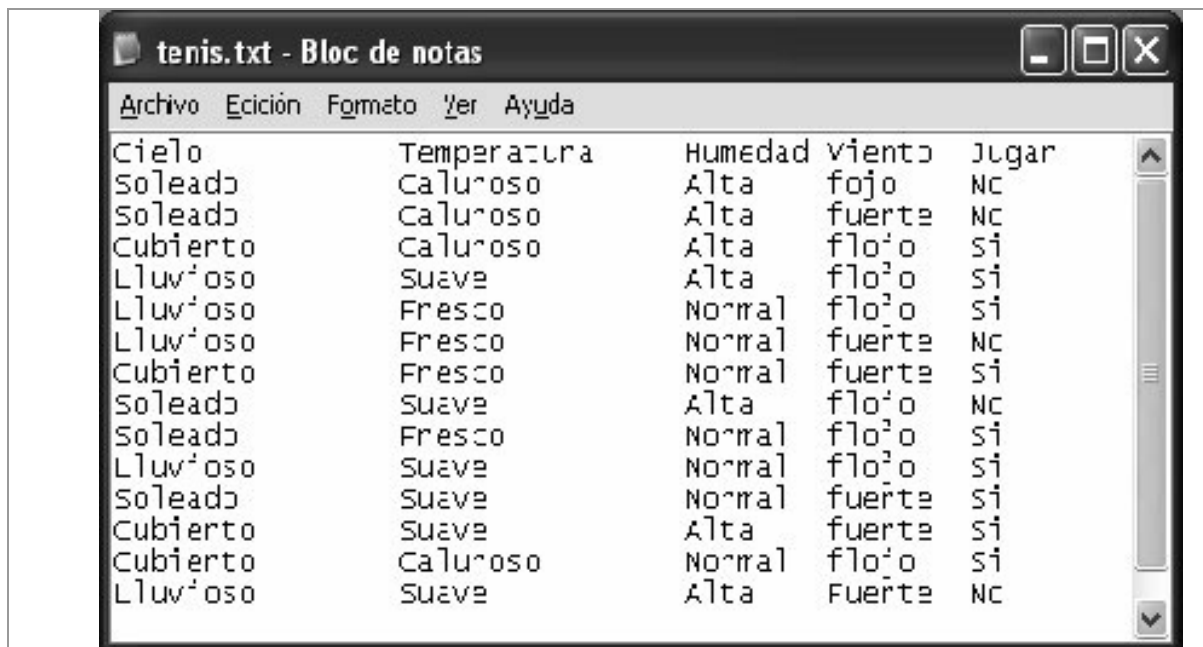


Figura 3-22

EJEMPLO DE TRABAJO CON IBM SPSS MODELER

Con los datos del fichero en formato ASCII de nombre *tenis.txt* situado en el subdirectorio de trabajo, que contiene información acerca de los días que se ha podido jugar al tenis en función de diversos aspectos meteorológicos, se trata de construir un modelo basado en árboles de decisión que permita predecir si a día de hoy es posible jugar al tenis. Los datos se muestran en la Figura 3-30.



Cielo	Temperatura	Humedad	Viento	Jugar
Soleado	Caluroso	Alta	fofo	NC
Soleado	Caluroso	Alta	fuerte	NC
Cubierto	Caluroso	Alta	flo'o	si
Lluvioso	Suave	Alta	flo'o	si
Lluvioso	Fresco	Normal	flo'o	si
Lluvioso	Fresco	Normal	fuerte	NC
Cubierto	Fresco	Normal	fuerte	si
Soleado	Suave	Alta	flo'o	NC
Soleado	Fresco	Normal	flo'o	si
Lluvioso	Suave	Normal	flo'o	si
Soleado	Suave	Normal	fuerte	si
Cubierto	Suave	Alta	fuerte	si
Cubierto	Caluroso	Normal	flo'o	si
Lluvioso	Suave	Alta	Fuerte	NC

Figura 3-23



Figura 3-24

Comenzamos abriendo Modeler mediante *Inicio* —> *Todos los programas IBM SPSS Modeler* —> *IBM SPSS Modeler* (Figura 3-24). Al abrir el programa, las dos áreas de trabajo (izquierda superior y derecha superior) aparecen en blanco (mejor dicho en azul y gris).

Insertar un nodo fuente (origen) de datos en el área de trabajo

Lo primero que vamos a hacer es *insertar un nodo fuente de datos* al área de trabajo. Para ello, pinchamos dos veces (o una vez en el nodo y después otra vez en el área de trabajo) en el nodo *Archivo variable* (Figura 3-25) que está en la categoría *Orígenes* en la parte inferior izquierda de la pantalla. Aparecerá el nodo en el área de trabajo, tal y como se muestra en la Figura 3-26.

Si fuese necesario *borrar un nodo*, simplemente se selecciona y se pulsa la tecla *Supr.* También se puede borrar con la opción *Eliminar* del menú de contexto asociado al nodo, el cual se abre pulsando el botón derecho sobre él (Figura 3-27).



Figura 3-25

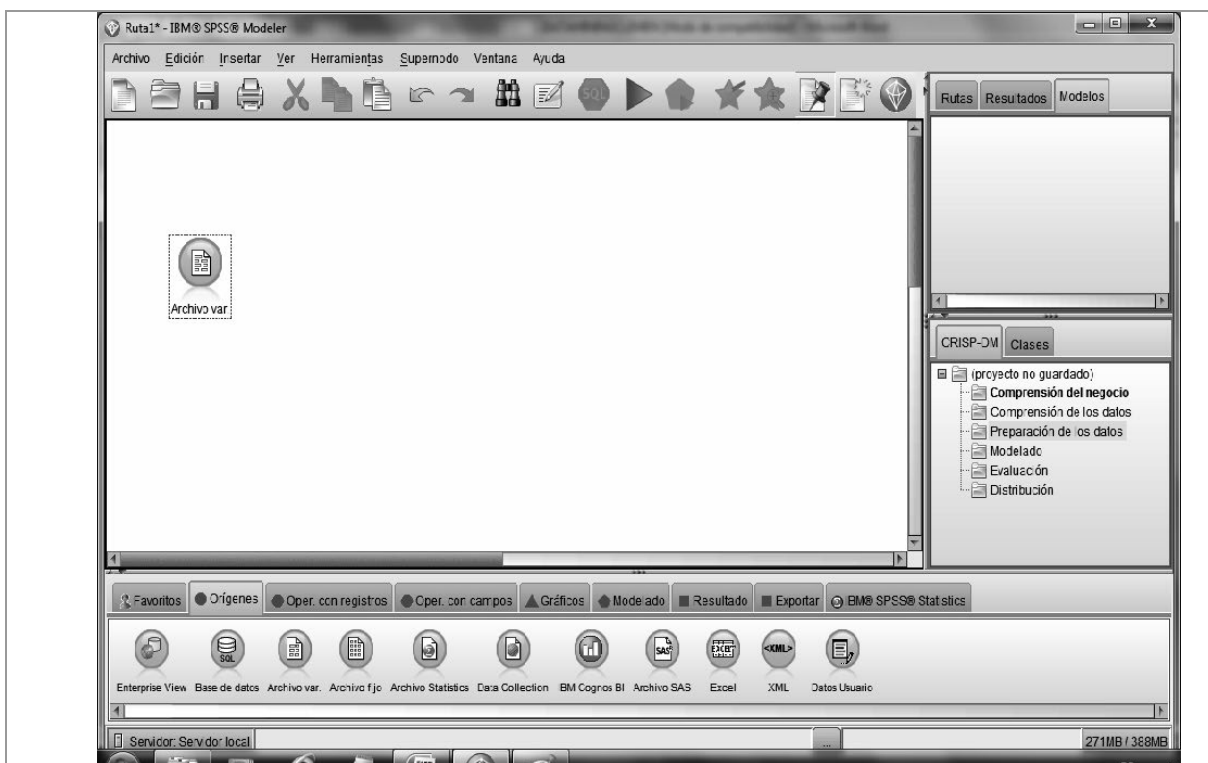


Figura 3-26

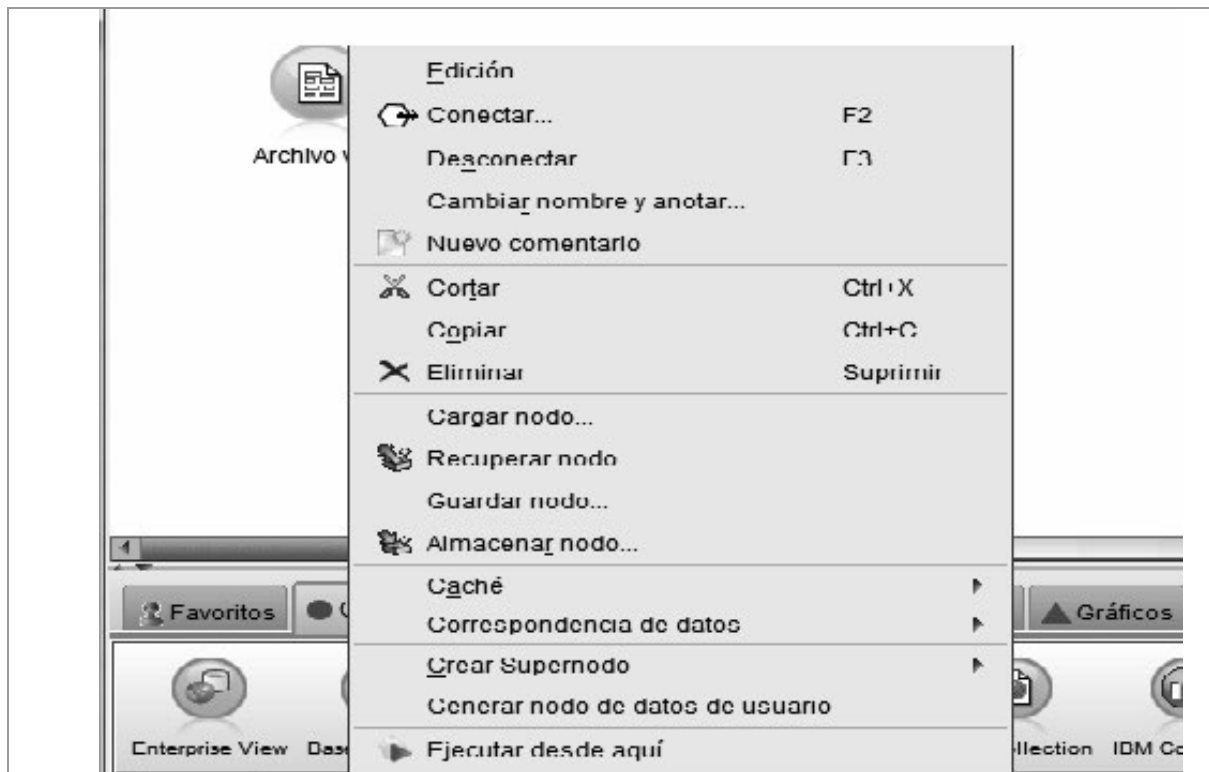


Figura 3-27

Enlazar un nodo con una fuente de datos

La siguiente tarea será enlazar el nodo con una fuente de datos. Para ello, hacemos clic con el botón derecho del ratón sobre el nodo archivo “variable” de la zona de trabajo y seleccionaremos Edición en la Figura 3-27. En la pantalla de edición (Figura 3-28) elegiremos el nombre del fichero, el directorio donde está y la forma de Importarlo (leyendo los nombres de campo en la primera fila del archivo y utilizando como delimitadores los tabuladores). Al hacer clic en Aceptar, el nodo archivo variable aparece ya etiquetado con el nombre del fichero origen de sus datos tenis.txt (Figura 3-29).

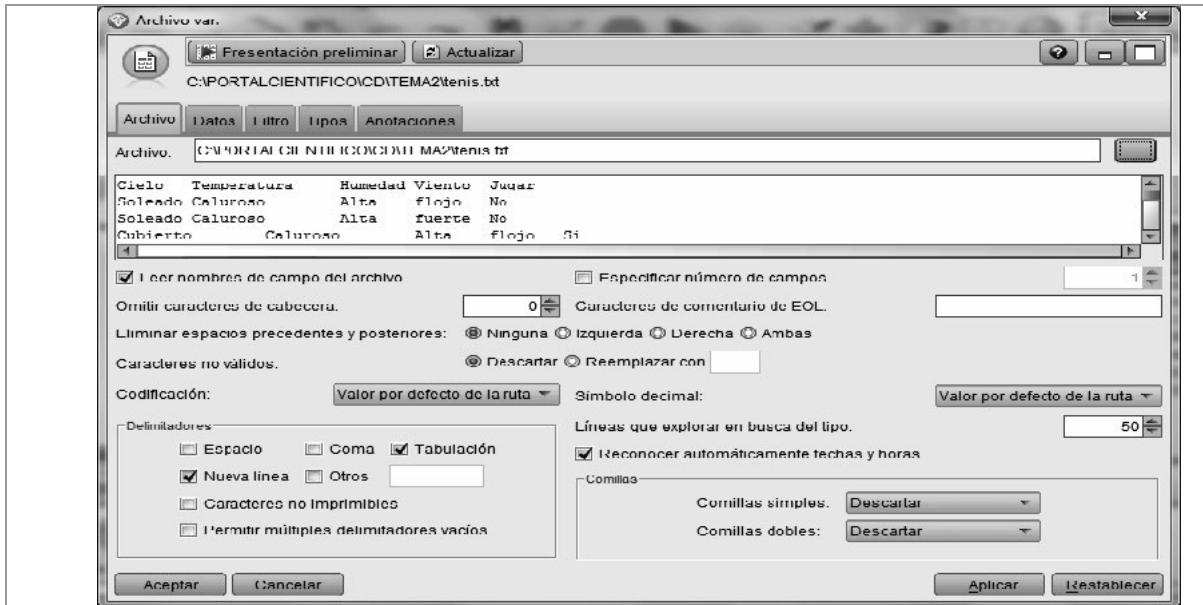


Figura 3-28

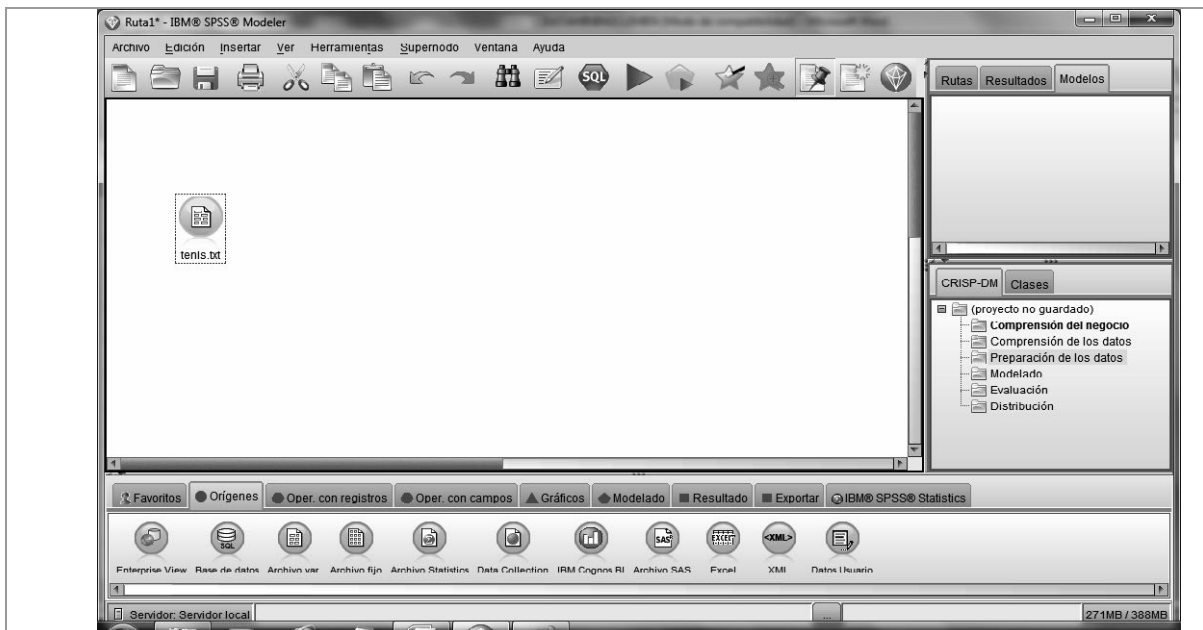


Figura 3-29

Controlar la carga de datos con un nodo Tabla

La siguiente tarea es controlar la carga de los datos añadiendo un nodo *Tabla* de la categoría *Resultados* haciendo doble clic sobre él (Figura 3-30).

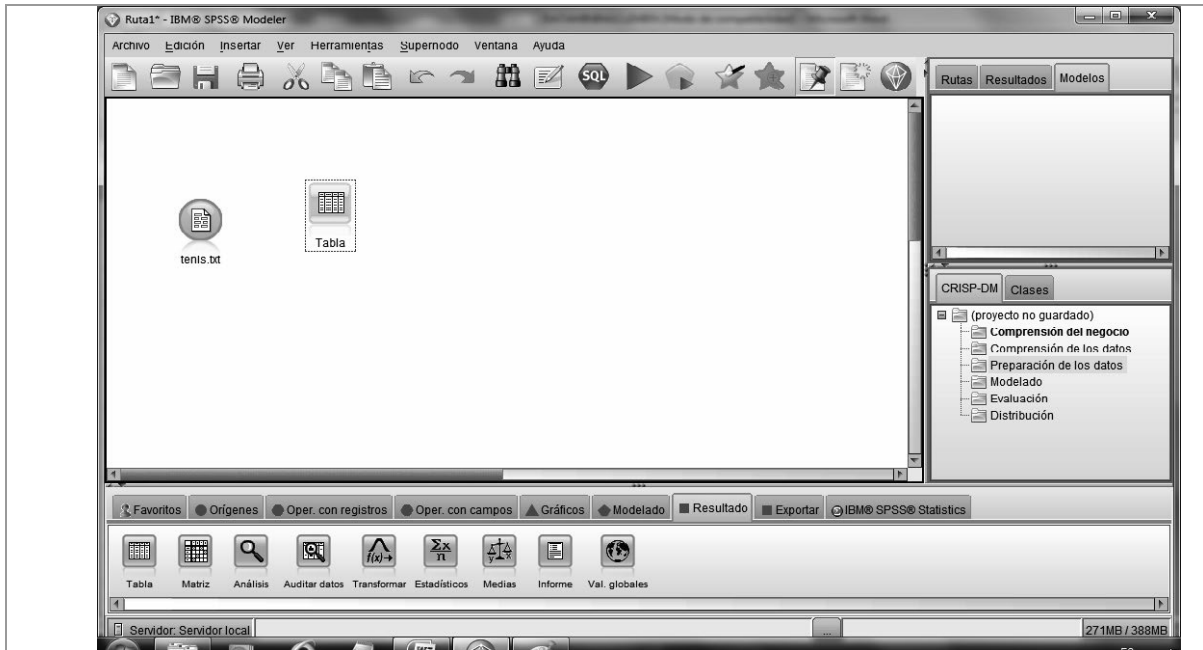


Figura 3-30

Una vez que aparece el nodo *Tabla* en la zona de trabajo, hay que enlazarlo al nodo *tenis.txt*. Para *enlazar dos nodos* en Modeler, se han de pulsar el botón izquierdo y derecho a la vez sobre el nodo origen y arrastrar el ratón hasta el nodo destino (Figura 3-31), soltando en este momento los dos botones. Si el ratón tiene botón del medio, también se puede utilizar este botón. También se puede seleccionar el nodo origen, hacer clic con el botón derecho del ratón y seleccionar la opción *Conectar* en el menú emergente resultante (Figura 3-32).

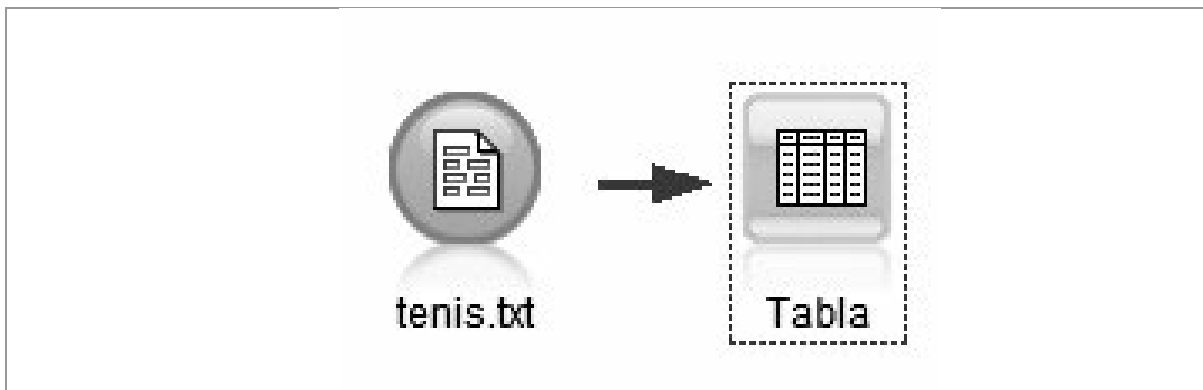


Figura 3-31

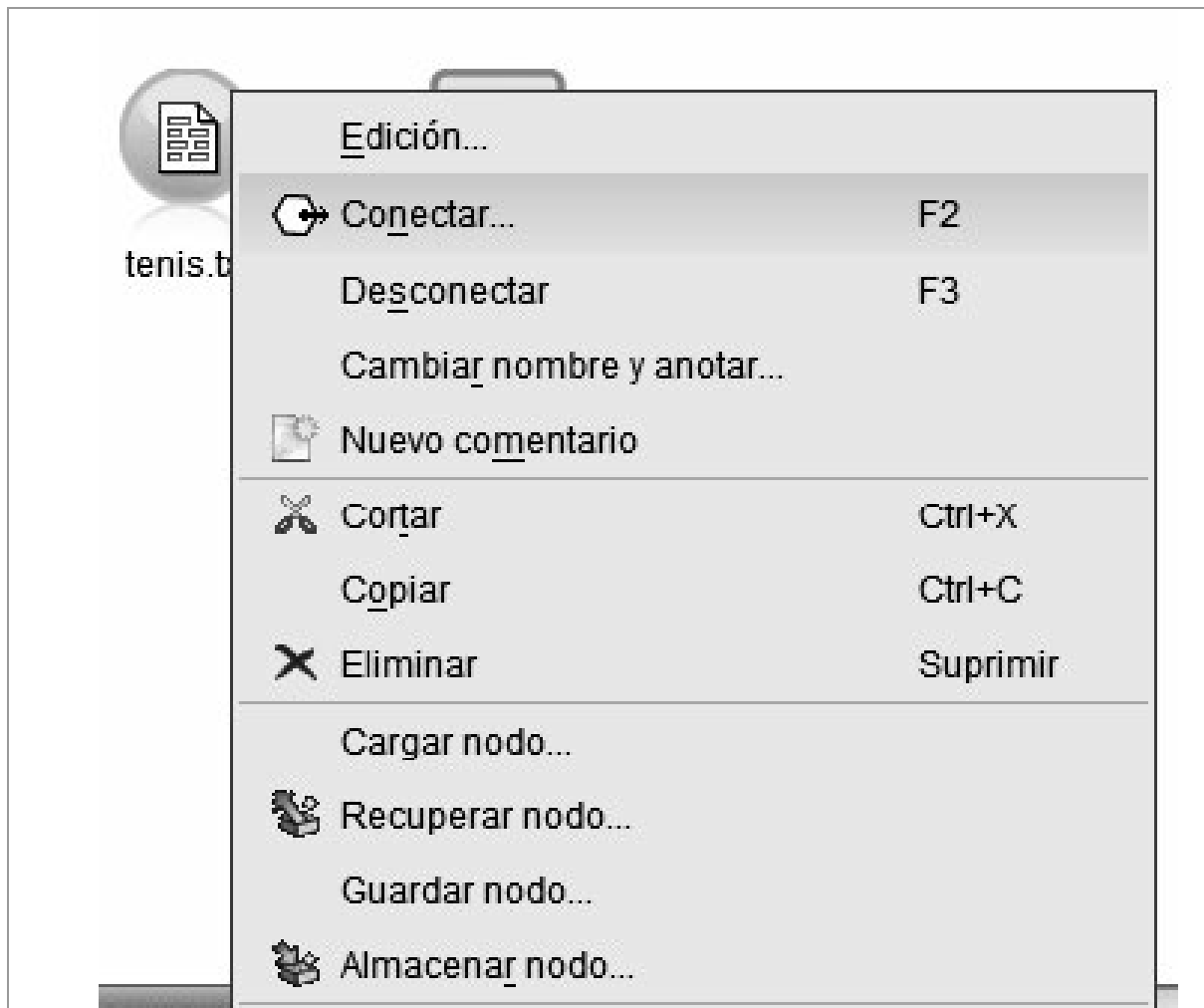


Figura 3-32

Para eliminar un enlace, simplemente se hace clic con el botón derecho en el enlace y en el menú contextual resultante (Figura 3-32) se elige *Eliminar*.

Si una vez conectados los dos nodos, pulsamos con el botón derecho del ratón sobre el nodo *Tabla* y elegimos la opción *Ejecutar* (Figura 3-33), se mostrará el contenido de los datos importados en una tabla (Figura 3-34). También podemos hacer clic en el botón *Ejecutar* de la barra de iconos de opciones de la parte superior de la pantalla.

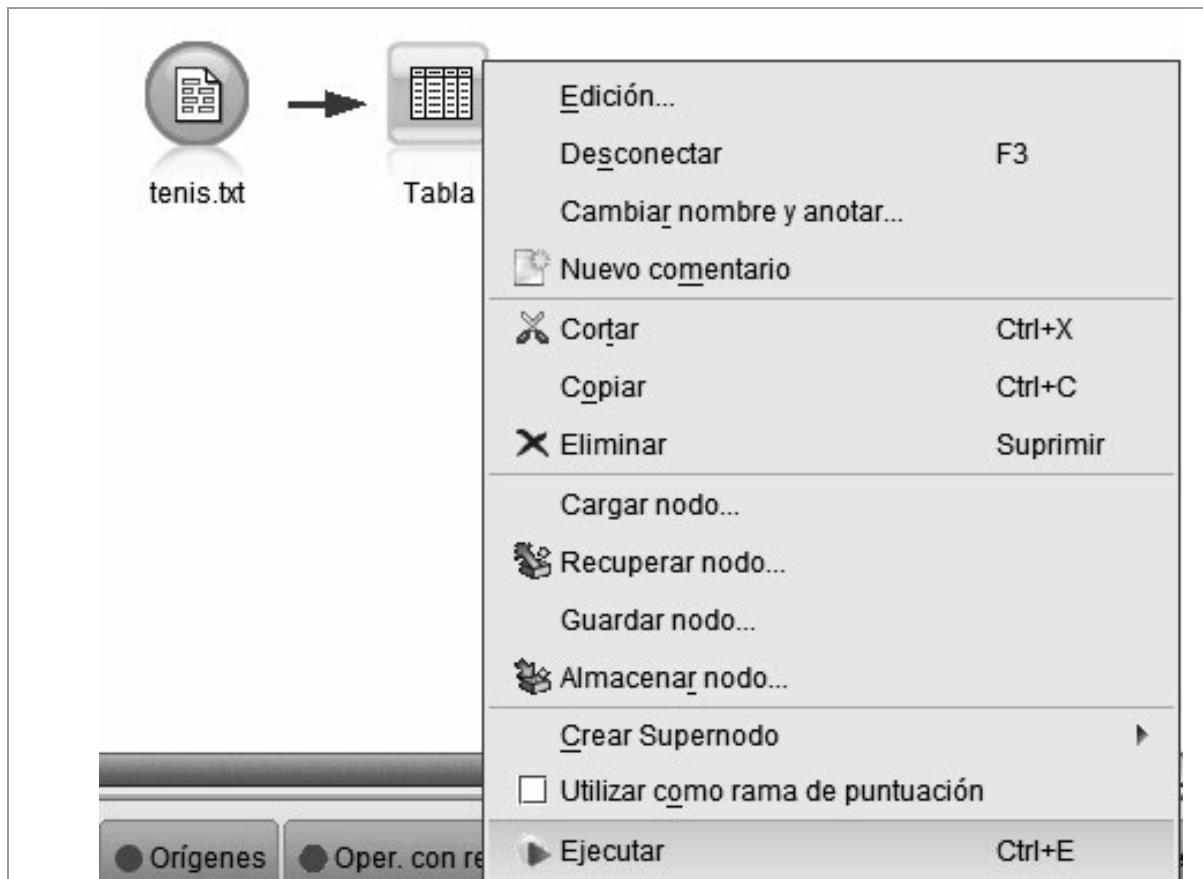


Figura 3-33



Figura 3-34

Definir variables predictoras con el nodo Tipo

Nuestro problema era ver si hoy podemos jugar al tenis. Para poder

abordar este problema hemos de definir los campos *Cielo*, *Temperatura*, *Humedad* y *Viento* como predictores (es decir, de entrada), mientras que el campo *Jugar* es la clase a predecir, o sea, el resultado (es decir, la salida).

Para ello vamos a añadir un nodo *Tipo* que se encuentra en la categoría *Oper. con campos* haciendo doble clic sobre él. A continuación enlazamos el nodo *tenis.txt* con el nodo *tipo* haciendo clic sobre el primero con el botón derecho del ratón, eligiendo conectar en el menú emergente resultante y haciendo clic en el segundo nodo. Se obtiene la Figura 3-35 con los nodos enlazados.

Ahora hacemos clic con el botón derecho del ratón en el nodo *Tipo* y elegimos *Edición* en el menú emergente resultante (Figura 3-36) y se obtiene la tabla *Tipo* de la Figura 3-37. Como vemos en la última columna todos los nodos tienen dirección ENTRADA (están definidos como predictores). Como la salida va a ser la variable *Jugar*, modificamos su dirección a OBJETIVO haciendo clic sobre ENTRADA. Se observa ya la tabla *Tipo* con las propiedades adecuadas (Figura 3-38). A continuación hacemos clic en *Aplicar* y en *Aceptar*.

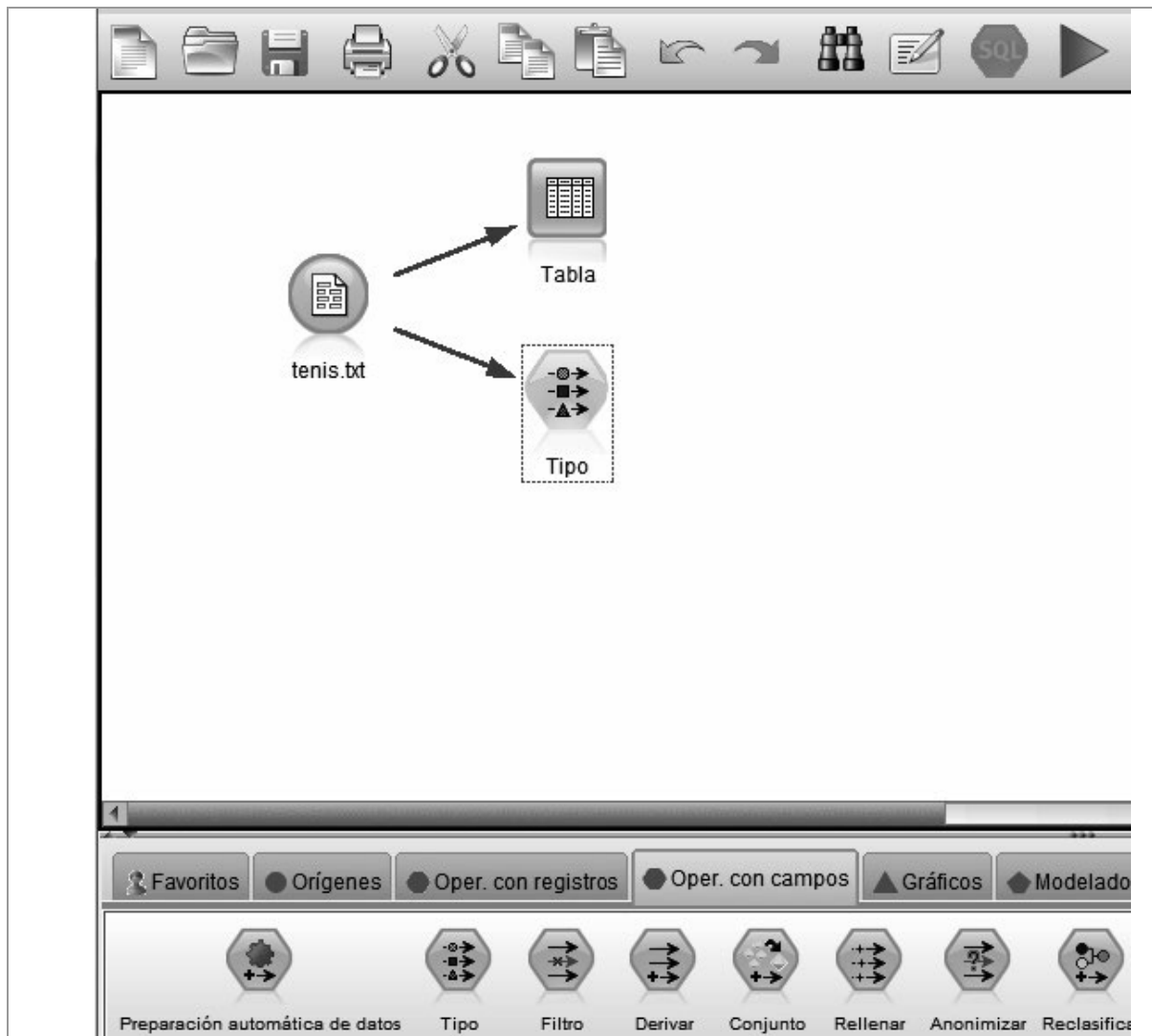


Figura 3-35

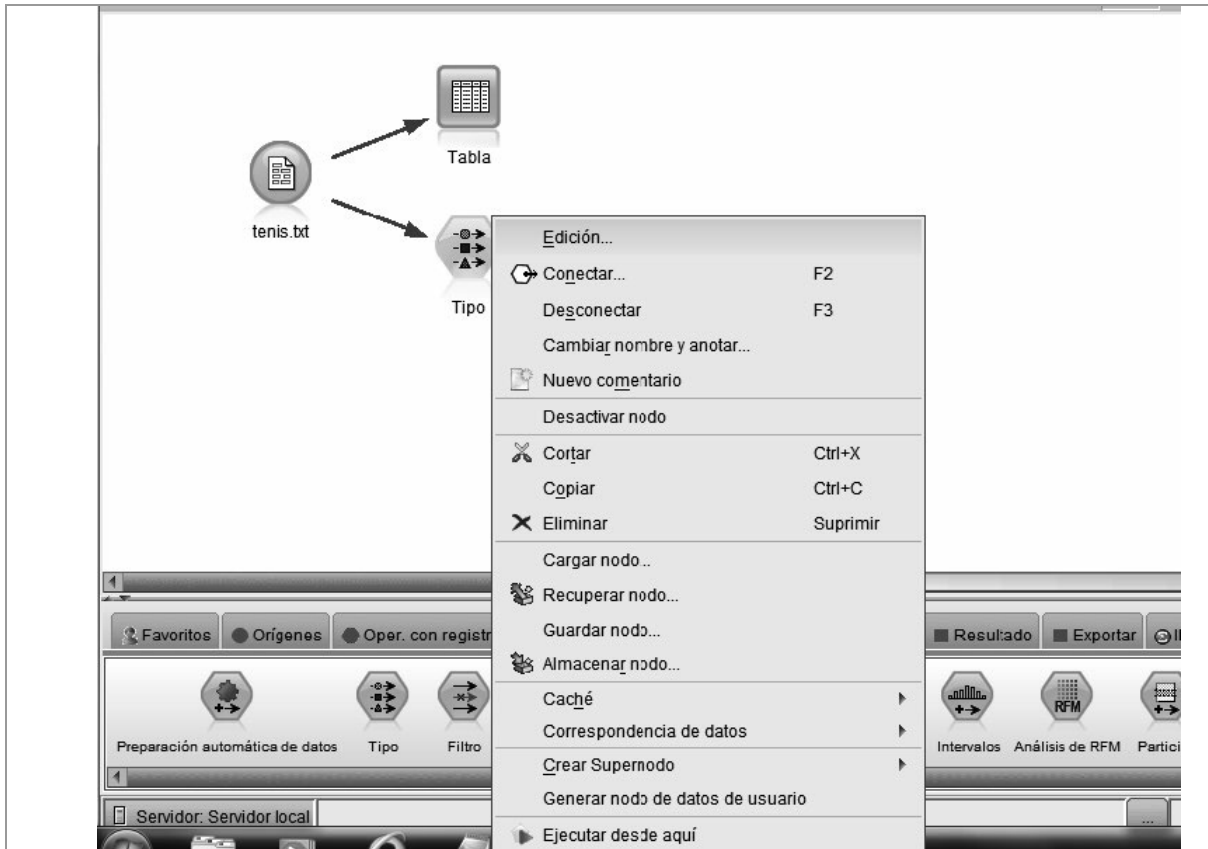


Figura 3-36



Figura 3-37

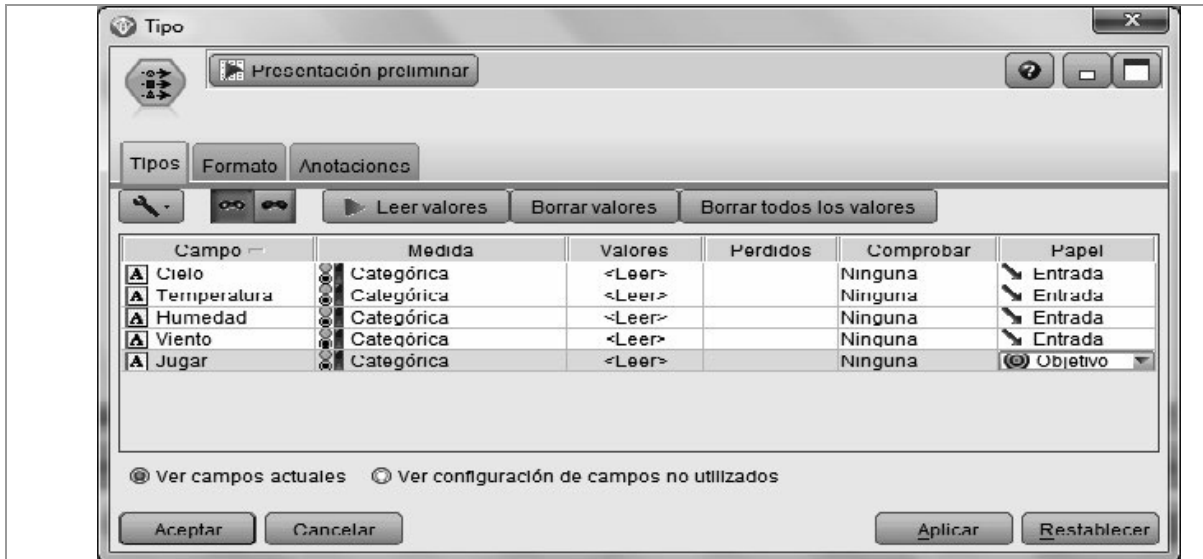


Figura 3-38

Utilizar un nodo de modelado

Ahora ya estamos en situación de intentar utilizar un modelo a partir de los datos, en este caso una función, de modo que dados unos determinados valores de los atributos de entrada obtengamos un valor para la salida. Para ello añadimos el nuevo nodo *Crear C5.0* de la categoría *Modelado* con el objeto de construir un árbol de decisión con los datos. A continuación conectamos el nodo *tipo* con el nodo *Crear C5.0* (Figura 3-39) que pasa a llamarse *Jugar*.

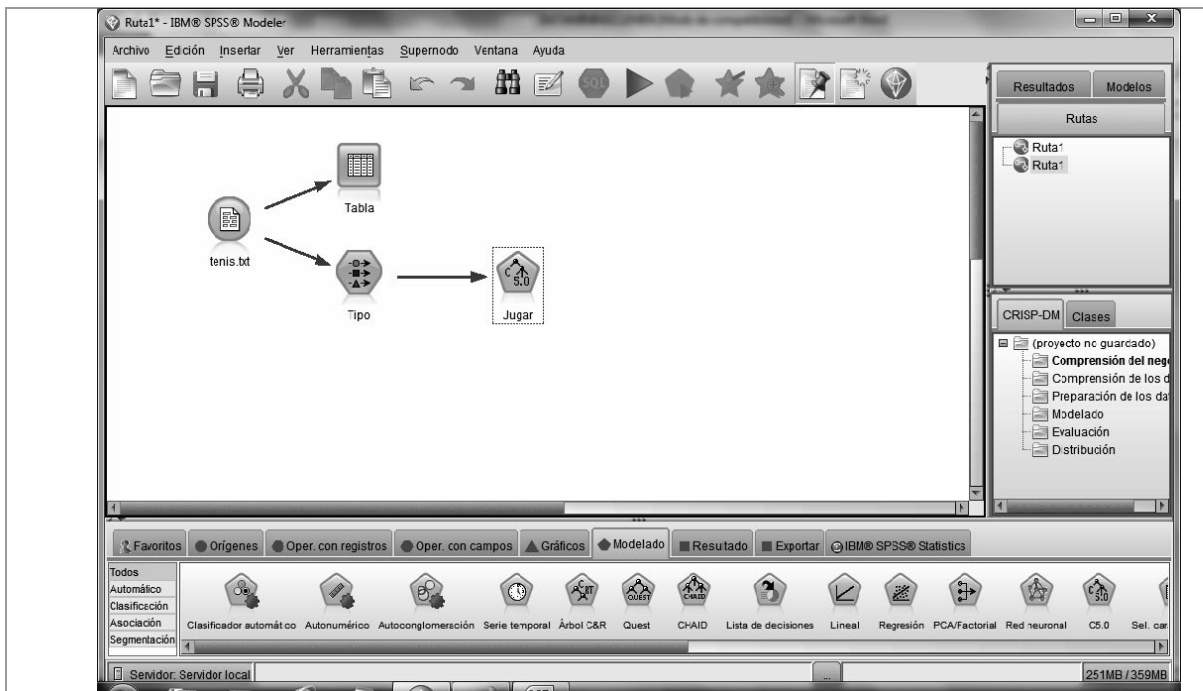


Figura 3-39

Ejecutar de una ruta

Ahora ya estamos en disposición de trabajar con el árbol de decisión). Para ello, hacemos clic en el nodo *Jugar* con el botón derecho del ratón y elegimos la opción Ejecutar en el menú emergente resultante (Figura 3-40) o hacemos clic en el icono *Ejecutar* de la barra de iconos de la parte superior de la pantalla. Se observa que se ha generado un nuevo icono en la pestaña *Modelos* situada en el área de trabajo de la parte superior derecha, con la forma de un diamante (Figura 3-41).

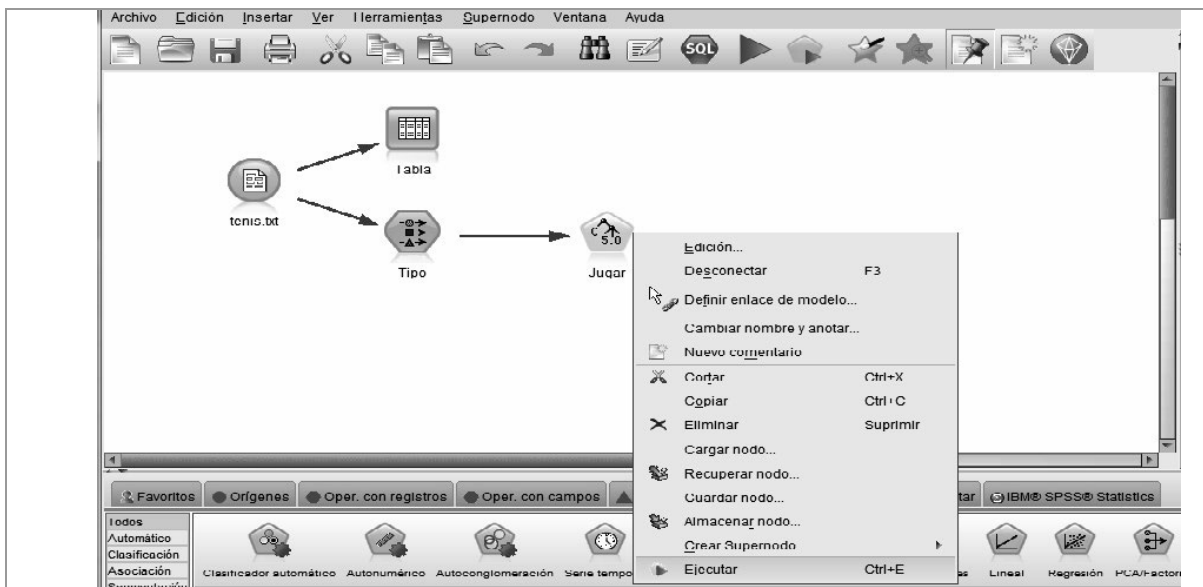


Figura 3-40

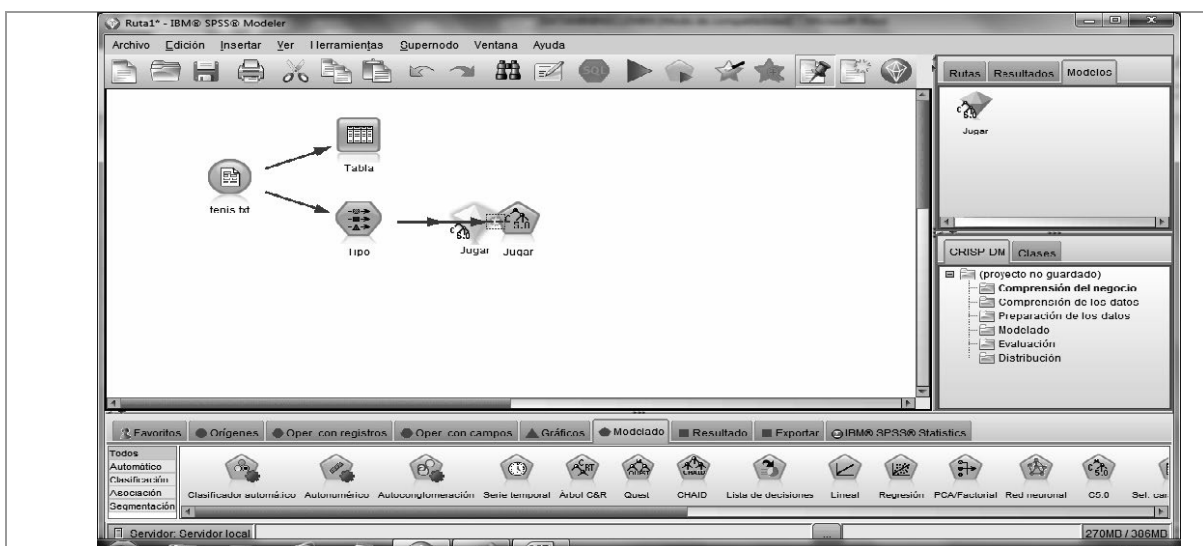


Figura 3-41

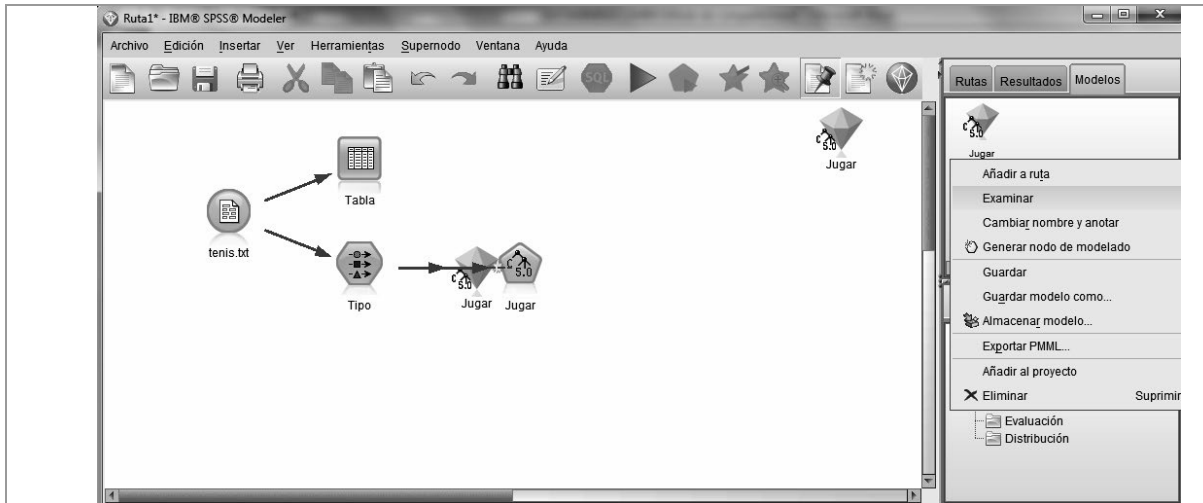


Figura 3-42

A continuación hacemos clic con el botón derecho del ratón en el diamante de la pestaña *Modelos* y, en el menú emergente resultante (Figura 3-42) elegimos *Examinar*. Se obtiene una ventana donde podemos ver el árbol de decisión creado (Figura 3-43) en la pestaña *Visor*.

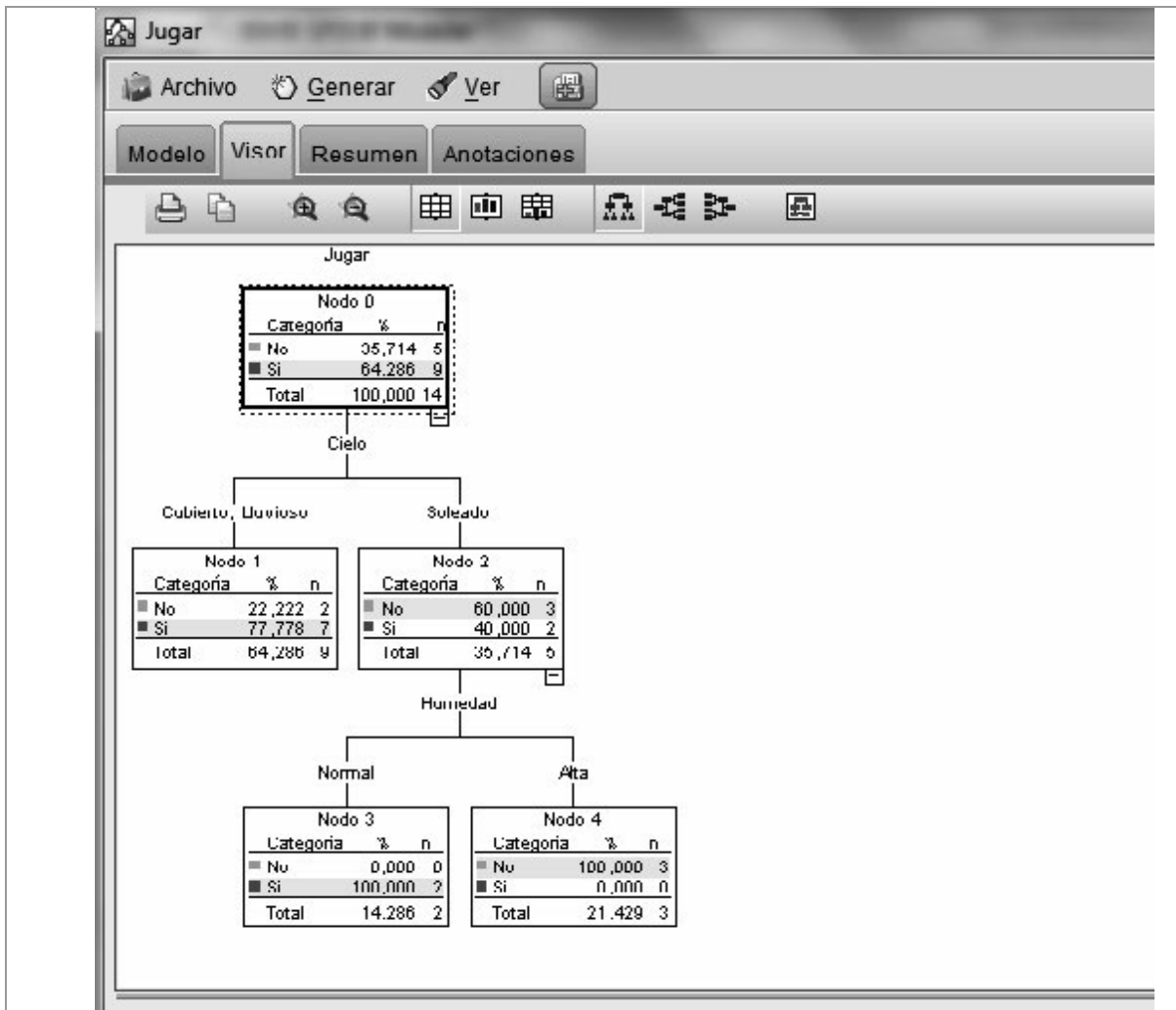


Figura 3-43

Para ver el árbol en forma de reglas y un gráfico con la Importancia de las categorías del predictor se hace clic en la pestaña *Modelo* de la parte superior izquierda de la pantalla del visor (Figura 3-44).

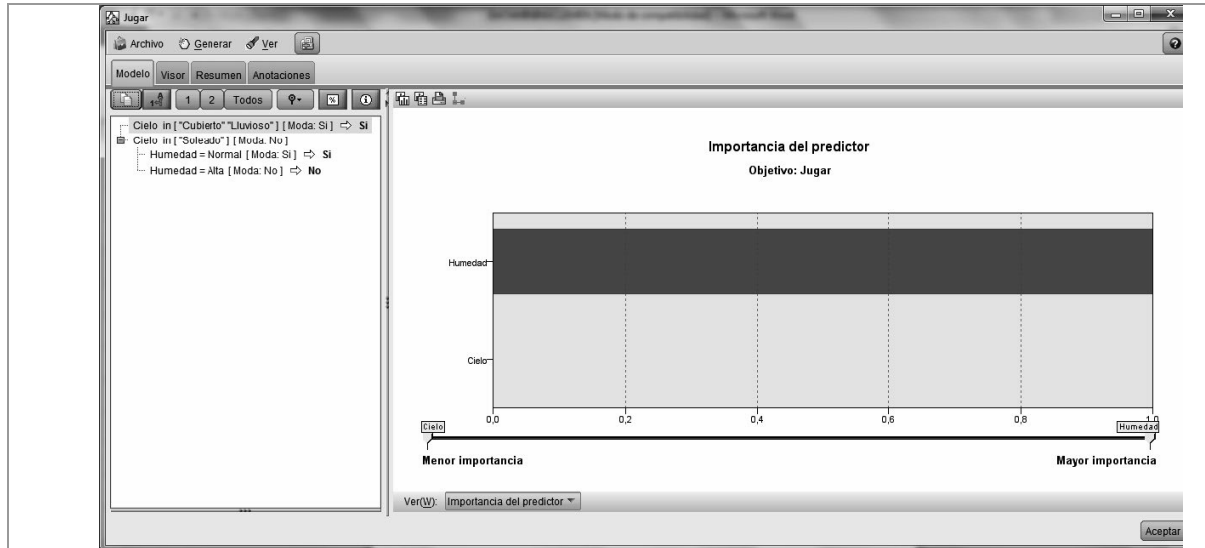


Figura 3-44

Predecir con un modelo

La finalidad última de nuestro modelo de árbol de decisión es predecir si podemos jugar o no jugar al tenis hoy según el tiempo que haga.

Observando el árbol de la Figura 3-43 se deduce que si hoy el cielo está cubierto se jugará al tenis con una confianza del 77,7%. Si hoy el cielo está lluvioso se jugará al tenis con una confianza del 40%. Si hoy el cielo está soleado y la humedad es normal se jugará al tenis con una confianza del 100%, pero si el cielo está soleado y la humedad es alta no se jugará al tenis con una confianza también del 100%.

Guardar un modelo

Mediante *Archivo* —> *Guardar ruta* (Figura 3-45) se almacena la ruta que hemos seguido para construir y utilizar el modelo. De esta forma será posible utilizarlo posteriormente mediante *Archivo* —> *Abrir ruta* y el botón *Ejecutar*.

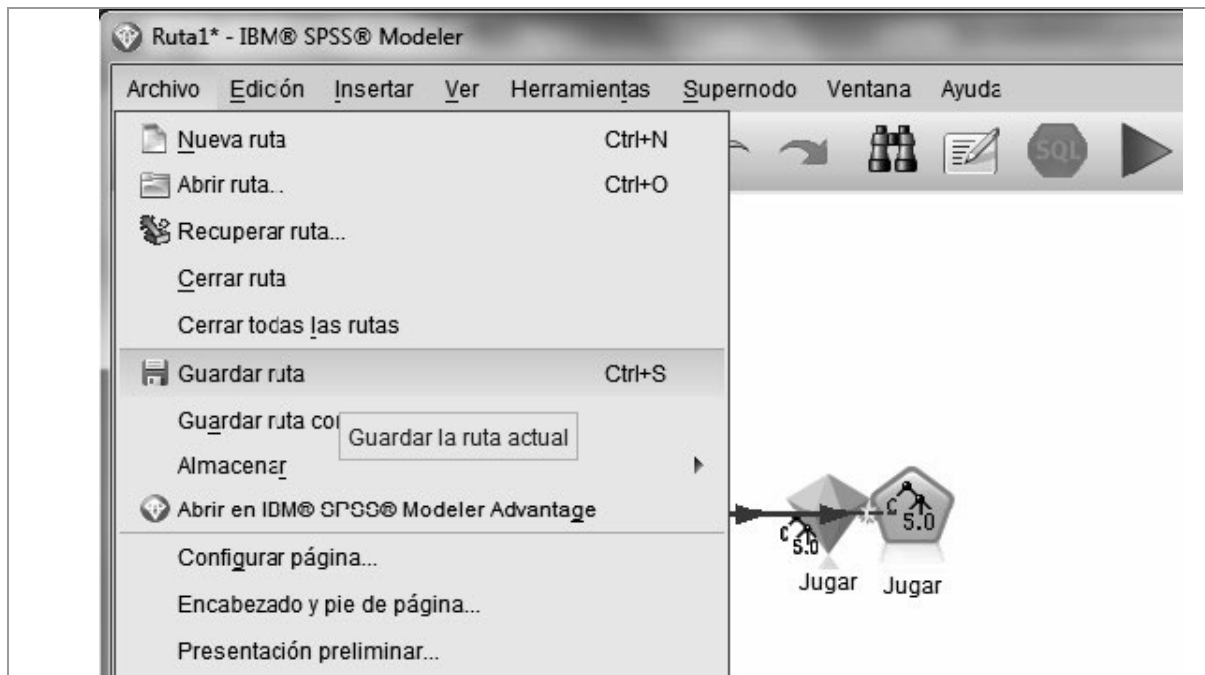


Figura 3-45

NODOS DE ORÍGENES DE DATOS

Modeler ofrece las opciones de obtención de datos de diversas fuentes a través de los nodos de orígenes de datos (*Orígenes*) que se muestran en la paleta *Orígenes* de la Figura 3-46.



Figura 3-46

Los nodos más importantes de la paleta *Orígenes* son los siguientes:

- *Archivo variable*: permite obtener datos ASCII en formato libre.
- *Archivo fijo*: permite obtener datos ASCII en formato fijo.
- *Bases de datos*: permite obtener datos de bases de datos vía ODBC.
- *Excel*: importa datos de formato Excel.
- *Archivo SAS*: importa datos de formato SAS.

Las opciones de la paleta *IBM Statistics* permiten importar datos de SPSS. La más importante es *Archivo Statistics* para importar archivos SPSS.

NODOS DE OPERACIONES CON REGISTROS

Modeler presenta un grupo de nodos cuya finalidad es la selección y transformación de los datos, que como ya sabemos es una fase previa a la aplicación de las técnicas de minería de datos. Modeler permite operaciones con registros y operaciones con campos. Los nodos relativos a operaciones con registros, que son las que nos ocupan en este apartado, se muestran en la paleta *Operaciones con registros* (Figura 3-47).



Figura 3-47

La paleta Operaciones con registros contiene los siguientes nodos:

- *Seleccionar*: permite seleccionar un subconjunto de registros según una condición especificada, tanto para incluirlos como para excluirllos del análisis.
- *Muestrear*: permite obtener una muestra de los registros iniciales.
- *Fundir*: permite combinar en un solo archivo registros provenientes de distintos archivos que tienen campos diferentes, con la condición de que haya un campo común para poder hacer la fusión.
- *Equilibrar*: permite corregir desajustes de registros en conjuntos de datos.
- *Ordenar*: permite ordenar registros de modo ascendente o descendente de acuerdo a los valores de uno o más campos.
- *Distinguir*: permite ignorar registros duplicados.
- *Agregar*: permite reemplazar una secuencia de registros de entrada por un resumen suyo.
- *Añadir*: permite concatenar conjuntos de registros. Se utiliza para unir conjuntos de datos con estructuras similares.

NODOS DE OPERACIONES CON CAMPOS

Dentro de los nodos cuya finalidad es la selección, preparación y transformación de los datos, que como ya sabemos es una fase previa a la aplicación de las técnicas de minería de datos, Modeler dispone de nodos relativos a operaciones con campos, que se muestran en la paleta *Operaciones con campos* (Figura 3-48).



Figura 3-48

Entre los nodos de la paleta Operaciones con campos destacan los siguientes:

- *Filtrar*: permite filtrar la información de múltiples campos simultáneamente.
- *Derivar*: permite obtener nuevos campos en función de otros campos.
- *Tipo*: permite especificar determinadas propiedades de los campos como su tipo, naturaleza (predictor o predicho) y definición de blancos.
- *Rellenar*: permite reemplazar blancos según una condición.
- *Histórico*: permite crear nuevos campos conteniendo datos de registros previos. Se usa para datos secuenciales, como por ejemplo las series temporales.
- *Preparación automática de datos*: prepara automáticamente los datos para modelado.
- *Intervalo tiempo*: crea intervalos de tiempo igualmente espaciados.
- *Reorganizar campos*: cambia el orden de clasificación de los campos.
- *Transponer*: transpone los registros a campos y los campos a registros.

- *Reestructurar*: crea nuevos campos a partir de uno o más campos categóricos.
- *Crear*: crea nuevos campos de marcas a partir de uno o más campos establecidos.
- *Partición*: divide los datos en subconjuntos independientes.
- *Análisis RFM*: realiza análisis de actualidad, frecuencia y monetario de los datos agregados.
- *Intervalos*: convierte los datos numéricos en conjuntos.
- *Reclasificar*: vuelve a categorizar los valores del conjunto.
- *Anonimizar*: anonimiza nombres y valores de campos.
- *Conjunto*: permite a los modelos múltiples ofrecer un resultado combinado.

NODOS PARA GRÁFICOS

Determinadas fases de la minería de datos necesitan de las representaciones gráficas. Por ejemplo, es posible conectar un nodo gráfico a un conjunto de datos para ver su distribución.

Dentro de la paleta *Gráficos*, Modeler dispone de los que se muestran en la Figura 3-49.



Figura 3-49

La funcionalidad de los nodos de la paleta *Gráficos* es la siguiente:

- *Tablero*: crea gráficos basados en campos seleccionados.
- *Gráfico*: permite crear gráficos de líneas y de dispersión.
- *Distribución*: permite graficar la distribución de los valores de una variable que puede ser cualitativa.
- *Histograma*: permite graficar la distribución de los valores de una variable cuantitativa.
- *Malla*: permite graficar las relaciones entre los valores de dos o más variables cualitativas.
- *Colección*: permite crear histogramas que muestran la distribución de los valores de una variable numérica relativos a cada valor de otra.
- *Gráfico múltiple*: permite realizar varios gráficos de líneas sobre los mismos ejes.
- *Gráfico de tiempo*: representa una o varias series respecto al tiempo.
- *Evaluación*: permite evaluar y comparar modelos predictivos eligiendo el mejor modelo para su aplicación.

NODOS PARA MODELADO

Los nodos de modelado constituyen el corazón del proceso de minería de datos. Modeler ofrece gran variedad de métodos de modelado asociados con las distintas técnicas de *data mining*. Dentro de la paleta *Modelado*, Modeler agrupa los nodos por funcionalidades: *Modelado automático* (Figura 3-50), *Clasificación* (Figura 3-51), *Asociación* (Figura 3-52) y *Segmentación* (Figuras 3-53).



Figura 3-50



Figura 3-51



Figura 3-52



Figura 3-53

Entre los nodos más importantes de la paleta *Modelado* destacan los siguientes:

- *Red neuronal*: permite crear y entrenar una red neuronal (perceptrón multicapa).
- *C5.0*: permite construir árboles de decisión y conjunto de reglas utilizando el algoritmo C5.0.
- *Kohonen*: permite crear y entrenar redes neuronales de Kohonen, que suelen usarse para crear clusters cuando no se conoce el número inicial de grupos.
- *Regresión*: permite crear y estimar un modelo de regresión lineal simple o múltiple.
- *A priori*: permite descubrir reglas de asociación en los datos mediante cinco métodos distintos utilizando un esquema sofisticado de indexado para procesos eficientes con grandes conjuntos de datos.
- *K-Medias*: permite realizar el método K-Medias de análisis cluster.
- *Logística*: permite crear y ajustar modelos de regresión logística con la finalidad de clasificar registros.
- *PCA/Factor*: permite ejecutar técnicas de reducción de la dimensión como el análisis factorial y las componentes principales.
- *Bietápico*: permite realizar análisis cluster por el método de las dos fases, que suele utilizarse cuando se mezclan variables cualitativas y cuantitativas.
- *Discriminante*: permite construir modelos predictivos de análisis discriminante.

NODOS DE RESULTADO

Los nodos de resultado o salida permiten obtener información acerca de los datos y modelos mediante la presentación de tablas, análisis de modelos, estadísticas, exportación de datos, etc. Dentro de la paleta *Resultado*, Modeler dispone de los nodos que se muestran en la Figura 3-54.



Figura 3-54

La paleta Salida contiene los siguientes nodos:

- *Tabla*: permite crear una tabla con los datos de un análisis para mostrarlos o guardarlos en un fichero.
- *Matriz*: permite crear una tabla que muestra las relaciones entre dos campos.
- *Análisis*: permite analizar, evaluar y comparar modelos predictivos.
- *Auditar datos*: crea una visualización de resumen de estadísticas y permite manejar los valores atípicos, extremos y perdidos.
- *Transformar*: realiza transformaciones de datos interactivas.
- *Estadísticos*: calcula estadísticos para los campos seleccionados.
- *Medias*: compara las medias entre pares de campos o entre grupos de un campo.
- *Val. globales*: permite computar estadísticos de campos para usar en expresiones.
- *Informe*: permite obtener informes con formato de texto fijo y expresiones.

NODOS DE EXPORTACIÓN

Los nodos de exportación envían resultados a otras aplicaciones y formatos. Dentro de la paleta Exportación, Modeler dispone de los nodos que se muestran en la Figura 3-55.



Figura 3-55

A continuación se presentan las funcionalidades de los nodos de exportación más importantes.

- *Archivo plano*: permite escribir datos en un archivo ASCII plano.
- *Base de datos*: permite exportar datos a bases de datos vía ODBC.
- *Statistics Export*: permite exportar datos a formato IBM SPSS.
- *Data Collection Export*: permite llamar a un procedimiento SPSS para analizar datos.
- *Excel*: permite exportar datos a formato Excel.
- *Exportar SAS*: permite exportar datos a formato SAS.
- *Exportación a XML*: permite exportar datos a XML.

IBM SPSS MODELER E IBM SPSS STATISTICS

IBM SPSS Modeler puede utilizarse conjuntamente con IBM SPSS Statistics (software general de IBM para Estadística). Desde la propia instalación de Modeler puede ejecutarse *Statistics* mediante la ruta *Todos los programas -> IBM SPSS Modeler -> Utilidad de Statistics* (Figura 3-56).

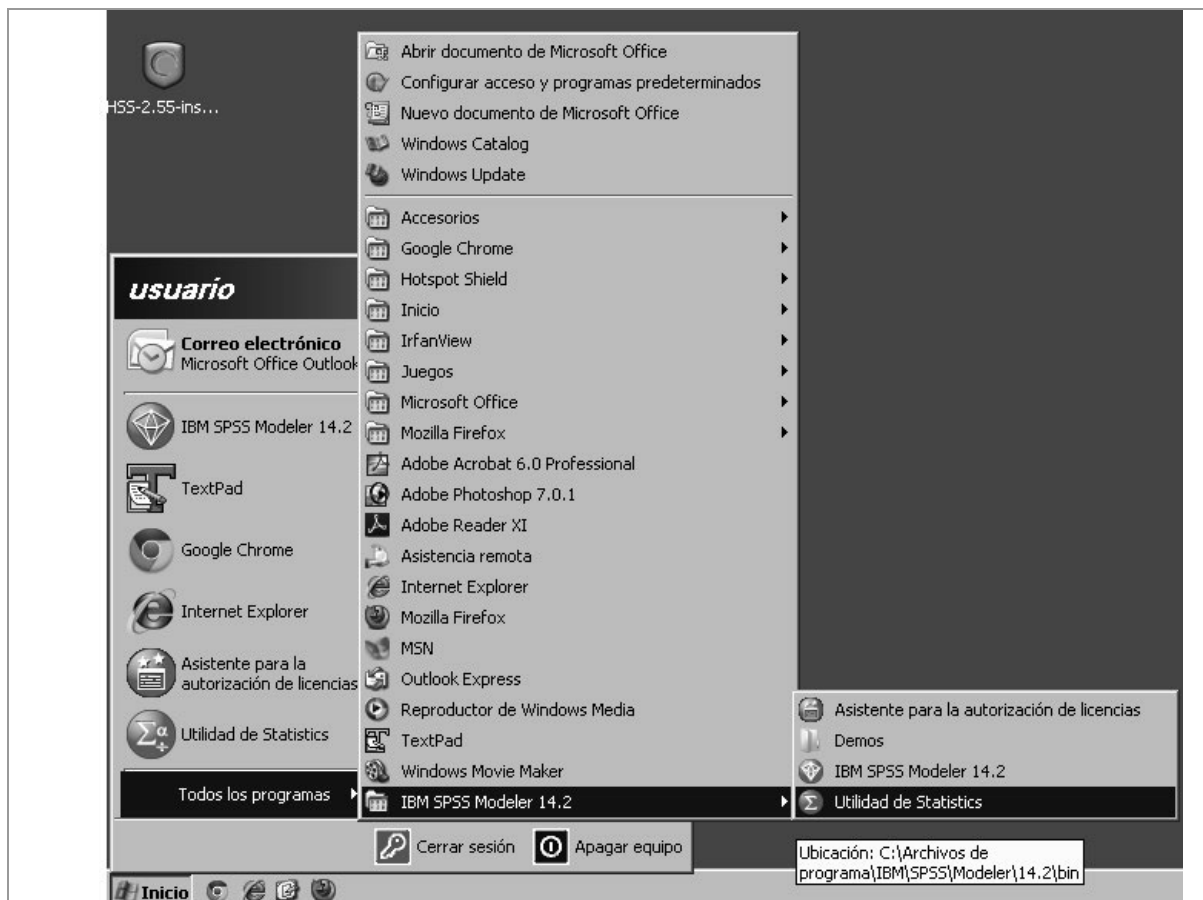


Figura 3-56

Por otra parte, también pueden utilizarse los nodos del grupo IBM SPSS Statistics (Figura 3-57) para interrelacionar entre Modeler y Statistics. El nodo *Archivo Statistics* permite leer directamente en Modeler un archivo de datos de IBM SPSS Statistics, el nodo *Transformación de Statistics* permite realizar transformaciones de datos mediante la sintaxis de comandos de IBM SPSS Statistics, el nodo *Modelo de Statistics* permite construir en Modeler un modelo de IBM SPSS Statistics, el nodo

Resultados de Statistics permite realizar análisis de datos mediante la sintaxis de comandos de IBM SPSS Statistics y el nodo *Statistics Export* permite exportar datos en formato de almacenamiento de Statistics.



Figura 3-57

Por lo tanto existe gran interrelación entre ambos programas para complementarse a la hora de realizar tareas de minería de datos. A lo largo de este texto explotaremos dicha relación.

CAPÍTULO 4

BIG DATA CON HERRAMIENTAS DE ORACLE

ORACLE Y EL BIG DATA

Hoy en día, las empresas utilizan los datos para modelar y controlar los procesos y manejar el negocio. Este torrente de nuevos datos ofrece una oportunidad de obtener una visión sin precedentes y probar nuevas ideas de manera rápida. También proporciona el poder para cambiar las operaciones comerciales de manera fundamental.

Big Data es la electricidad del siglo XXI: un nuevo tipo de poder que transforma todo lo que toca en los negocios, el gobierno y la vida privada.

Mientras que la electricidad tardó más de 100 años en transformar el mundo, los grandes volúmenes de datos están revolucionando la forma en que las empresas y el gobierno funcionan prácticamente día y noche.

Oracle dispone de herramientas para aprovechar el poder de los grandes volúmenes de datos de los negocios. Estas herramientas permiten maximizar el valor de los grandes volúmenes de datos y extraer el conocimiento contenido en ellos.

Los sistemas de ingeniería de Oracle se envían preintegrados para reducir los costes y la complejidad de las infraestructuras TI. Aceleran la implementación, aumentan el rendimiento y reducen el riesgo de los proyectos de grandes volúmenes de datos, NoSQLy Hadoop.

Oracle aporta capacidades adicionales a sus bases de datos clásicas con la incorporación de las herramientas de Blg Data. Son destacables las siguientes características:

- La estrecha Integración de las herramientas de Blg Data con la base de datos Oracle.

- Recursos Informáticos Hadoop para los datos en HDFS.
- Habilitar Oracle SQL para acceder y cargar datos Hadoop.
- Habilitar carga rápida y muy eficiente de Hadoop en Base de Datos Oracle.
- Habilitar partición de tablas Hive durante la carga y consulta.
- Disponer de las interfaces gráficas de usuario de Oracle Data Integrator para los flujos de trabajo de transformación de datos sobre Hadoop.
- Transformar automáticamente los programas de R en trabajos Hadoop.
- Procesar grandes volúmenes de archivos XML en consultas XQuery paralelas.
- Realizar cargas en la base de datos con seguridad y autenticación Kerberos.
- Entregar rápidamente aplicaciones de descubrimiento del conocimiento a los usuarios de negocio.
- Realizar consultas con Oracle SQL en lugar de Hadoop.
- Cargar datos de modo extremadamente rápido entre Hadoop y Oracle Database y reducir al mínimo la utilización de CPU de base de datos durante la carga.
- Capacitar a los científicos de datos para utilizar R en los datos en Hadoop y combinar con análisis avanzados en bases de datos muy grandes.
- Procesar grandes volúmenes de datos XML en Hadoop.
- Reducir la complejidad de Hadoop a través de herramientas gráficas.
- Integrar y probar las herramientas de Big Data en Big Data Appliance
- Facilitar el uso de Fladoop para desarrolladores Oracle.

Entre las herramientas de Oracle para Big Data destacan las siguientes:

- Oracle Big Data Appliance

- Oracle Exadata
- Oracle NoSQL Database
- Oracle Exalytics
- Oracle Business Intelligence Enterprise Edition
- Oracle Endeca Information Discovery
- Oracle Data Integrator

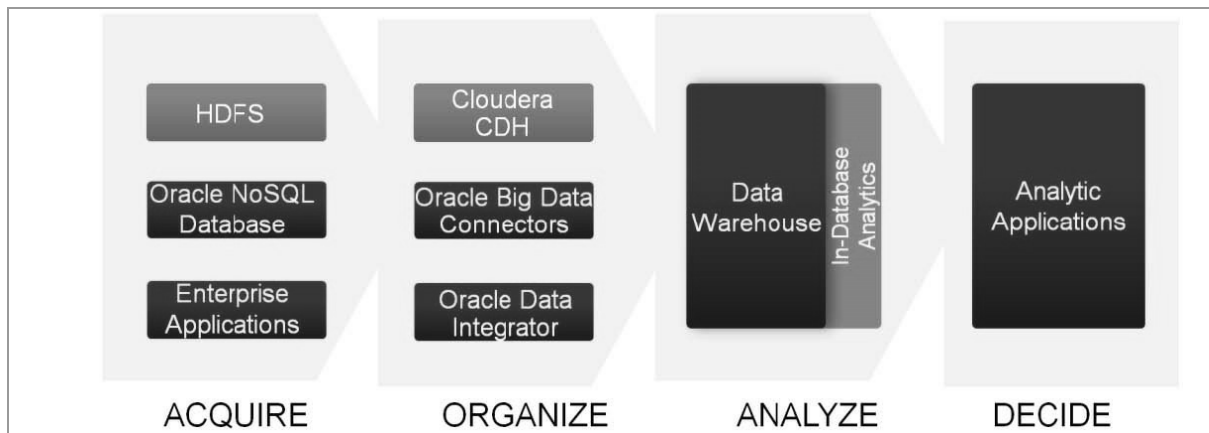
ORACLE BIG DATA APPLIANCE

Oracle Big Data Appliance es un sistema de ingeniería que brinda capacidades de Big Data seguras e integrales para la empresa a un coste aceptable. Está integrado, optimizado y configurado para acortar los tiempos de implementación y reducir el riesgo.

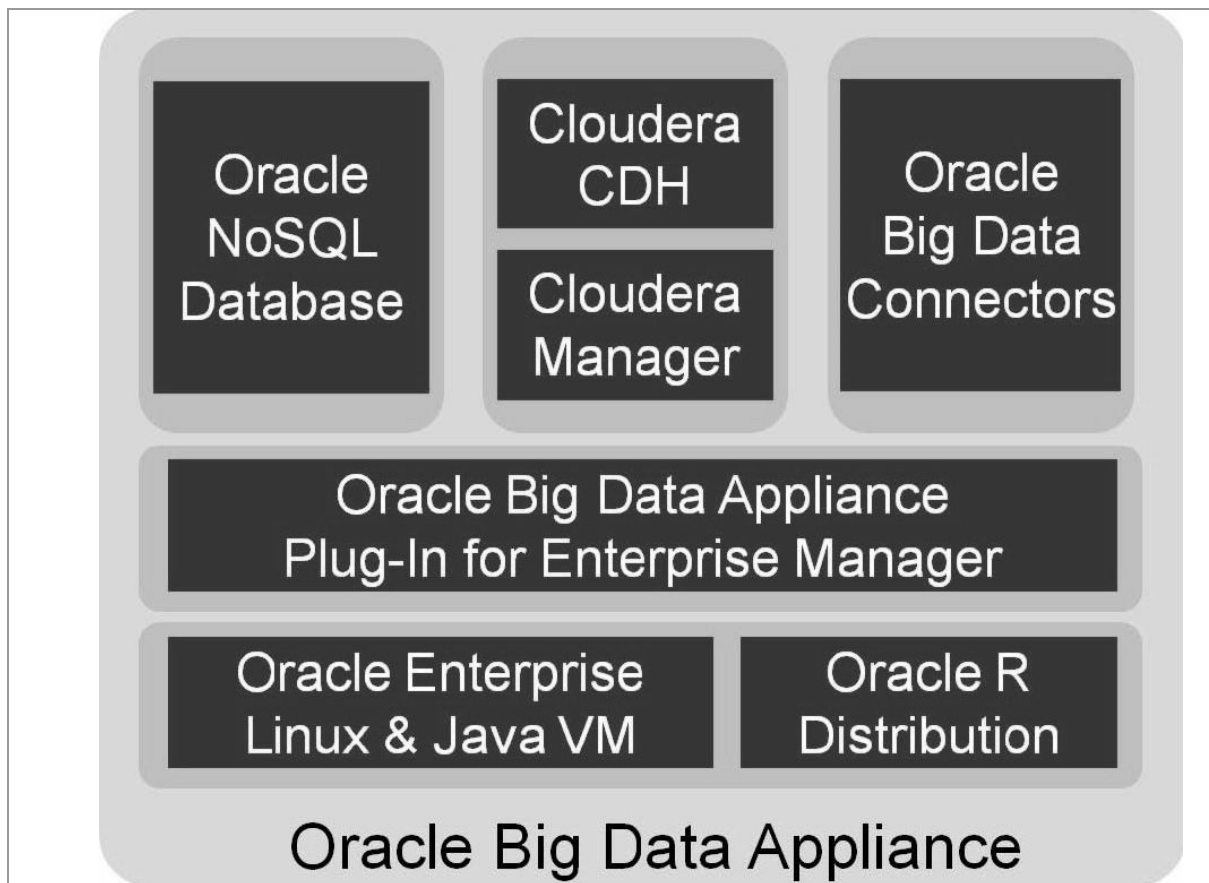
A continuación se enumeran los recursos de esta herramienta de Big Data:

- Oracle Big Data Appliance X4-2 es una configuración de bastidores preintegrada con 18 de los servidores Sun de Oracle que incluyen conectividad de InfiniBand y Ethernet para simplificar la implantación y la gestión.
- Oracle Big Data Appliance X4-2 Starter Rack contiene seis servidores Oracle Sun dentro de un bastidor de tamaño natural con switches redundantes de Infiniband y unidades de distribución de poder. Oracle Big Data Appliance X4- 2 In-Rack Expansión incluye un paquete de seis servidores adicionales para expandir la configuración mencionada en el punto anterior a 12 nodos y luego a un bastidor completo de 18 nodos.
- Todo el software Cloudera Enterprise Technology, inclusive Cloudera CDH, Cloudera Manager y Cloudera RTQ(Impala).
- Oracle NoSQL Database.
- Seguridad integral con capacidades de auditoría, autorización y autenticación.
- Software de sistema adicional, incluidos Oracle Linux y Oracle Java Hotspot VM, y la distribución de Oracle R.

El esquema siguiente organiza los recursos de Oracle Big Data Appliance de acuerdo a la cuatro acciones fundamentales relativas a la adquisición, organización, análisis y toma de decisiones.



No obstante, el esquema siguiente amplía el anterior incluyendo la mayoría de los recursos de Oracle Big Data Appliance y sus relaciones.



Oracle Big Data Connectors

Oracle Big Data Connectors es una suite de software que integra Apache Hadoop con Oracle Database y Oracle Data Integrator. Las organizaciones pueden usar Apache Hadoop para la adquisición de datos y el procesamiento inicial, luego vincular los datos empresariales en Oracle Database para análisis integrados.

Cada vez es más habitual recoger grandes volúmenes de datos y tratarlos en Hadoop, mientras que los sistemas de TI de la empresa se centran en los almacenes de datos relacionales. Oracle Big Data Connectors relaciona el procesamiento de datos en Hadoop con las base de datos Oracle, proporcionando la capacidad crucial para unificar los datos a través de estos sistemas. La combinación de preprocesamiento de grandes volúmenes de datos en bruto y no estructurados en Hadoop con los análisis avanzados, gestión de datos complejos y capacidades de consulta en tiempo real de Oracle Database, proporciona un gran avance en el análisis de la información.

Oracle Big Data Connectors ofrece características que apoyan el descubrimiento de información, análisis profundo y una rápida integración de todos los datos de la empresa. Los componentes de esta suite de software son:

Oracle SQL Connectorfor Hadoop Distributed File System

Oracle Loaderfor Hadoop

Oracle Data Integrator Application Adapterfor Hadoop

Oracle R Advanced Analytics for Hadoop

Oracle XQuery for Hadoop

Oracle Big Data Connectors trabaja con sistemas de ingeniería de Oracle (Oracle Big Data Appliance y Oracle Exadata), así como con las distribuciones de Hadoop y las versiones de bases de datos de Oracle.

Oracle SQL Connectorfor Hadoop Distributed File System

Oracle SQL Connector for Hadoop Distributed File System es un conector de alta velocidad para la carga y consulta de datos de Oracle Database en Hadoop. Conector Oracle SQL para HDFS toma los datos en la base de datos Oracle seleccionándolos a través de SQL. Los usuarios pueden cargar datos en la base de datos, o consultar los datos en Hadoop con Oracle SQL mediante tablas externas. La velocidad de carga de Oracle Big Data Appliance de Oracle Exadata es de 15 TB / hora. El acceso de consulta completa utilizando Oracle SQL permite a los usuarios aplicar el SQL más rico de la industria a los datos almacenados en Hadoop y en la

base de datos Oracle. Conector Oracle SQL para HDFS puede consultar o cargar datos en archivos de texto o tablas Hive más archivos de texto.

Oracle Loaderfor Hadoop

Oracle Loader for Hadoop es un conector eficiente y de alto rendimiento para cargar los datos de Hadoop en Oracle Database. Las transferencias de datos se inician en Hadoop y se trasladan a la base de datos. Oracle Loader for Hadoop se aprovecha de los recursos informáticos de Hadoop para ordenar, particionar y convertir los datos en tipos de datos de Oracle listos para la carga. El pre- procesamiento de datos en Hadoop reduce el uso de CPU de base de datos al cargar datos. Esto reduce al mínimo el impacto en las aplicaciones de bases de datos y alivia la competencia por los recursos, un problema común en el tratamiento de grandes volúmenes de datos. El conector Oracle Loader for Hadoop particularmente útil para cargas continuas y frecuentes utiliza una técnica de muestreo Innovadora distribuyendo los datos en tareas que permiten la carga de los datos en la base de datos en paralelo de forma inteligente. Esto reduce al mínimo los efectos en el rendimiento de desvío de datos, una preocupación común en aplicaciones paralelas.

Oracle Loader for Hadoop puede cargar datos de una amplia gama de formatos de entrada y fuentes de entrada. Nativamente puede cargar datos de archivos de texto, tablas Hive, archivos analizados por una expresión regular de registro y de base de datos Oracle NoSQL. Al cargar datos desde Hive optimiza particiones realizando la carga de forma selectiva. A través de la integración con Hive, Oracle Loader for Hadoop puede cargar desde una variedad de formatos de entrada accesibles a Hive y variedad de clientes fuentes de entrada como por ejemplo HBase. Además, Oracle Loader for Hadoop puede leer formatos de datos propietarios a través de implementaciones de formatos de entrada personalizados proporcionados por el usuario.

Oracle Data Integrator Application Adapter for Hadoop

Oracle Data Integrator Application Adapter for Hadoop proporciona

integración Hadoop nativa dentro de ODI. Se incluyen dentro de Application Adapter ODI para Hadoop módulos específicos de conocimiento ODI optimizados para operaciones en Hadoop. Los módulos de conocimiento pueden ser utilizados para construir metadatos Hadoop dentro de ODI, cargar datos en Hadoop, transformar los datos en Hadoop, y cargar los datos en la base de datos Oracle utilizando Oracle Conector for Hadoop y Oracle SQL Conector para HDFS.

Las implementaciones de Hadoop a menudo requieren código complejo MapReduce Java para ser escrito y ejecutado en el cluster Hadoop. Con el uso de ODI y el adaptador de Aplicación ODI para Hadoop, los desarrolladores utilizan una Interfaz gráfica de usuario para crear estos programas. ODI genera HiveQL optimizado que a su vez genera programas MapReduce nativos que se ejecutan en Hadoop.

Oracle R Advanced Analytics for Hadoop

Oracle R Advanced Analytics for Hadoop ejecuta código R en un cluster Hadoop para análisis escalables. Oracle R Advanced Analytics for Hadoop acelera el análisis avanzado de datos en grandes volúmenes de Información ocultando las complejidades de la Informática basada en Hadoop a los usuarios finales R. El conector se integra con Oracle Advanced Analytics for Oracle Database para ejecutar R y realizar cálculos de minería de datos directamente en la base de datos.

Oracle R Advanced Analytics for Hadoop trabaja con rapidez y alto rendimiento implementando técnicas estadísticas paralelas y de predicción comunes, aprovechando el cluster Hadoop sin necesidad de duplicación de datos o movimiento de datos. La escalabilidad transparente está marcada por la ejecución de código R desde aplicaciones de escritorio Independientes, desarrollados en cualquier IDE que el usuario R elige, de forma paralela en Hadoop. Oracle R Advanced Analytics for Hadoop permite un rápido desarrollo con capacidades de depuración de código R en paralelo en los escritorios de los usuarios, con el apoyo de la simulación de paralelismo. El conector permite a los analistas combinar datos de varios entornos de escritorio cliente, HDFS, de bases de datos Oracle y estructuras de datos R en memoria, todo en el contexto de una sola ejecución de la tarea analítica, lo que simplifica en gran medida la recopilación de datos y preparación. Oracle Advanced Analytics R para

Hadoop también proporciona un marco general para el cálculo y ejecución de código R en paralelo. El rendimiento de E / S de trabajos MapReduce basados en R coincide con el de los programas de MapReduce con el apoyo de la representación binaria de entrada basada en Rdata y Java puro.

Oracle XQuery for Hadoop

Oracle XQuery for Hadoop permite que se utilice XQuery para procesar y transformar texto, XML, JSON y contenido almacenado en un Avro Hadoop Cluster. Oracle XQuery para Hadoop aprovecha al máximo la gran cantidad de CPUs presentes en un cluster típico.

Oracle XQuery para Hadoop se basa en un Hadoop optimizado implementado en Java con motor XQuery de Oracle Database. El motor XQuery evalúa automáticamente las expresiones XQuery W3C estándar en paralelo, aprovechando el marco de MapReduce para distribuir una expresión XQuery para todos los nodos del cluster. Esto permite utilizar expresiones XQuery para ser evaluados en el procesamiento de los datos, en lugar de llevar los datos al procesador XQuery. Este método de evaluación de consultas proporciona mucho mayor rendimiento que está disponible con otras soluciones de XQuery.

Casos de uso típicos para Oracle XQuery para Hadoop incluyen análisis y transformación de registro web sobre operaciones de texto, XML, JSON, y el contenido de Avro. Después de procesar los datos se pueden cargar en la base de datos o indexarlos con Cloudera Search.

Oracle NoSQL Database

Se trata de una base de datos escalada horizontalmente de valor clave para servicios de Internet y nube. Oracle NoSQL Database proporciona un modelo de transacción poderoso y flexible que simplifica enormemente el proceso de desarrollo de una aplicación basada en NoSQL. Realiza escalado horizontalmente con mayor disponibilidad y balance de carga transparente aun cuando agrega una nueva capacidad dinámicamente.

Entre sus recursos destacan:

- Modelo de datos simple por medio de pares de valor clave con

índices secundarios.

- Modelo de programación simple con transacciones ACID, modelos de datos tubulares y soporte JSON.
- Seguridad de aplicaciones con autenticación y cifrado SSL de nivel de session.
- Integrada con Oracle Database, Oracle Wallet y Hadoop.
- Datos geodistribuidos con soporte para múltiples centros de datos.
- Disponibilidad alta con sincronización y fallas remotas y locales.
- Rendimiento escalable y latencia segura.

Continuando con su liderazgo en soluciones de bases de datos de clase empresarial, Oracle anunció Oracle NoSQL Database 3.0. Con esta última versión, Oracle ofrece a los desarrolladores una solución NoSQL mejorada para la construcción de alto rendimiento, aplicaciones de próxima generación. La combinación de seguridad, disponibilidad, escalabilidad y flexibilidad del modelo de datos ofrece una solución de alto rendimiento NoSQL integral para los desarrolladores de aplicaciones.

Oracle NoSQL Database 3.0 es la última versión de base de datos clave-valor distribuida de Oracle, diseñada para simplificar el desarrollo de aplicaciones de próxima generación, al tiempo que permite una mayor seguridad y disponibilidad en tiempo real, a escala web, las cargas de trabajo.

Las mejoras de Oracle NoSQL Database 3.0 mejoran la seguridad, facilidad de uso y el rendimiento necesarios para apoyar el desarrollo empresarial y las necesidades de TI, incluyendo:

- **Mayor seguridad:** autenticación de usuario OS-independiente en todo el cluster basado en contraseñas y Oracle Wallet integración permite una mayor protección contra el acceso no autorizado a datos sensibles. Además, las restricciones de sockets seguros a nivel de sesión Layer (SSL) y del puerto de red ofrecen una mayor protección contra intrusiones en la red.
- **Usabilidad y facilidad de desarrollo:** El apoyo a los modelos de datos tabulares simplifica el diseño de aplicaciones y permite una perfecta integración con las aplicaciones basadas en SQL familiares. Indexación Secundaria ofrece un rendimiento mejorado

dramáticamente para las consultas.

- **Mejoras en el rendimiento del centro de datos:** la conmutación por error automática a centros de datos secundaria del área metropolitana permite una mayor continuidad del negocio para las aplicaciones. Zonas de servidores secundarios también se pueden utilizar para descargar de solo lectura, como las cargas de trabajo de análisis, generación de informes y el intercambio de datos para mejorar la gestión de carga de trabajo.

Oracle NoSQL Database 3.0 Enterprise Edition y Oracle NoSQL Database 3.0 Community Edition ya están disponibles para su descarga a través de Oracle Technology NetWork.

Como proveedor líder de soluciones de gestión de datos, Oracle se ha comprometido a ofrecer a sus clientes las herramientas más amplias para hacer frente a las necesidades en evolución de la gestión de datos empresariales, incluyendo NoSQL. Oracle NoSQL 3.0 ayuda a las organizaciones a llenar la brecha en las capacidades, la seguridad y el rendimiento. La base de datos NoSQL es quizá la mejor de clase empresarial de la industria que permite a los desarrolladores de bases de datos y administradores de bases trabajar de manera fácil, intuitiva y segura para construir y desplegar aplicaciones de próxima generación con confianza.

La nueva versión de la base de datos Oracle NoSQL agrega valores atractivos para los desarrolladores de TI, como seguridad y mejoras de centros de datos de monitoreo que son críticos para las implementaciones empresariales. Las nuevas características de indexación son también una adición bienvenida a la última versión.

Oracle NoSQL ya es un componente crítico de los entornos de producción, y las mejoras de esta nueva versión llevan el producto a un nuevo nivel. Se ha potenciado el uso de los índices secundarios y las funcionalidades de tabla, proporcionando significativas ganancias de productividad y rendimiento, así como la simplificación del código que se usa. Las mejoras de seguridad también son muy destacables.

ORACLE EXADATA DATABASE

Oracle Exadata Database Machine está diseñada para ser una plataforma de máximo rendimiento y disponible para ejecutar una base de datos Oracle. Oracle Exadata funciona en todas las clases de cargas de trabajo de bases de datos, incluso Online Transaction Processing (OLTP), Data Warehousing (DW) y la consolidación de cargas de trabajo mixtas. Simple y de rápida implementación, Oracle Exadata Database Machine respalda y protege sus bases de datos más importantes. Es el cimiento ideal para una base de datos consolidada en la nube.

Oracle Exadata Database Machine X4-2 está disponible en configuraciones de un octavo, un cuarto, medio y un rack completo para cumplir con los distintos requisitos de aplicación y permitirle escalar fácilmente a medida que sus requisitos cambian. El anaquel completo viene con ocho servidores de bases de datos de 2 sockets, 14 Oracle Exadata Storage Servers, switches InfiniBand y más de 44 terabytes de Exadata Smart Flash Cache para dar soporte sumamente rápido a los tiempos de respuesta de transacciones y alta producción.

Exadata Database Machine X3-8 está diseñada para las implementaciones de base de datos que requieren grandes cantidades de datos, que proporcionen rendimiento extremo y escalabilidad a petabyte para todas las aplicaciones, como el procesamiento de transacciones en línea (OLTP), el almacén de datos (DW) y la consolidación de cargas de trabajo mixtas. Viene completo con dos servidores de bases de datos de 8 sockets, 14 Oracle Exadata Storage Servers, switches InfiniBand y más de 44 terabytes de Exadata Smart Flash Cache para dar soporte sumamente rápido a los tiempos de respuesta de transacciones y alta producción.

Oracle Exadata Storage Expansión Rack le permite aumentar la capacidad de almacenamiento de Exadata y el ancho de banda de Exadata Database Machine X4-2 y X3-8, y Oracle SPARC SuperCluster. Fue diseñado para las implementaciones de base de datos que requieran grandes cantidades de datos, como: datos históricos o archivados; respaldos y archivo de datos de Oracle Exadata Database Machine; documentos, imágenes, datos de archivo y XML; grandes datos no estructurados de LOB y otros.

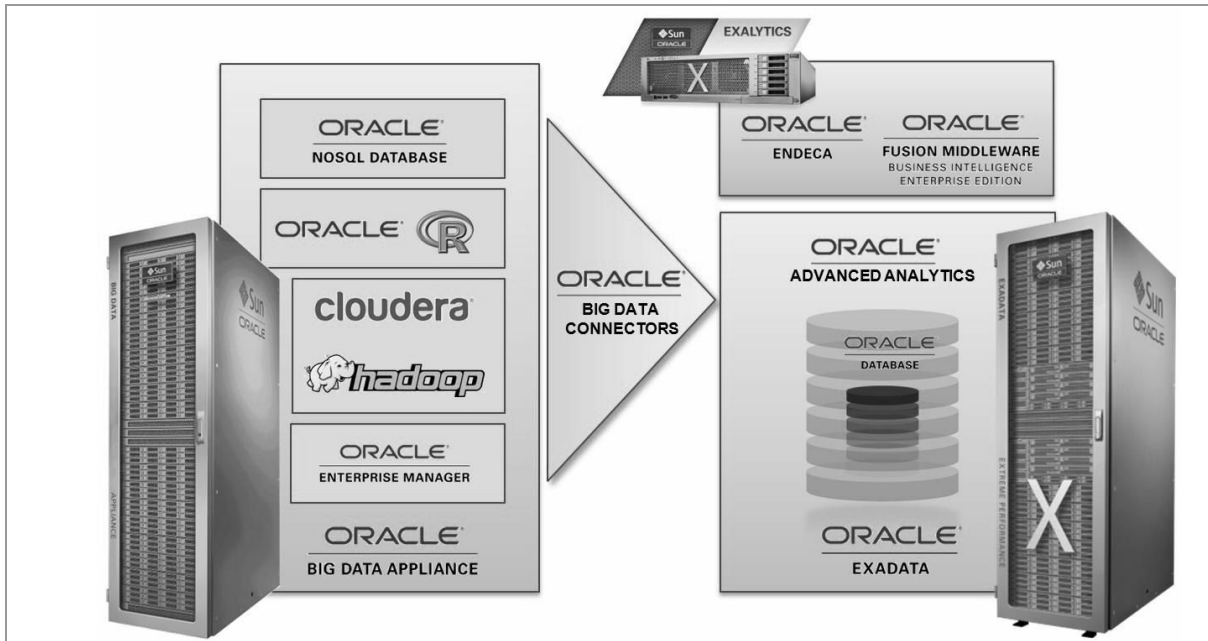
ORACLE EXALYTICS IN-MEMORY MACHINE

Se trata de un sistema de diseño de ingeniería para análisis extremo. Oracle Exalytics es el primer sistema de ingeniería del sector para el análisis en memoria que ofrece un rendimiento extremo para las aplicaciones Business Intelligence y Enterprise Performance Management. Construido con lo mejor en hardware, software de inteligencia de negocios líder en el mercado y tecnología de base de datos en memoria, Oracle Exalytics es un sistema optimizado que ofrece una velocidad de análisis de pensamiento con inteligencia, sencillez y capacidad de gestión inigualables.

Sus características más importantes son las siguientes:

- Excelente plataforma de BI empresarial, software de análisis en memoria y hardware optimizado para trabajar en conjunto.
- Avanzada visualización y exploración de datos, que permite obtener rápidamente conocimiento útil a partir de grandes volúmenes de datos.
- Herramienta muy rápida para aplicaciones de descubrimiento, inteligencia comercial, modelado, pronóstico y planificación.
- Aplicación masiva y consolidación de servidores.
- Acceso a todas las fuentes de datos de Oracle y de otros fabricantes.

La figura siguiente engloba la mayoría de los componentes de Oracle Big



ORACLE BUSINESS ANALYTICS

El valor añadido de las aplicaciones de grandes volúmenes de datos y análisis es que permiten a su organización aprovechar una amplia gama de información para operar más eficientemente, ofrecer nuevos servicios y adelantarse a la competencia. Las soluciones Oracle Business Analytics ayudan a las organizaciones de todos los tamaños a crecer al permitirles descubrir nuevas formas de crear estrategias, planificar, optimizar operaciones comerciales y capturar nuevas oportunidades del mercado.

Como líder del mercado en software de análisis comercial, Oracle brinda las soluciones más completas e integradas que permiten a sus clientes encontrar valor en los grandes volúmenes de datos, obtener conocimiento sobre cada aspecto de su empresa, planificar con anticipación y actuar con confianza en cualquier momento y lugar, desde cualquier dispositivo.

Oracle Business Intelligence Foundation Suite

Las herramientas de esta suite permiten obtener visibilidad instantánea en el rendimiento de los negocios y lograr mejores resultados mediante un amplio conjunto de prestaciones de informes, análisis, modelado y pronósticos. Todo ello diseñado para la velocidad de rendimiento del pensamiento en cualquier dispositivo.

Oracle Business Intelligence Foundation Suite incluye las siguientes capacidades:

Enterprise BI Platform. Transforma las Tecnologías de la Información (TI) de un centro de coste a un activo de negocio mediante la estandarización en una sola plataforma de BI y escalable que permite a los usuarios de negocio crear fácilmente sus propios informes con información relevante para ellos. Proporciona el conjunto más robusto de presentaciones de informes, consultas ad-hoc, análisis OLAP y cuadro de mandos con una experiencia de usuario rica que incluye la visualización, colaboración, alertas y otras opciones.

Entre sus características destacan:

- Habilitar un acceso a los datos corporativos más fácil para los usuarios de negocio.
- Proporcionar una infraestructura común para la producción y entrega de informes empresariales, scorecards, dashboards, análisis ad-hoc y análisis OLAP.
- Incluir visualización rica, tableros interactivos, una amplia gama de opciones de gráficos animados, interacciones de estilo OLAP y la búsqueda de innovaciones y capacidades de colaboración viables para aumentar la adopción del usuario.
- Reducir el costo de una arquitectura orientada a servicios basada en la web probada que se integra con la infraestructura existente de TI.

OLAP Analytics. Aporta el procesamiento analítico en línea multidimensional. Se trata de un servidor líder en la industria (OLAP),

diseñado para ayudar a los usuarios de negocio a pronosticar niveles de rendimiento empresarial probables y realizar análisis de las condiciones variables. Permite la presentación de informes y análisis para miles de usuarios con acceso a conjuntos de datos muy grandes. Además, se trata de un apoyo para descubrir rápidamente y destacar las tendencias en estos grandes conjuntos de datos. También permite desarrollar, mantener e informar de los modelos altamente dimensionales.

Scorecard and Strategy Management. Permite definir metas y objetivos que pueden conectarse en cascada a todos los niveles estratégicos de la empresa, permitiendo a los empleados entender su impacto en el logro del éxito y alinear sus acciones en consecuencia.

Entre sus capacidades destacadas tenemos:

- Proporcionar un marco que organiza el pensamiento estratégico y la medición del desempeño.
- Construir un consenso sobre la dirección estratégica y la comunicará a todos los niveles en la organización.
- Ver y gestionar la estrategia y las iniciativas por línea de negocio, geografía u otras estructuras organizativas.
- Apoyar la planificación estratégica a través de análisis de las relaciones métricas y el aprendizaje organizacional.

Mobile BI. Permiter llevar la información crítica del negocio allí donde se desee. El análisis de negocios no tiene por qué quedar atrapado dentro del PC. Necesitará ser portable en dispositivos móviles. Oracle Business Intelligence (BI) Mobile portfolio es compatible con una de las tecnologías de más rápido crecimiento en el mercado hoy por dar a los clientes accesibilidad al análisis en sus dispositivos móviles.

La Inteligencia de Negocios de Oracle Business Intelligence (BI) Mobile portfolio incorpora información analítica basada en datos a los teléfonos inteligentes y las tabletas, sin comprometer la integridad de los datos o la seguridad. Los usuarios de negocios no tienen que perder el impulso en su jornada de trabajo a medida que salen de su oficina o de viaje de negocios. Es necesario tener acceso instantáneo y constante a la inteligencia crítica para el negocio se esté donde se esté.

Oracle ha lanzado en este campo dos aplicaciones móviles de Business Intelligence: Oracle BI Mobile HD y Oracle BI Mobile App Designer

Oracle BI Mobile HD es la aplicación principal para acceder a la información de inteligencia de negocios desde un teléfono inteligente o tableta. Totalmente integrado con Oracle Business Intelligence Fundación Suite, Oracle BI Mobile HD proporciona acceso inmediato a contenido existente analítico como cuadros de mando e informes sin necesidad de ninguna modificación. El contenido entregado se optimiza automáticamente para la interacción en los dispositivos móviles a través de gestos multi-touch.

Oracle BI Mobile App Designer convierte al usuario de negocios en un diseñador auto-suficiente de aplicaciones profesionales, especialmente diseñadas para móviles analíticos. Una interfaz de usuario de código cero con arrastrar y soltar permite a los usuarios crear rápida y fácilmente aplicaciones impresionantes. Existen móviles interactivos de análisis para cualquier tipo de negocio. Independientemente del sistema operativo o tamaño de la pantalla, las aplicaciones de nueva creación están disponibles en cualquier dispositivo móvil.

Enterprise Reporting. Proporciona una única plataforma basada en la web, para la creación, administración y entrega de informes interactivos, cuadros de mando y todo tipo de documentos con formato complejo.

En esta herramienta destacan las siguientes capacidades:

- Destacar los datos importantes en los informes y cuadros de mando con la resolución necesaria.
- Crear informes diseños a través de Internet o en herramientas de escritorio familiares.
- Ir más allá de la presentación de informes tradición y crear todo tipo de documentos con formato.
- Arrastrar y soltar un par de indicadores de negocio en una hoja en blanco para crear nuevos informes.
- Entregar la información correcta a las personas correctas en el momento adecuado.

Ante el crecimiento exponencial de Big Data y de analítica de negocio, las organizaciones buscan cómo utilizar y convertir volúmenes masivos de datos en bruto en conocimiento inteligente. Hoy en día, los negocios pueden analizar, supervisar y predecir resultados más rápido que nunca antes y aquellos que lo hacen consiguen un rendimiento superior al de sus

competidores.

Power Systems es una herramienta de vanguardia en el suministro de herramientas que obtienen información más rápida a partir de la analítica de información estructurada y de Big Data no estructurada, como vídeo, imágenes y contenido procedente de dispositivos móviles, redes sociales y sensores. Para extraer información y tomar mejores decisiones, las compañías precisan una herramienta, software de sistemas abierto y una plataforma flexible y segura como Power Systems para dar soporte a la continuada carga de datos, ejecutar múltiples consultas simultáneas y ofrecer analítica en tiempo real soportada por un ancho de banda de E/S masivo.

Power System incorpora hardware, software y herramientas adecuadas para el trabajo en Big Data.

Enterprise Performance Management

Las soluciones de planificación comercial de Oracle que alinean finanzas y operaciones permiten planificar y pronosticar con mayor precisión. Asimismo, aceleran los procesos de cierre financiero e informes a la vez que brindan una mayor transparencia y confianza en los números.

Las herramientas incluidas en Enterprise Performance Management son las siguientes:

Oracle Strategy Management

Oracle Financial Close and Reporting

Oracle Planning, Budgeting, and Forecasting

Oracle Profitability and Cost Management

Aplicaciones analíticas

Oracle dispone de herramientas que permiten el enfoque en lo que importa y el descubrimiento de soluciones para los retos de negocio mediante análisis predefinidos basados en la experiencia de Oracle a través de cientos de automatizaciones CRM y ERP. Podríamos clasificar estas herramientas en los siguientes grupos:

Aplicaciones analíticas para su rol comercial

- **Oracle Human Resources Analytics.** Herramienta para el análisis y gestión de recursos humanos.
- **Oracle Sales Analytics.** Herramienta para el análisis de ventas.
- **Oracle Project Analytics.** Herramienta para la gestión de proyectos.
- **Oracle Procurement & Spend Analytics.** Ayuda a las organizaciones a optimizar su rendimiento del lado de la oferta mediante la integración de datos de toda la cadena de valor de la empresa, permitiendo así que los ejecutivos, gerentes y empleados de primera línea tomen decisiones más informadas y procesables.
- **Oracle Financial Analytics.** Ayuda a los gerentes de primera línea a mejorar el rendimiento financiero con información completa, analizando los gastos de sus departamentos y contribuciones de ingresos.
- **Oracle Marketing Analytics.** Permite a la organización obtener los mejores resultados de sus inversiones de marketing, proporcionando todo su equipo de marketing con una imagen completa de las preferencias del cliente, el comportamiento de compra y la rentabilidad.
- **Oracle Service Analytics.** Permite a las organizaciones con visión de gran alcance analizar todos los aspectos del rendimiento del centro de servicio y tomar medidas para aumentar la satisfacción del cliente.
- **Oracle Supply Chain and Order Management Analytics.** Permite a las organizaciones evaluar los niveles de inventario, las necesidades de cumplimiento de productos probables e identificar rápidamente posibles problemas de cartera de pedidos.

Aplicaciones analíticas para su sector

Oracle dispone de herramientas de análisis enfocadas a los diferentes sectores. A continuación se citan las más importantes para cada sector.

- **Ventas al pormenor**

Oracle Retail Merchandising Analytics

Oracle Retail Customer Analytics

- **Servicios financieros**

Oracle Financial Analytics

- **Health Care**

Oracle Enterprise Healthcare Analytics

Oracle Clinical Development Analytics

Oracle Operating Room Analytics

- **Ciencias de la salud**

Oracle Health Sciences Clinical Development Analytics

Oracle Argus Analytics

- **Comunicación**

Oracle Communications Data Model

- **Educación**

Oracle Student Information Analytics

- **Sector público**

Oracle Tax Analytics

- **Fabricación**

Oracle Manufacturing Analytics

- **Valoración de activos**

Oracle Enterprise Asset Management Analytics

- **Otros productos relacionados**

Analytic Applications for Your Business Role

Analytic Applications for Your Product Line

Oracle Business Intelligence Tools and Technology

Oracle Exalytics In-Memory Machine

Aplicaciones analíticas para su línea de producto

Con Oracle ERP y CRM Analytics, se obtienen los conocimientos que

se necesitan para optimizar los recursos, reducir costos y mejorar la eficacia de las actividades que van desde las ventas a los recursos humanos y la contratación. Y gracias a la pre-integración con Oracle de PeopleSoft y JD Edwards, es posible beneficiarse de una implementación rápida y económica.

Information Discovery

A veces se esconden nuevas oportunidades de negocio en las fuentes de datos que se extienden más allá de su almacén de datos. Busque y explore datos estructurados y no estructurados, dentro de la empresa y más allá a través de la herramienta Oracle Endeca Information Discovery.

Advanced Analytics

Oracle permite ejecutar poderosos algoritmos estadísticos y de minería de datos en miles de millones de conjuntos de datos de fila en cuestión de segundos que proporcionan una idea de los temas clave de negocio tales como predicción de rotación, recomendaciones de productos y alerta de fraude. Las herramientas de Oracle adecuadas para estas tareas son:

Oracle Advanced Analytics

Oracle Real Time Decisions

Nube

Oracle permite transformar la forma de administrar su negocio con sistemas ejecutivos de gestión de rendimiento basado en nube compuestos de las mejores aplicaciones a nivel empresarial. La herramienta de Oracle adecuada para esta tarea es Enterprise Performance Management Cloud.

SOLUCIONES DE DATOS RÁPIDOS DE ORACLE

Las soluciones de datos rápidos de Oracle descubren el valor económico de los datos con alta velocidad y alto volumen, permitiendo análisis rápidos y acciones también más rápidas.

Estas soluciones permiten maximizar el valor de datos a alta velocidad. Hoy en día, el mundo se mueve más rápido que nunca, con más volúmenes de datos, más dispositivos y más conexiones. Para mantener el ritmo, se necesita procesar grandes volúmenes de intercambio dinámico de datos. Además, no solo los datos se desplazan dentro de la organización, sino también deben estar en constante actualización fuera del firewall. Los datos de alta velocidad aportan gran valor, sobre todo a los procesos de negocios volátiles. Sin embargo, algunos de estos datos pierden su valor operativo en un corto tiempo. Para extraer el máximo valor de los datos altamente dinámicos y perecederos, es necesario procesar mucho más rápidamente y tomar las medidas oportunas.

La obtención de conocimientos en tiempo real permite:

- Habilitar nuevos servicios que antes no eran posibles.
- Proporcionar mejores experiencias a los clientes en tiempo real a través de una interacción más personalizada.
- Manejar con eficacia los recursos del sistema.
- Lograr una mayor calidad en las operaciones, aumentando la visibilidad y previsibilidad.

Con múltiples tecnologías que trabajan mano a mano para crear valor a partir de la alta velocidad en grandes volúmenes de datos, las soluciones de datos rápidos de Oracle están diseñadas para optimizar la eficiencia y la escala para el procesamiento de eventos de gran volumen y transacciones. A continuación se citan las soluciones de datos rápidos más importantes de Oracle de acuerdo a su finalidad.

- **Filtrar y correlacionar.** Con **Oracle Event Processing**, puede utilizar las reglas predefinidas para filtrar y correlacionar datos a través

de grandes fuentes de datos. Cuando se integra con Oracle Coherence, se puede ejecutar en memoria para optimizar el rendimiento y la escala.

- **Mover y transformar.** Con **Oracle Data Integrator Enterprise Edition** y **Oracle GoldenGate** puede capturar datos (estructurados o no) y pasar inmediatamente la información donde se necesita y en el formato adecuado para apoyar mejor la toma de decisiones.
- **Analizar.** **Oracle Business Analytics** permite realizar análisis en tiempo real y ampliar lo que sea posible para su negocio.
- **Actúe.** **Oracle Real-Time Decisions**, junto con **Oracle BPM**, ayuda a apoyar tanto la toma de decisiones automatizadas, como las interacciones más complejas, de origen humano para la gestión inteligente de los procesos de negocio.

Todos estos componentes se ejecutan en un entorno de un rico nivel de datos que admite tanto Hadoop/NoSQL y SQL y escala elásticamente plataformas Java basadas en estándares.

Oracle cuenta con una solución completa e integrada con los mejores componentes de su clase para eventos de procesamiento, datos y análisis en tiempo real. Oracle puede ejecutar estos componentes de datos rápidos en diversas arquitecturas de implementación, incluidas puertas de enlace de dispositivos, entornos de grandes volúmenes de datos o sistemas de ingeniería.

ORACLE SOCIAL CLOUD

Oracle Social Cloud es un servicio en nube que ayuda a administrar y escalar la relación con clientes en canales de medios sociales. Oracle ha integrado los mejores componentes de la gestión de relaciones sociales (SRM) en su clase, escucha social, compromiso social, publicación social, aplicaciones de contenido social y análisis social en un servicio de nube unificado para proporcionarle la solución de SRM más completa del mercado.

Oracle puede conectar cada interacción que su cliente tiene con su marca. El objetivo es ayudar a brindar la mejor experiencia de cliente donde sea que el cliente use su marca. Oracle tiene la plataforma social que incluye todas las capacidades para administrar una estrategia de medios sociales, junto con integraciones con otras soluciones de Oracle para aprovechar el poder de los medios sociales en la empresa. Oracle Social Cloud está integrada con la automatización de marketing, servicios, ventas y sistemas de comercio de Oracle, lo que facilita agregar funcionalidades sociales a la inversión existente de Oracle.

Oracle Social Relationship Management (SRM) permite a las organizaciones habilitar socialmente la forma de hacer negocios, sin el coste y la complejidad de almacenes sociales. Es una estrategia para ser más atractivo y sensible a escala, escuchar y responder a la velocidad del desarrollo social, con una consistencia y transparencia que los clientes valorarán. Oracle SRM Incluye las siguientes soluciones sociales:

- Oracle Social Engagement & Monitoring Cloud Service

- Oracle Social Marketing Cloud Service

- Oracle Social Network

- Social Relationship Management Services

CAPÍTULO 5

BIG DATA CON HERRAMIENTAS DE MICROSOFT

MICROSOFT Y EL BIG DATA

Las organizaciones de hoy enfrentan desafíos crecientes para extraer valor del negocio a partir de los datos. En primer lugar, continúa el crecimiento desmedido de los datos que las organizaciones almacenan y pueden acceder. En segundo lugar, la complejidad de los datos aumenta conforme los clientes almacenan no solo datos estructurados en formato relacional, sino también datos no estructurados tales como archivos Word, PDF, imágenes, vídeos y datos geoespaciales. De hecho, analistas de la industria indican que más del 80% de los datos capturados son no estructurados. Finalmente, los clientes también se enfrentan a la velocidad de datos: las organizaciones que procesan datos transmitidos en tiempo real por sitios web requieren actualizar los datos en tiempo real para, por ejemplo, ofrecer el anuncio correcto o presentar las ofertas correctas a sus clientes.

Microsoft ha estado trabajando con Big Data desde mucho antes de que fuera una megatendencia. Por ejemplo, en Bing se analizan más de 100 petabytes de datos para ofrecer resultados de búsqueda de alta calidad. Microsoft proporciona una gama de soluciones para ayudar a los clientes a enfrentar los desafíos de Big Data.

La familia de soluciones de data warehouse de Microsoft cuenta con una amplia gama de productos como Microsoft SQL Server 2012 R2, Microsoft SQL Server 2008 R2, SQL Server Fast Track Data Warehouse, Business Data Warehouse y Microsoft SQL Server Parallel Data Warehouse, lo que ofrece una plataforma sólida y escalable para

almacenar y analizar datos en sistemas data warehouse. El sistema de Parallel Dataware House (PDW) ofrece a los clientes rendimiento de clase empresarial que maneja volúmenes masivos a más de 600 TB. También proporcionamos LINQ para HPC (High Performance Computing) con un tiempo de ejecución distribuido y un modelo de programación para computación técnica.

Además de las capacidades tradicionales, en Microsoft se está adoptando Apache Hadoop, como parte de un mapa de solución para cumplir con la visión de ofrecer soluciones de negocios para usuarios de todo tipo mediante la activación de nuevos tipos de datos de cualquier tamaño.

SOLUCIÓN BIG DATA DE MICROSOFT

La visión de Microsoft es proporcionar conocimiento de negocios a partir de cualquier tipo de datos, incluyendo conocimiento previamente escondido en datos no estructurados. Para lograr este objetivo, Microsoft ofrece distribuciones de Windows Server y Windows Azure basadas en Apache Fladoop, acelerando su adopción en las empresas.

Esta nueva distribución basada en Fladoop por Microsoft permite a los clientes obtener una visión de negocios sobre datos estructurados y no estructurados de cualquier tamaño y activar nuevos tipos de datos. Esta información extraída de Hadoop se puede combinar perfectamente con la plataforma de Business Intelligence de Microsoft.

Beneficios clave:

- Ampliar y facilitar el acceso a Hadoop por medio de una instalación y configuración sencilla, además de programación simplificada con JavaScript.
- Aportar una distribución Hadoop “lista para la empresa”, con mayor seguridad (integrada al Directorio Activo de Microsoft) y facilidad de administración (con una consola única con System Center).
- Facilidad para descubrir y aplicar información del negocio, mediante el uso de herramientas conocidas como PowerPivot para Excel, SQL Server Analysis y Reporting Services de SQL Server, se puede tener una integración y explotación de datos en poco tiempo y con grandes resultados.

La solución de Big Data de Microsoft también ofrece interoperabilidad con otras distribuciones de Hadoop, permitiendo obtener información de varias fuentes.

- Dos conectores de Hadoop: que permiten a los clientes mover datos fácilmente entre Hadoop y SQL Server o SQL Server Parallel Data Warehouse. Estos conectores ya se encuentran disponibles.

- Controlador Hive ODBC, además de Excel Hive Add-In: ofrecemos un nuevo controlador de Hive ODBC y un Hive de Excel complementario que permiten a los clientes mover datos directamente en Excel, o herramientas de BI de Microsoft tales como PowerPivot, para análisis.

ACCESO A HADOOP

Las técnicas de Big Data permiten transformar datos sin procesar en Información reveladora y almacena todos los datos (estructurados, sin estructurar y de streaming) desde dentro y fuera de la organización. Asimismo los usuarios empresariales tienen la posibilidad de obtener información reveladora con herramientas familiares para ellos, como Excel u Office 365. Sus decisiones serán mejores y más rápidas.

Con cada foto cargada, con cada tweet, con cada compra, con cada desplazamiento de GPS, se crean datos. En la actualidad, el 85% de los datos se genera automáticamente con sensores y dispositivos. Nos encontramos en el universo de los datos grandes (Big Data) y los resultados pueden ser grandes también, siempre que se disponga de las herramientas necesarias para controlarlo.

Microsoft trabaja con técnicas de Big Data utilizando el poder de Hadoop en las bases de datos principales para dar vida a datos estructurados y no estructurados a través de visualizaciones de datos 3D enriquecidas con las herramientas que más se utilizan en el negocio.

Con las soluciones Big Data de Microsoft, puede implementar un cluster de Hadoop y estar preparado en unos minutos para consultar y combinar datos relacionales y no relacionales, con los mismos conocimientos que emplea con SQL Server.

Además, cualquiera puede ser un experto de BI y usar Excel y Power BI para crear eficaces y hermosas visualizaciones o contar historias con datos.

Microsoft está comprometido en ampliar la accesibilidad y el uso de Hadoop para usuarios, desarrolladores y profesionales de TI. La nueva distribución de Windows Azure basada en Hadoop facilita las cosas al personal de TI, simplificando la experiencia de adquisición, instalación y configuración. Las mejoras en el empaquetamiento de Hadoop y sus herramientas permiten instalarlo y desplegarlo en cuestión de horas en lugar de días.

Los usuarios finales pueden utilizar el controlador Hive ODBC o Hive Add-in for Excel para analizar los datos de Hadoop usando herramientas

conocidas como Microsoft Excel y clientes de BI como PowerPivot para Excel.

Para los desarrolladores, Microsoft está invirtiendo en hacer que JavaScript sea un lenguaje de primera clase para Big Data, permitiendo escribir tareas Map/Reduce de alto desempeño en JavaScript. Además, nuestra consola de JavaScript permitirá a los usuarios crear con JavaScript desde su navegador tareas Map/Reduce, así como queries de Pig-Latin y Hive para ejecutarse en Hadoop. Este es el tipo de innovación que Microsoft espera contribuir como propuesta a la comunidad.

ADAPTACIÓN DE HADOOP PARA LA EMPRESA

Para acelerar su adopción en las empresas, Microsoft adaptará Hadoop para la empresa gracias a lo siguiente:

- Integración con Active Directory para manejo de seguridad.
- Mejoras en desempeño para grandes cantidades de datos.
- Integración con System Center para simplificar la administración.
- Integración con soluciones de Business Intelligence.

Adicionalmente, las opciones de despliegue de Windows Server y Windows Azure ofrecen gran flexibilidad y poder de elección:

- Libertad para elegir qué datos se mantienen in-house o en la nube.
- Menor costo total de propiedad (TCO) al desplegar Hadoop en la nube.
- Elasticidad para satisfacer la demanda, además de tener la opción de ampliar una solución de Hadoop in-house con Hadoop en Azure para satisfacer períodos de alta demanda.
- Mayor desempeño ya que nuestra solución permite a los clientes procesar datos más cerca de donde estos nacen, ya sea en sitio o en la nube.

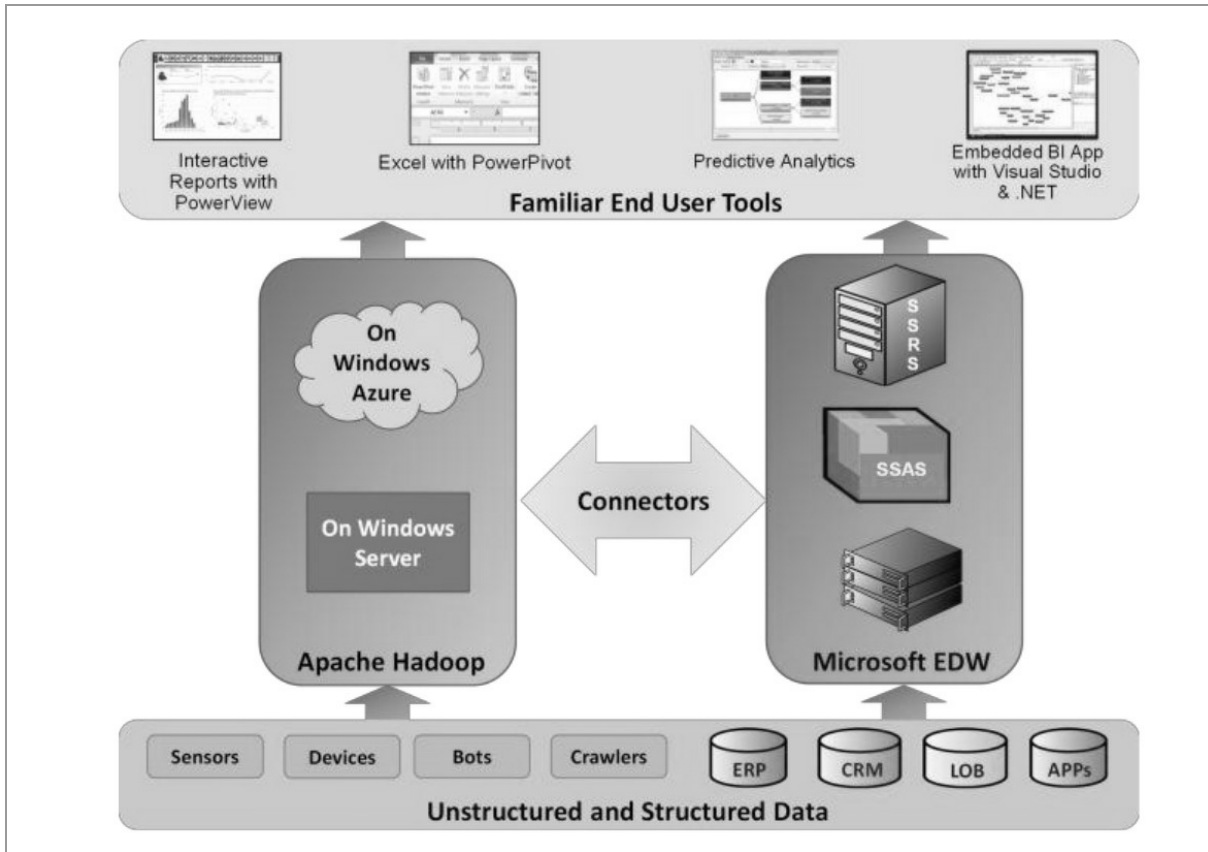
Todo esto se hace sin perder la compatibilidad con las herramientas existentes de Hadoop tales como Pig, Hive y Java. Nuestro objetivo es garantizar que las aplicaciones creadas en Apache Hadoop puedan migrar fácilmente a nuestra distribución para ejecutarse en Windows Azure o Windows Server.

APROVECHAMIENTO DE INFORMACIÓN

La solución de Big Data de Microsoft mejora significativamente la detección y aprovechamiento de información al permitir combinar datos relacionales de bases de datos con datos no estructurados de Hadoop. La distribución de Windows Server y Windows Azure de Microsoft basada en Hadoop permite:

- Analizar datos de Hadoop con herramientas familiares a los usuarios como Excel, gracias a su Hive Add-in para Excel.
- Reducir el tiempo de solución mediante la integración de herramientas de Hive y Microsoft BI como PowerPivot y PowerView.
- Construir soluciones corporativas de BI que incluyen datos de Hadoop, mediante la integración de Hive y herramientas líderes de BI como SQL Server Analysis Services y Reporting Services.
- El controlador Hive ODBC permite a los clientes mover datos desde Hive directamente en Microsoft Excel o Herramientas de BI herramientas como SQL Server Analysis Services, Reporting Services, PowerPivot y PowerView para una visualización de datos enriquecidos. Estas vistas pueden incorporar paneles de control para los tomadores de decisiones.

La figura siguiente ilustra el panorama de la solución de Big Data de Microsoft.



EL PAPEL DE SQL SERVER

Con la finalidad de trabajar con grandes volúmenes de datos, SQL Server permite manejar grupos de alta disponibilidad para clusters, mirroring, log shipping y diagnósticos, ofreciendo múltiples servidores secundarios en modo activo y múltiples bases de datos para tolerancia a fallos, escalabilidad bajo demanda y distribuyendo cargas de trabajo en servidores secundarios con una velocidad mejorada gracias a la tecnología de ColumnStore Index. De esta forma se facilita el trabajo relativo a Big Data.

Para ayudar la detección y aprovechamiento de grandes volúmenes de información, SQL Server dispone de una exploración y visualización de informes, reportes y datos mejorada agregando nuevas tecnologías como Power View que permite ver reportes de una forma gráfica y dinámica, así como Power Pivot que permite explotar millones de registros vía Excel, permitiendo al usuario un uso efectivo de sus datos en Excel y SharePoint Server. Se habilita el trabajo en la nube creando y escalando soluciones de negocios de forma rápida a través de servidores, nube privada o pública.

SQL Server ofrece la agilidad necesaria para crear y escalar soluciones de forma rápida que permitan resolver los desafíos y habilitar nuevas oportunidades de negocio desde el servidor a la nube pública o privada vinculando herramientas comunes para optimizar la productividad y desarrollo. Una de las ventajas con las herramientas de datos de SQL Server es escribir una vez, ejecutar en cualquier lugar ya sea en servidor o en la nube, no hay que reescribir el código.

SQL Server también habilita el trabajo In-Memory con la finalidad de desempeñar de forma óptima el tratamiento de los grandes volúmenes de datos necesario en las tareas de Big Data.

SQL Server está diseñado para soportar las cargas de trabajo más exigentes del mercado, ofreciendo los más altos niveles disponibilidad, desempeño, alta escalabilidad, seguridad y una experiencia mejorada para Inteligencia de Negocios y Big Data.

LOS ORÍGENES DE HADOOP. LA NUBE

Apache Hadoop es software de código abierto que sirve para almacenar y analizar cantidades masivas de datos, tanto estructurados como sin estructurar: terabytes o más de correo electrónico, lecturas del sensor, registros de servidor, fuentes de Twitter, señales de GPS... cualquier tipo de datos que pueda imaginar. Con Hadoop puede procesar grandes conjuntos de datos desordenados y obtener información y respuestas a partir de estos, de ahí la expectación creada.

Creado en 2005 por Mike Cafarella y Doug Cutting (que le puso el nombre del elefante de juguete de su hijo), Hadoop estaba destinado originalmente a datos de búsqueda en Internet. Hoy en día, es un proyecto de código abierto comunitario de Apache Software Foundation que se usa en todo tipo de organizaciones e industrias. Microsoft contribuye activamente a los esfuerzos de desarrollo de la comunidad.

Una de las razones del éxito de Hadoop es una simple cuestión económica. El procesamiento de conjuntos de datos de gran tamaño solía requerir equipos de alto rendimiento y hardware adicional especializado de precio elevado. Con Hadoop es posible realizar tareas de procesamiento confiable, escalable y distribuido en servidores estándar del sector, con capacidad para abordar petabytes de datos y sin que los presupuestos más reducidos supongan un problema. Hadoop también está diseñado para escalar de un único servidor a miles de máquinas, así como para detectar y controlar errores en la capa de aplicación para mayor confiabilidad.

Según ciertas estimaciones, hasta un 80% de los datos con los que las organizaciones trabajan hoy en día no vienen perfectamente clasificados en columnas y filas. Más bien se trata de una avalancha desordenada de correos electrónicos, fuentes de medios sociales, imágenes de satélites, señales de GPS, registros de servidor y otros archivos no relacionales sin estructurar. Hadoop puede administrar prácticamente cualquier archivo o formato (su otra gran ventaja), de manera que las organizaciones puedan plantear cuestiones que nunca creyeron posible.

Con Windows Azure, HDInsight y SQL Server, podemos recopilar,

analizar y generar inteligencia empresarial prácticamente en tiempo real a partir de datos Big Data recopilados de fuentes de medios sociales, señales de GPS y datos de sistemas gubernamentales.

Puede implementar Hadoop en un centro de datos tradicional en la oficina. Algunas empresas (incluida Microsoft) también ofrecen Hadoop como servicio en la nube. Una pregunta evidente sería: ¿por qué usar Hadoop en la nube? A continuación veremos por qué cada vez más organizaciones eligen esta opción.

El código abierto no significa que todo sea gratis. La implementación de Hadoop localmente requiere el uso de servidores, así como de expertos en Hadoop, para configurarlos, adaptarlos y mantenerlos. Un servicio en la nube permite poner en marcha un cluster de Hadoop en cuestión de minutos sin costo inicial alguno.

En la nube de Microsoft Azure solamente se paga por el almacenamiento y los procesos que se utilicen, cuando se usen. Puede arrancar un cluster de Hadoop, analizar los datos y cerrarlo una vez terminado para detener el contador.

Cree un cluster de Hadoop en cuestión de minutos y agregue nodos a petición. La nube ofrece a las organizaciones un tiempo de amortización inmediato.

HDINSIGHT

HDInsight de Microsoft Azure es un servicio basado 100% en Apache Hadoop en la nube de Azure. Ofrece todas las ventajas de Hadoop, junto con la capacidad de integración con Excel, los clusters de Hadoop locales y el ecosistema de software y servicios empresariales de Microsoft. Entre las ventajas podríamos destacar las siguientes:

- Escalar a petabytes a petición
- Procesar datos no estructurados y semiestructurados
- Desarrollar en Java, .NET, etc.
- No hay hardware para comprar o mantener
- Pagar solo por lo que se usa
- Poner en marcha un cluster de Hadoop en cuestión de minutos
- Visualizar los datos de Hadoop en Excel
- Integrar fácilmente en clusters de Hadoop locales

Escalamiento con total flexibilidad a petición

HDInsight es una distribución de Hadoop diseñada por la nube. Esto significa que HDInsight se ha diseñado para poder hacer frente a cualquier cantidad de datos, con la capacidad de escalar de terabytes a petabytes a petición. Puede arrancar un número de nodos cualquiera en cualquier momento. Solamente se cobra por los recursos de proceso y almacenamiento que realmente usa.

En campos como la auditoría, se conservan los datos durante siete años. Además, determinada información debe conservarse hasta 30 años. Con HDInsight, es posible almacenar más datos y consultarlos según sea necesario.

Análisis de datos semiestructurados, estructurados y no estructurados

Dado que es 100% Apache Hadoop, HDInsight puede procesar datos no estructurados o semiestructurados desde secuencias de clics web, medios sociales, registros de servidor, dispositivos, sensores, etc. Esto permite analizar nuevos conjuntos de datos que descubren nuevas posibilidades de negocio para impulsar a su organización.

Con una solución basada en SQL Server y el servicio HDInsight de Azure, es posible capturar datos escritos en lenguaje natural y usarlos para mejorar nuestros servicios.

Desarrollo en el lenguaje favorito. Hardware

HDInsight tiene extensiones de programación eficaces para lenguajes como CU, Java, .NET, etc. Así, en Hadoop, podrá usar el lenguaje de programación de su elección para crear, configurar, enviar y supervisar trabajos de Hadoop.

Con HDInsight, puede implementar Hadoop en la nube sin comprar nuevo hardware u otros costos iniciales. Además, la instalación y la configuración se realizan de forma rápida. Azure se encarga de todo. Puede iniciar su primer cluster en minutos.

Dado que con Windows Azure nos encontramos en una nube elástica, ya no es necesario preocuparse por la configuración de infraestructura o la posibilidad de ampliar la capacidad actual de nuestros centros de datos.

Excel para visualizar datos de Hadoop

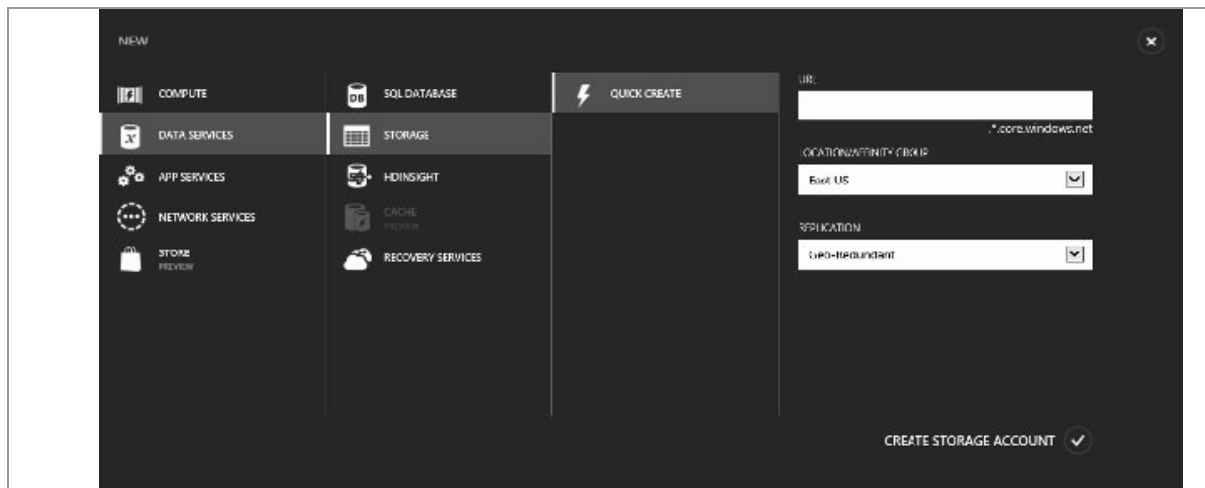
Dado que se integra con Excel, HDInsight le permite visualizar y analizar los datos de Hadoop de nuevas y convincentes formas en una herramienta conocida para sus usuarios finales. Desde Excel, los usuarios pueden seleccionar Azure HDInsight como origen de datos.

Los clusters locales de Hadoop y la nube

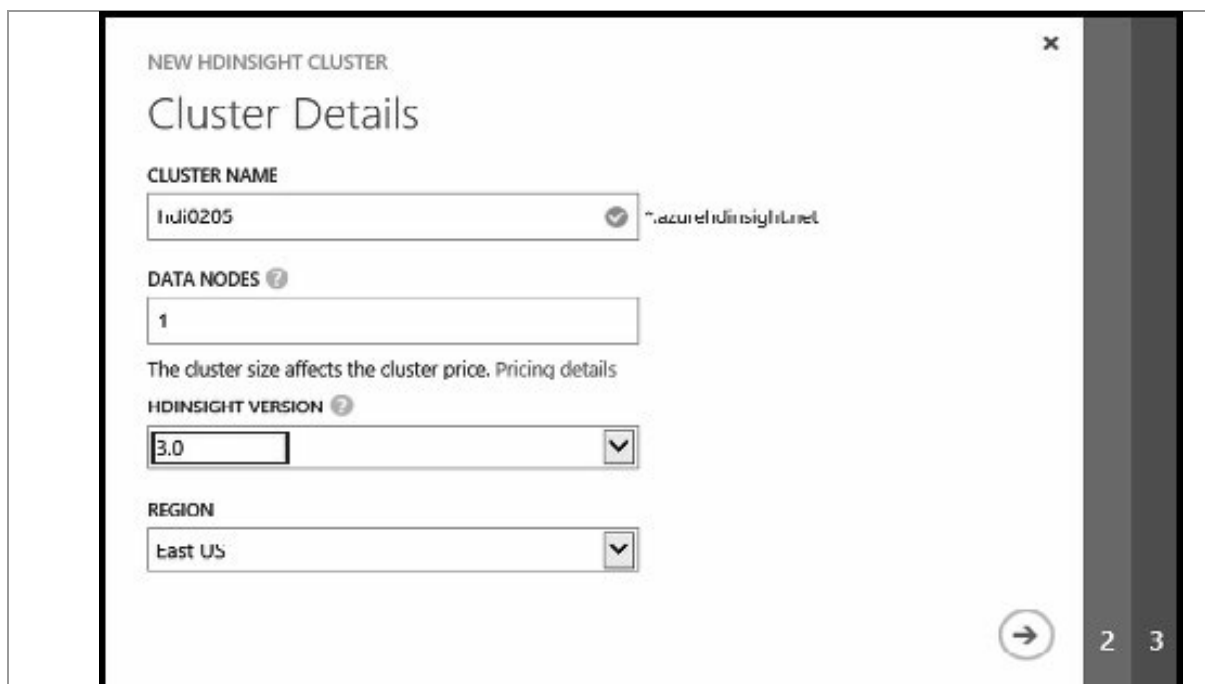
HDInsight también se integra con Hortonworks Data Platform, por lo que puede mover datos de Hadoop de un centro de datos de la oficina a la nube de Azure para escenarios de copia de seguridad, desarrollo y prueba

y ampliación a la nube. Con Microsoft Analytics Platform System, puede incluso realizar consultas a sus clusters de Hadoop locales y en la nube simultáneamente.

Podría sorprenderse de lo fácil que es preparar un cluster Hadoop en la nube. Con tan solo 15 clics, o en un tiempo aproximado de 20 minutos, puede crear un cluster Hadoop de HDInsight en la nube. Con HDInsight, puede tener incluso varios clusters Hadoop en el mismo conjunto de datos.



Haga clic en NUEVO en la esquina inferior izquierda y, después, haga clic en SERVICIOS DE DATOS, en HDINSIGHT y, por último, en CREACIÓN PERSONALIZADA.



HDInsight y Hbase

HDInsight también incluirá Apache HBase, una base de datos NoSQL columnar que se ejecuta sobre el sistema de archivos distribuido de Hadoop (HDFS). Esto le permitirá llevar a cabo procesos transaccionales grandes (OLTP) de datos no relacionales que permiten el uso de casos como tener sitios web interactivos o escritura de datos de sensor en el almacenamiento de blobs de Azure.

CONCEPTOS ESENCIALES EN AZURE HDINSIGHT

Azure HDInsight es un servicio que implementa y aprovisiona clusters Apache Hadoop® en la nube con el fin de proporcionar un marco de software que se ha diseñado para realizar tareas de administración, análisis y generación de informes en relación con datos de gran tamaño.

Datos de gran tamaño

Los datos se describen como “datos de gran tamaño” para indicar que se están recopilando en volúmenes de escala continua, a velocidades cada vez mayores, y para una variedad cada vez más amplia de formatos no estructurados y contextos semánticos variables. La recopilación de datos de gran tamaño no proporciona valor por sí mismo a una empresa. Para que los datos de gran tamaño proporcionen valor en forma de conocimientos o datos de inteligencia procesables, no solo es necesario formular las preguntas correctas y recopilar datos relevantes para los problemas en cuestión, sino que también los datos deben ser accesibles y se deben limpiar y analizar para presentarlos de forma útil, normalmente en combinación con datos de otras fuentes diversas que establezcan una perspectiva y un contexto en relación con lo que actualmente se denomina mashup.

Apache Hadoop

Apache Hadoop es un marco de software que facilita la administración y el análisis de datos de gran tamaño. El núcleo de Apache Hadoop proporciona almacenamiento de datos fiable con el sistema de archivos distribuidos Hadoop (HDFS, Hadoop Distributed File System) y un modelo de programación MapReduce sencillo para procesar y analizar, en paralelo, los datos almacenados en este sistema distribuido. HDFS utiliza la replicación de datos para resolver los problemas ocasionados por errores de hardware que surgen al implementar dichos sistemas altamente

distribuidos.

MapReduce

Para simplificar las complejidades del análisis de datos no estructurados procedentes de diversas fuentes, el modelo de programación MapReduce proporciona una abstracción del núcleo que sobrescribe el cierre del mapa y reduce las operaciones. El modelo de programación MapReduce visualiza todos los trabajos como cálculos sobre los conjuntos de datos que estén compuestos de pares clave-valor. Por lo tanto, tanto los archivos de entrada como los de salida deben contener conjuntos de datos que estén compuestos de pares clave-valor únicamente. Lo más importante de esta limitación es que los trabajos de MapReduce, en consecuencia, admiten composición.

Otros proyectos relacionados con Hadoop, por ejemplo Pig y Hive, se compilan sobre el sistema HDFS y el marco MapReduce. Proyectos como estos se utilizan para proporcionar una manera de administrar un cluster más sencilla que si se trabajara directamente con programas MapReduce. Pig, por ejemplo, permite escribir programas, mediante un lenguaje de procedimientos denominado Pig Latín, que se compilan en programas MapReduce en el cluster. También proporciona controles fluidos para administrar el flujo de datos. Hive es una infraestructura de almacenes de datos que proporciona una abstracción de datos en tablas para los archivos que estén almacenados en un cluster, con el fin de poder realizar consultas mediante instrucciones de tipo SQL en un lenguaje declarativo que se denomina HiveQL.

HDInsight

Azure HDInsight pone Apache Hadoop a disposición de los usuarios como servicio en la nube. Proporciona el marco de software HDFS/MapReduce y los proyectos relacionados, como Pig y Hive, en un entorno más sencillo, escalable y rentable.

Uno de los aspectos más eficientes que se han introducido en HDInsight es el modo de administración y almacenamiento de los datos. HDInsight utiliza el almacenamiento de blobs de Azure como sistema de

archivos predeterminado. El almacenamiento de blobs y el sistema HDFS son sistemas de archivos distintos que se han optimizado, respectivamente, para el almacenamiento de datos y la ejecución de cálculos sobre esos datos.

El almacenamiento de blobs de Azure proporciona una opción de almacenamiento altamente escalable y disponible, económica, duradera e intercambiable para los datos que se deben procesar con HDInsight.

Los clusters de Hadoop que HDInsight implementa en HDFS están optimizados para ejecutar tareas de cálculo de MapReduce sobre esos datos.

Los clusters de HDInsight se implementan en Azure, en nodos de proceso, para ejecutar tareas MapReduce y los usuarios pueden descartarlos una vez que estas tareas se hayan completado. El mantenimiento de datos en los clusters de HDFS después de haber completado los cálculos supondría un alto coste para el almacenamiento de estos datos. El almacenamiento de blobs es una solución de almacenamiento en Azure robusta y de uso general. Por lo tanto, el almacenamiento de datos en blobs permite que los clusters que se utilizan para realizar cálculos se puedan eliminar de forma segura sin pérdida de datos del usuario. No obstante, el almacenamiento en blobs no solo es una solución de bajo coste. Proporciona una interfaz de sistema de archivos HDFS completa que ofrece una experiencia inigualable a los clientes mediante la habilitación de un conjunto de componentes completo en el ecosistema Hadoop para procesar directamente (de forma predeterminada) los datos que administra.

HDInsight utiliza Azure PowerShell para configurar, ejecutar y procesar posteriormente los trabajos de Hadoop. HDInsight también proporciona un conector Sqoop que se puede utilizar para importar datos de una base de datos SQL de Azure al sistema HDFS o exportar datos a una base de datos SQL de Azure desde el sistema HDFS.

Microsoft Power Query para Excel está disponible para importar datos de Azure HDInsight o cualquier sistema HDFS a Excel. Este complemento mejora la experiencia de Business Intelligence (BI) de autoservicio en Excel al simplificar la detección de datos y el acceso a los mismos para una gran variedad de orígenes de datos. Además de Power Query, Microsoft Hive ODBC Driver está disponible para integrar herramientas

de BI, como Excel, SQL Server Analysis Services y Reporting Services, con el fin de facilitar y simplificar el análisis de datos de extremo a extremo.

EL ECOSISTEMA HADOOP EN AZURE

HDInsight ofrece un marco para implementar la solución basada en la nube de Microsoft para el tratamiento de datos de gran tamaño. Este ecosistema federado administra y analiza grandes cantidades de datos para explotar las capacidades de procesamiento en paralelo del modelo de programación MapReduce. Las tecnologías Hadoop compatibles con Apache que se pueden utilizar con HDInsight se presentan una a una y se describen brevemente en esta sección.

HDInsight proporciona implementaciones de Hive y Pig para integrar el procesamiento de datos y las capacidades de almacenamiento. La solución para datos de gran tamaño de Microsoft se integra con las herramientas de BI de Microsoft, como SQL Server Analysis Services, Reporting Services, PowerPivot y Excel. De este modo es posible realizar un procesamiento de BI sencillo con los datos que almacena y administra HDInsight en el almacenamiento de blobs.

También se pueden descargar y utilizar con HDInsight otras tecnologías compatibles con Apache y tecnologías similares que forman parte del ecosistema Hadoop y que se han compilado para su ejecución sobre los clusters de Hadoop. Entre ellas se incluyen tecnologías de código abierto, como Sqoop, que integra HDFS con el almacenamiento de datos relacionales.

Pig

Pig es una plataforma de alto nivel para el procesamiento de datos de gran tamaño en clusters de Hadoop. Pig consta de un lenguaje de flujo de datos, denominado Pig Latín, compatible con la escritura de consultas en grandes conjuntos de datos y un entorno que ejecuta programas desde una consola. Los programas escritos en Pig Latín constan de una serie de transformaciones de conjuntos de datos que se convierten en una serie de programas MapReduce. Las abstracciones de Pig Latín proporcionan estructuras de datos más detalladas que las de MapReduce y hacen en

Hadoop lo que SQL hace en los sistemas de administración de bases de datos relacionales (RDBMS). Pig Latín es completamente extensible. Las funciones definidas por el usuario (UDF), escritas en Java, Python, Ruby, CU o JavaScript, se pueden llamar para personalizar cada fase de ruta de procesamiento al elaborar el análisis.

Hive

Hive es un almacén de datos distribuidos que administra datos almacenados en un sistema HDFS. Es el motor de consultas de Hadoop. Hive proporciona a los analistas expertos en SQL una interfaz parecida a SQL y un modelo de datos relacionales. Hive utiliza un lenguaje denominado HiveQL, un dialecto de SQL. Hive, al igual que Pig, es una abstracción situada en un nivel superior a MapReduce y, al ejecutarse, traduce las consultas en una serie de trabajos de MapReduce. Los escenarios de Hive se parecen mucho a los de RDBMS, por lo que son apropiados para el uso con datos más estructurados. Para los datos no estructurados, Pig es una mejor opción.

Sqoop

Sqoop es una herramienta que transfiere grandes cantidades de datos entre Hadoop y bases de datos relacionales, como SQL u otros almacenes de datos estructurados, de la forma más eficiente posible. Puede utilizar Sqoop para importar datos de almacenes de datos estructurados externos al sistema HDFS o a sistemas similares como Hive. Sqoop también puede extraer datos de Hadoop y exportar los datos extraídos a bases de datos relacionales externas, almacenes de datos empresariales o cualquier otro tipo de almacén de datos estructurados.

Herramientas de Business Intelligence y conectores

Las herramientas de Business Intelligence habituales (como Excel, PowerPivot, SQL Server Analysis Services y Reporting Services) recuperan, analizan y generan informes de datos integrados en HDInsight con el complemento Power Query o Microsoft Hive ODBC Driver.

ESCENARIOS DE DATOS DE GRAN TAMAÑO EN HDINSIGHT

Un análisis puntual, por lotes, de un conjunto de datos no estructurado y completo que esté almacenado en nodos de Azure y que no requiera actualizarse periódicamente es un escenario ejemplar que ofrece un caso de aplicación de HDInsight.

Estas condiciones se aplican a una gran variedad de actividades empresariales, científicas y gubernamentales. Entre ellas se incluyen, por ejemplo, la supervisión de cadenas de suministro al por menor, los modelos de comercialización sospechosos en el ámbito financiero, los modelos de demanda de los servicios públicos, la calidad del aire y el agua en redes de sensores para el control medioambiental o los modelos delictivos en áreas metropolitanas.

Las tecnologías HDInsight (y las tecnologías Hadoop en general) son principalmente adecuadas para el tratamiento de grandes cantidades de datos registrados o archivados que no requieren actualizarse periódicamente una vez escritos y que se leen con frecuencia, normalmente para realizar un análisis completo. Este escenario complementa el de los datos que se tratan de una forma más adecuada con un RDBMS, con el que es necesario utilizar una menor cantidad de datos (gigabytes en lugar de petabytes) y que se deben actualizar continuamente o consultar en busca de datos específicos del conjunto total de datos. RDBMS funciona especialmente bien con datos estructurados que estén organizados y almacenados según un esquema fijo. MapReduce funciona bien con datos no estructurados y sin un esquema predefinido porque es capaz de interpretar esos datos durante el procesamiento.

INTRODUCCIÓN AL USO DE HDINSIGHT DE AZURE

HDInsight pone Apache Hadoop a disposición de los usuarios como servicio en la nube. De esta forma, el marco de software de MapReduce está disponible en un entorno de Azure más sencillo, escalable y rentable. HDInsight también ofrece una solución rentable para la administración y el almacenamiento de datos mediante el almacenamiento de blobs de Azure.

En este ejemplo se aprovisionará un cluster de HDInsight a través del Portal de administración de Azure, se enviará un trabajo de MapReduce de Hadoop mediante PowerShell y, después, se importarán los datos resultantes del trabajo de MapReduce en Excel para su análisis.

Además de poner HDInsight de Azure a disposición de los usuarios, Microsoft también ha lanzado el emulador de HDInsight para Azure, anteriormente conocido como Microsoft HDInsight Developer Preview. Este producto está destinado a los desarrolladores y, como tal, solo admite implementaciones de un solo nodo.

Requisitos previos

Antes de empezar el trabajo debe disponer de lo siguiente:

Una suscripción de Azure.

Un equipo que ejecute Windows 8, Windows 7, Windows Server 2012 o Windows Server 2008 R2. Este equipo se usará para enviar trabajos de MapReduce.

Office Professional Plus 2013, Office 365 Pro Plus, Excel 2013 Standalone u Office Professional Plus 2010.

Configuración de un entorno local para ejecutar PowerShell

Existen varias formas de enviar trabajos de MapReduce a HDInsight.

En este trabajo se usará Azure PowerShell. Para instalar Azure PowerShell, ejecute el instalador de plataforma web de Microsoft. Cuando se le solicite, haga clic en **Ejecutar**, a continuación, en **Instalar** y después siga las instrucciones. Para obtener más información, consulte Instalación y configuración de Azure PowerShell.

Los cmdlets de PowerShell requieren información de la suscripción para utilizarse a fin de administrar los servicios.

Para conectarse a la suscripción con Azure AD

Abra la consola de Azure PowerShell siguiendo las instrucciones del apartado Instalación de Azure PowerShell.

Ejecute el siguiente comando: Add-AzureAccount

En la ventana, escriba la dirección de correo electrónico y la contraseña asociadas a su cuenta. Azure autentica y guarda las credenciales y, a continuación, cierra la ventana.

Aprovisionamiento de un cluster de HDInsight

El proceso de aprovisionamiento de HDInsight requiere el uso de una cuenta de almacenamiento de Azure como sistema de archivos predeterminado. Dicha cuenta debe estar ubicada en el mismo centro de datos que los recursos de proceso de HDInsight. En la actualidad, solo pueden aprovisionarse clusters de HDInsight en los siguientes centros de datos:

Sudeste asiático

Europa del Norte

Europa occidental

Este de EE. UU.

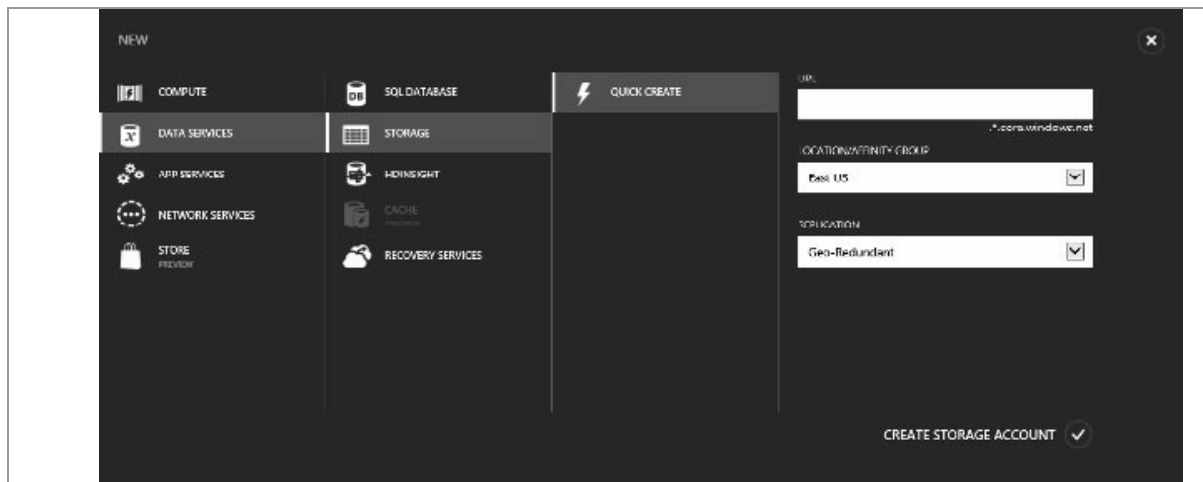
Oeste de EE. UU.

Debe elegir uno de estos cinco centros de datos para la cuenta de almacenamiento de Azure.

Para crear una cuenta de Almacenamiento de Azure

- Inicie sesión en el Portal de administración de Azure. 2. Haga clic

en **NEW** en la esquina inferior izquierda, seleccione **DATA SERVICES, STORAGE** y, a continuación, haga clic en **QUICK CREATE**.



- Escriba los detalles de **URL, LOCATION** y **REPLICATION** y, a continuación, haga clic en **CREATE STORAGE ACCOUNT**. No se admiten grupos de afinidad. La nueva cuenta de almacenamiento aparecerá en la lista de almacenamiento.

- Espere hasta que la característica **STATUS** de la nueva cuenta de almacenamiento cambie a **Online**.

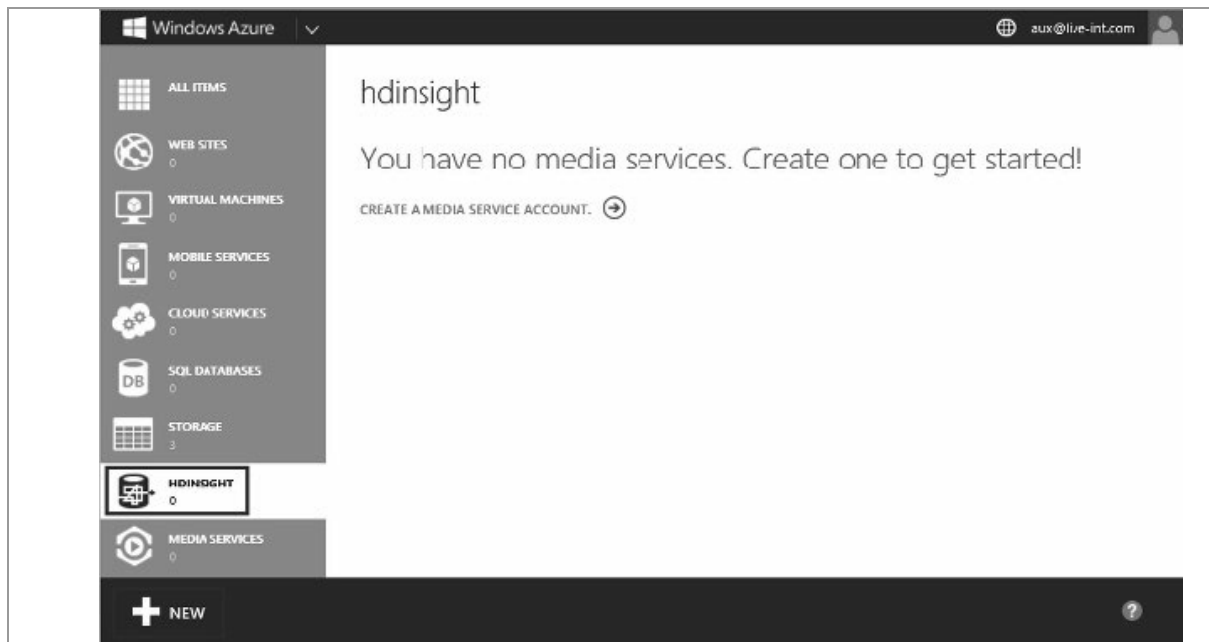
- Haga clic en la nueva cuenta de almacenamiento en la lista para seleccionarla.

- Haga clic en **MANAGE ACCESS KEYS** en la parte inferior de la página.

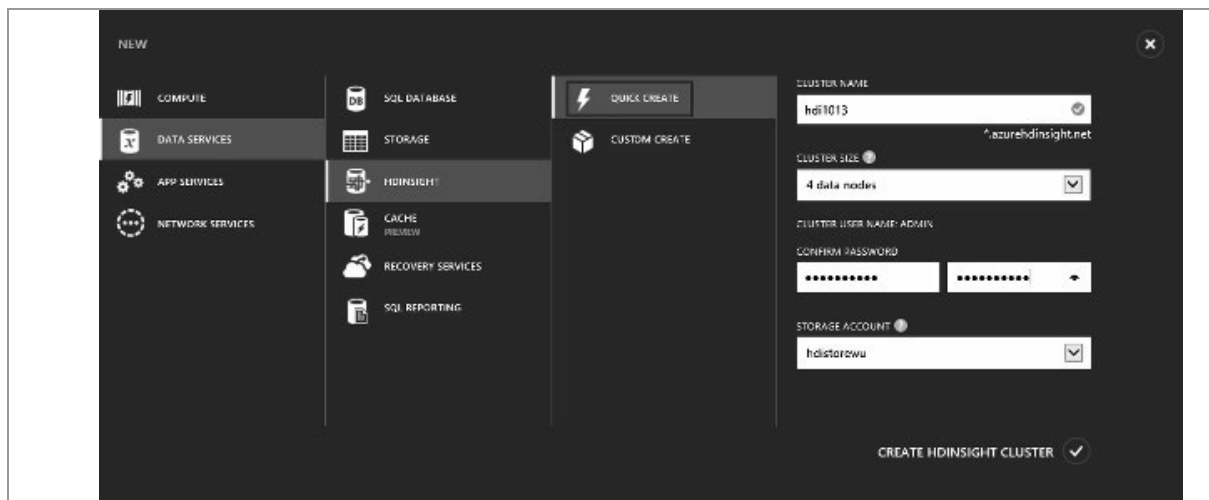
- Tome nota de los valores de los campos **STORAGE ACCOUNT NAME** y **PRIMARY ACCESS KEY**. Los necesitará más adelante en el tutorial.

Para aprovisionar un cluster de HDInsight:

- Inicie sesión en el Portal de administración de Azure.
- Haga clic en la opción **HDInsight**, que aparece a la izquierda, para ver el estado de los clusters en la cuenta. En la captura de pantalla siguiente no hay ningún cluster de HDInsight.



- Haga clic en **NEW** en la esquina inferior izquierda y, después, en **Data Services, HDInsight** y en **Quick Create**.



- Escriba o seleccione los siguientes valores:

NOMBRE

VALOR

Cluster Name

Nombre del cluster.

Cluster Size

Número de nodos de datos que desea implementar. El valor predeterminado es 4, aunque también hay disponibles clusters de nodos de datos de 8, 16 y 32 en el menú desplegable. Al usar la opción Custom

Create puede especificarse un número cualquiera de nodos de datos. Hay disponible información sobre las tarifas de facturación para varios tamaños de cluster. Haga clic en el símbolo ? justo encima del cuadro desplegable y siga el vínculo del elemento emergente.

Password (administrador de clusters)

La contraseña del administrador de la cuenta. El nombre del usuario del cluster se especifica como "admin" de forma predeterminada al usar la opción Quick Create. Esta solo puede cambiarse mediante el asistente Custom Create. El campo de contraseña debe tener un mínimo de 10 caracteres y contener una letra mayúscula, una minúscula, un número y un carácter especial.

Storage Account

Seleccione la cuenta de almacenamiento que ha creado en el cuadro desplegable. > [WACOM.NOTE] > Una vez elegida una cuenta de almacenamiento, esta no se puede cambiar. En caso de quitarla, el cluster dejará de estar disponible para su uso. La ubicación del cluster de HDInsight será la misma que la de la cuenta de almacenamiento.

- Haga clic en **Create HDInsight Cluster** en la parte inferior derecha. Una vez completado el proceso de aprovisionamiento, la columna de estado mostrará **Running**.

Ejecución de un trabajo WordCount de MapReduce

Ahora ya se ha aprovisionado un cluster de HDInsight. El paso siguiente consiste en ejecutar un trabajo de MapReduce para contar las palabras de un archivo de texto.

Para ejecutar un trabajo de MapReduce se requieren los siguientes elementos:

Un programa de MapReduce. En este tutorial se usará el ejemplo WordCount incluido con la distribución del cluster de HDInsight, por lo que no tendrá que escribir uno propio. Se encuentra en */example/jars/hadoop-examples.jar*. Para obtener instrucciones acerca de cómo escribir su propio trabajo de MapReduce, consulte Desarrollo de programas MapReduce de Java para HDInsight.

Un archivo de entrada. Se usará */example/data/gutenberg/davinci.txt*

como este tipo de archivo.

Una carpeta de archivo de salida. Se usará */example/data/WordCountOutput* como la carpeta mencionada. El sistema creará la carpeta en caso de que esta no exista.

El esquema URI para obtener acceso a los archivos del almacenamiento de blobs es:

```
wasb [s] ://<containername>@<storageaccountname>.blob.core.windows.net/<path>
```

El esquema URI proporciona tanto acceso no cifrado con el prefijo *wasb:* como acceso SSL cifrado con *WASBS*. Se recomienda usar *wasbs* siempre que sea posible, incluso al obtener acceso a los datos que residen en el mismo centro de datos de Azure.

Dado que HDInsight usa un contenedor de almacenamiento de blobs como sistema de archivos predeterminado, puede consultar los archivos y directorios de dicho sistema de archivos mediante rutas de acceso relativas o absolutas.

Por ejemplo, para obtener acceso al archivo *hadoop-examples.jar*, puede usar una de las siguientes opciones:

- *wasb://<containername>@<storageaccountname>.blob.core.windows.net/example/jars/hadoop-examples.jar*
- *wasb:///example/jars/hadoop-examples.jar*
- */example/jars/hadoop-examples.jar*

El uso del prefijo *wasb://* en las rutas de acceso de estos archivos es necesario para indicar que el almacenamiento de blobs de Azure se está usando para los archivos de entrada y salida. El directorio de salida asume una ruta de acceso relativa predeterminada a la carpeta *wasb:///user/*.

Para ejecutar el ejemplo WordCount

- Abra **Azure PowerShell**. Para obtener instrucciones acerca de cómo abrir la ventana de la consola de Azure PowerShell, consulte *Instalación y configuración de Azure PowerShell*.
- Ejecute los siguientes comandos para establecer las variables:

```
$subscriptionName = "<SubscriptionName>"
```

```
$clusterName = "<HDInsightClusterName>"
```

- Ejecute los siguientes comandos para crear una definición del trabajo de MapReduce:

```
# Define the MapReduce job $wordCountJobDefinition = New-
```

```
AzureHDInsightMapReduceJobDefinition -JarFile
```

```
"wasb:///example/jars/hadoop-examples.jar" -ClassName
```

```
"wordcount" -Arguments
```

```
"wasb:///example/data/gutenberg/davinci.txt",
```

```
"wasb:///example/data/WordCountOutput"
```

El archivo `hadoop-examples.jar` se incluye con la distribución del cluster de HDInsight. Existen dos argumentos para el trabajo de MapReduce. El primero es el nombre del archivo de origen y, el segundo, la ruta de acceso del archivo de salida. El archivo de origen se incluye con la distribución del cluster de HDInsight y la ruta de acceso del archivo de salida se creará en tiempo de ejecución.

- Ejecute el siguiente comando para enviar el trabajo de MapReduce:

```
# Submit the job
```

```
Select-AzureSubscription $subscriptionName
```

```
$wordCountJob = Start-AzureHDInsightJob -Cluster
```

```
$clusterName -JobDefinition $wordCountJobDefinition
```

Además de la definición de trabajo de MapReduce, también debe proporcionarse el nombre del cluster de HDInsight en el que se desea ejecutar el trabajo de MapReduce.

Start-AzureHDInsightJob es una llamada no sincronizada. Para comprobar la finalización del trabajo, use el cmdlet *Wait-AzureHDInsightJob*.

- Ejecute el siguiente comando para comprobar la finalización del trabajo de MapReduce:

```
Wait-AzureHDInsightJob -Job $wordCountJob -
```

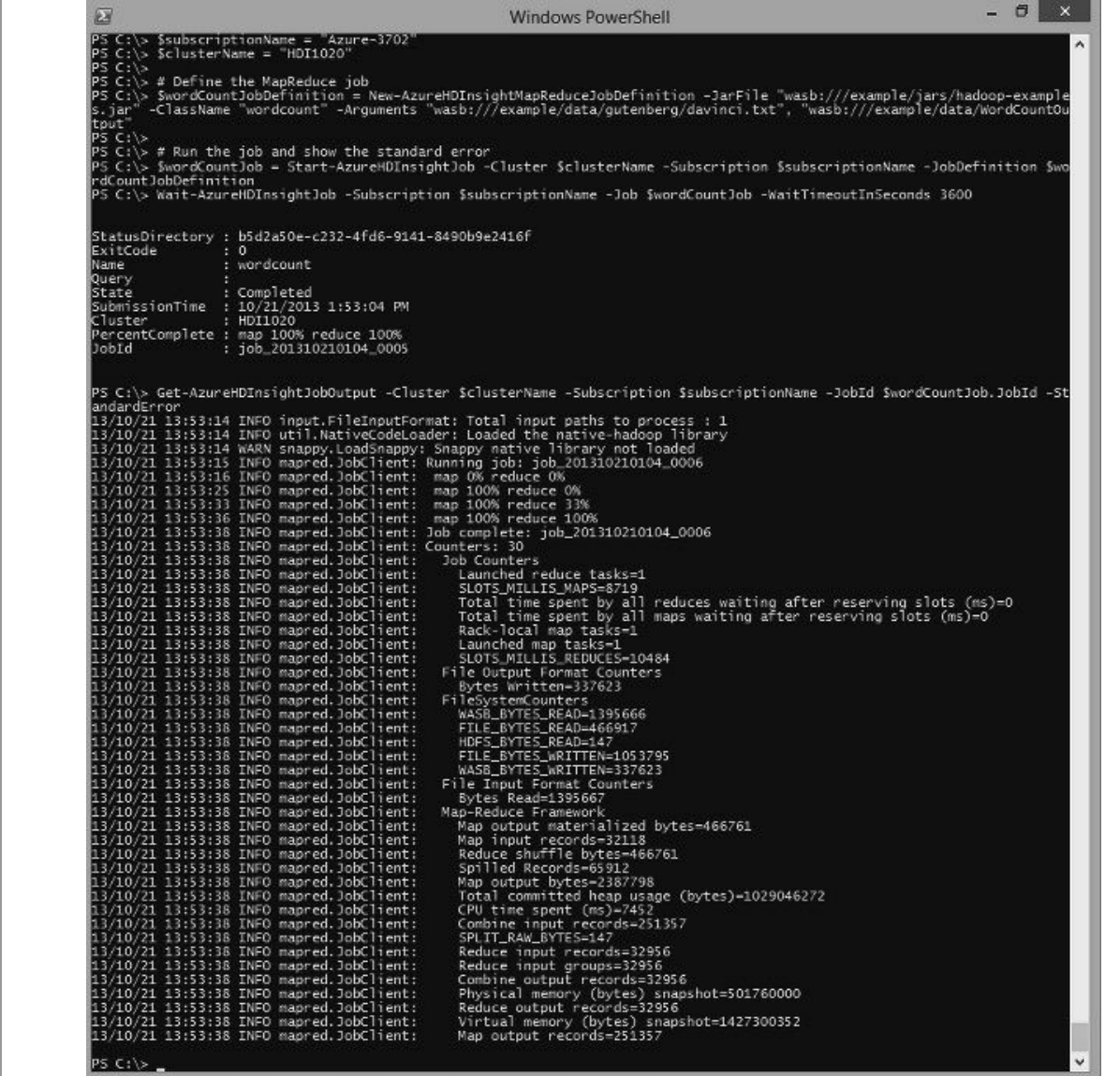
```
WaitTimeoutInSeconds 3600
```

- Ejecute el siguiente comando para comprobar posibles errores al ejecutar el trabajo de MapReduce:

Get the job output

```
Get-AzureHDInsightJobOutput -Cluster $clusterName -  
JobId $wordCountJob.JobId -StandardError
```

En la captura de pantalla de la página siguiente se muestra el resultado de una ejecución correcta. En caso contrario, aparecerán mensajes de error.



```
Windows PowerShell  
PS C:\> $subscriptionName = "Azure-3702"  
PS C:\> $clusterName = "HDI1020"  
PS C:\>  
PS C:\> # Define the MapReduce job  
PS C:\> $wordCountJobDefinition = New-AzureHDInsightMapReduceJobDefinition -JarFile "wasb:///example/jars/hadoop-example  
s.jar" -ClassName "wordcount" -Arguments "wasb:///example/data/gutenberg/davinci.txt", "wasb:///example/data/WordCountOut  
put"  
PS C:\>  
PS C:\> # Run the job and show the standard error  
PS C:\> $wordCountJob = Start-AzureHDInsightJob -Cluster $clusterName -Subscription $subscriptionName -JobDefinition $wo  
rdCountJobDefinition  
PS C:\> wait-AzureHDInsightJob -Subscription $subscriptionName -Job $wordCountJob -WaitTimeoutInSeconds 3600  
  
StatusDirectory : b5d2a50e-c232-4fd6-9141-8490b9e2416f  
ExitCode         : 0  
Name            : wordcount  
Query           :  
State           : Completed  
SubmissionTime  : 10/21/2013 1:53:04 PM  
Cluster         : HDI1020  
PercentComplete: map 100% reduce 100%  
JobId           : job_201310210104_0005  
  
PS C:\> Get-AzureHDInsightJobOutput -Cluster $clusterName -Subscription $subscriptionName -JobId $wordCountJob.JobId -St  
andardError  
13/10/21 13:53:14 INFO input.FileInputFormat: Total input paths to process : 1  
13/10/21 13:53:14 INFO util.NativeCodeLoader: Loaded the native-hadoop library  
13/10/21 13:53:14 WARN snappy.LoadSnappy: Snappy native library not loaded  
13/10/21 13:53:15 INFO mapred.JobClient: Running job: job_201310210104_0006  
13/10/21 13:53:16 INFO mapred.JobClient: map 0% reduce 0%  
13/10/21 13:53:25 INFO mapred.JobClient: map 100% reduce 0%  
13/10/21 13:53:33 INFO mapred.JobClient: map 100% reduce 33%  
13/10/21 13:53:36 INFO mapred.JobClient: map 100% reduce 100%  
13/10/21 13:53:38 INFO mapred.JobClient: Job complete: job_201310210104_0006  
13/10/21 13:53:38 INFO mapred.JobClient: Counters: 30  
13/10/21 13:53:38 INFO mapred.JobClient:   Job Counters  
13/10/21 13:53:38 INFO mapred.JobClient:     Launched reduce tasks=1  
13/10/21 13:53:38 INFO mapred.JobClient:     SLOTS_MILLIS_MAPS=8719  
13/10/21 13:53:38 INFO mapred.JobClient:     Total time spent by all reduces waiting after reserving slots (ms)=0  
13/10/21 13:53:38 INFO mapred.JobClient:     Total time spent by all maps waiting after reserving slots (ms)=0  
13/10/21 13:53:38 INFO mapred.JobClient:     Rack-local map tasks=1  
13/10/21 13:53:38 INFO mapred.JobClient:     Launched map tasks=1  
13/10/21 13:53:38 INFO mapred.JobClient:     SLOTS_MILLIS_REDUCES=10484  
13/10/21 13:53:38 INFO mapred.JobClient:   File Output Format Counters  
13/10/21 13:53:38 INFO mapred.JobClient:     Bytes written=337623  
13/10/21 13:53:38 INFO mapred.JobClient:   FileSystemCounters  
13/10/21 13:53:38 INFO mapred.JobClient:     WASB_BYTES_READ=1395666  
13/10/21 13:53:38 INFO mapred.JobClient:     FILE_BYTES_READ=466917  
13/10/21 13:53:38 INFO mapred.JobClient:     HDFS_BYTES_READ=147  
13/10/21 13:53:38 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=1053795  
13/10/21 13:53:38 INFO mapred.JobClient:     WASB_BYTES_WRITTEN=337623  
13/10/21 13:53:38 INFO mapred.JobClient:   File Input Format Counters  
13/10/21 13:53:38 INFO mapred.JobClient:     Bytes Read=1395667  
13/10/21 13:53:38 INFO mapred.JobClient:   Map-Reduce Framework  
13/10/21 13:53:38 INFO mapred.JobClient:     Map output materialized bytes=466761  
13/10/21 13:53:38 INFO mapred.JobClient:     Map input records=32118  
13/10/21 13:53:38 INFO mapred.JobClient:     Reduce shuffle bytes=466761  
13/10/21 13:53:38 INFO mapred.JobClient:     Spilled Records=65912  
13/10/21 13:53:38 INFO mapred.JobClient:     Map output bytes=2387798  
13/10/21 13:53:38 INFO mapred.JobClient:     Total committed heap usage (bytes)=1029046272  
13/10/21 13:53:38 INFO mapred.JobClient:     CPU time spent (ms)=7452  
13/10/21 13:53:38 INFO mapred.JobClient:     Combine input records=251357  
13/10/21 13:53:38 INFO mapred.JobClient:     SPLIT_RAW_BYTES=147  
13/10/21 13:53:38 INFO mapred.JobClient:     Reduce input records=32956  
13/10/21 13:53:38 INFO mapred.JobClient:     Reduce input groups=32956  
13/10/21 13:53:38 INFO mapred.JobClient:     Combine output records=32956  
13/10/21 13:53:38 INFO mapred.JobClient:     Physical memory (bytes) snapshot=501760000  
13/10/21 13:53:38 INFO mapred.JobClient:     Reduce output records=32956  
13/10/21 13:53:38 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=1427300352  
13/10/21 13:53:38 INFO mapred.JobClient:     Map output records=251357  
PS C:\>
```

Para recuperar los resultados del trabajo de MapReduce

- Abra **Azure PowerShell**. 2. Ejecute los siguientes comandos para crear una carpeta C:\Tutorials y cambie el directorio a la carpeta:
mkdir \Tutorials
cd \Tutorials

El directorio predeterminado de Azure Powershell es el siguiente: `C:\Windows\System32\WindowsPowerShell\v1.0`. De forma predeterminada, no tiene permiso de escritura en esta carpeta. Debe cambiar el directorio a una carpeta en la que tenga permiso de escritura.

- Establezca tres variables en los comandos siguientes y, a continuación, ejecútelos:

```
$subscriptionName = "<SubscriptionName>"
```

```
$storageAccountName = "<StorageAccountName>"
```

```
$containerName = "<ContainerName>"
```

La cuenta de almacenamiento de Azure es la misma que se creó anteriormente en el tutorial. Esta se usa para hospedar el contenedor de blobs utilizado como sistema de archivos predeterminado del cluster de HDInsight. El nombre del contenedor de almacenamiento de blobs suele coincidir con el del cluster de HDInsight, a menos que se especifique un nombre distinto durante el aprovisionamiento del cluster.

- Ejecute los siguientes comandos para crear un objeto de contexto de almacenamiento de Azure:

```
# Create the storage account context object
Select-AzureSubscription $subscriptionName
$storageAccountKey = Get-AzureStorageKey
$storageAccountName | %{ $_.Primary }
$storageContext = New-AzureStorageContext -
StorageAccountName $storageAccountName -
StorageAccountKey $storageAccountKey
```

Select-AzureSubscription se usa para establecer la suscripción actual en caso de tener varias y no usar la suscripción predeterminada.

- Ejecute el siguiente comando para descargar el resultado del trabajo de MapReduce del contenedor de blobs a la estación de trabajo:

```
# Download the job output to the workstation
Get-AzureStorageBlobContent -Container $ContainerName -
Blob example/data/WordCountOutput/part-r-OOOOO -Context
$storageContext -Forcé
```

example/data/WordCountOutput es la carpeta de salida especificada al

ejecutar el trabajo de MapReduce. `part-r-00000` es el nombre de archivo predeterminado para el resultado del trabajo de MapReduce. El archivo se descargará a la misma estructura de carpetas de la carpeta local. Por ejemplo, en la captura de pantalla siguiente, la carpeta actual es la carpeta raíz `C`. El archivo se descargará en la carpeta `C:\example\data\WordCountOutput`.

- Ejecute el siguiente comando para imprimir el archivo de salida del trabajo de MapReduce:

```
cat ./example/data/WordCountOutput/part-r-00000 | findstr "there"
```

```

PS C:\> $subscriptionName = "Azure-3702"
PS C:\> $storageAccountName = "hdistorewu"
PS C:\> $containerName = "hdi1010"
PS C:\> ## Select the current subscription
PS C:\> Select-AzureSubscription $subscriptionName
PS C:\>
PS C:\> ## Create the storage account context object
PS C:\> $storageAccountKey = Get-AzureStorageKey $storageAccountName | %[_Primary ]
PS C:\> $storageContext = New-AzureStorageContext -StorageAccountName $storageAccountName -StorageAccountKey $storageAccountKey
PS C:\> Get-AzureStorageBlobContent -Container $containerName -Blob example/data/WordCountOutput/part-r-00000 -Context $storageContext -Force

Container Uri: https://hdistorewu.blob.core.windows.net/hdi1010
Name      BlobType Length      ContentType      LastModified      SnapshotTime
-----
example/data/WordC... BlockBlob 337623      application/octet... 10/11/2013 5:43:47...

PS C:\> cat ./example/data/WordCountOutput/part-r-00000 | findstr "there"
Maffeo;--there 1
feathered 1
gathered 3
lathered 1
smothered 1
there 361
there, 24
there. 16
there; 5
thereabout, 1
thereabouts; 1
thereby 2
therefore 83
therefore, 8
therein 2
therein. 2
withered 1
PS C:\>

```

El trabajo de MapReduce genera un archivo denominado `part-r-00000` con las palabras y los recuentos. El script usa el comando `findstr` para enumerar todas las palabras que contienen `"there"`.

Si se abre en el Bloc de notas `./example/data/WordCountOutput/part-r-00000`, resultado de varias líneas de un trabajo de MapReduce, observará que los saltos de línea no se representan correctamente. Se espera que esto sea así.

Conexión a las herramientas de inteligencia empresarial de Microsoft

El complemento Power Query de Excel se puede usar para exportar los resultados de HDInsight a Excel y aplicarles las herramientas de Microsoft

Business Intelligence (BI) para seguir procesándolos o mostrar los resultados. Al crear un cluster de HDInsight, se creó un contenedor predeterminado con el mismo nombre que dicho cluster en la cuenta de almacenamiento asociada durante la creación. Este se rellena automáticamente con un conjunto de archivos. Uno de estos archivos es una tabla de Hive de ejemplo. En esta sección se mostrará cómo importar en Excel los datos incluidos en dicha tabla para poder verlos y procesarlos aún más.

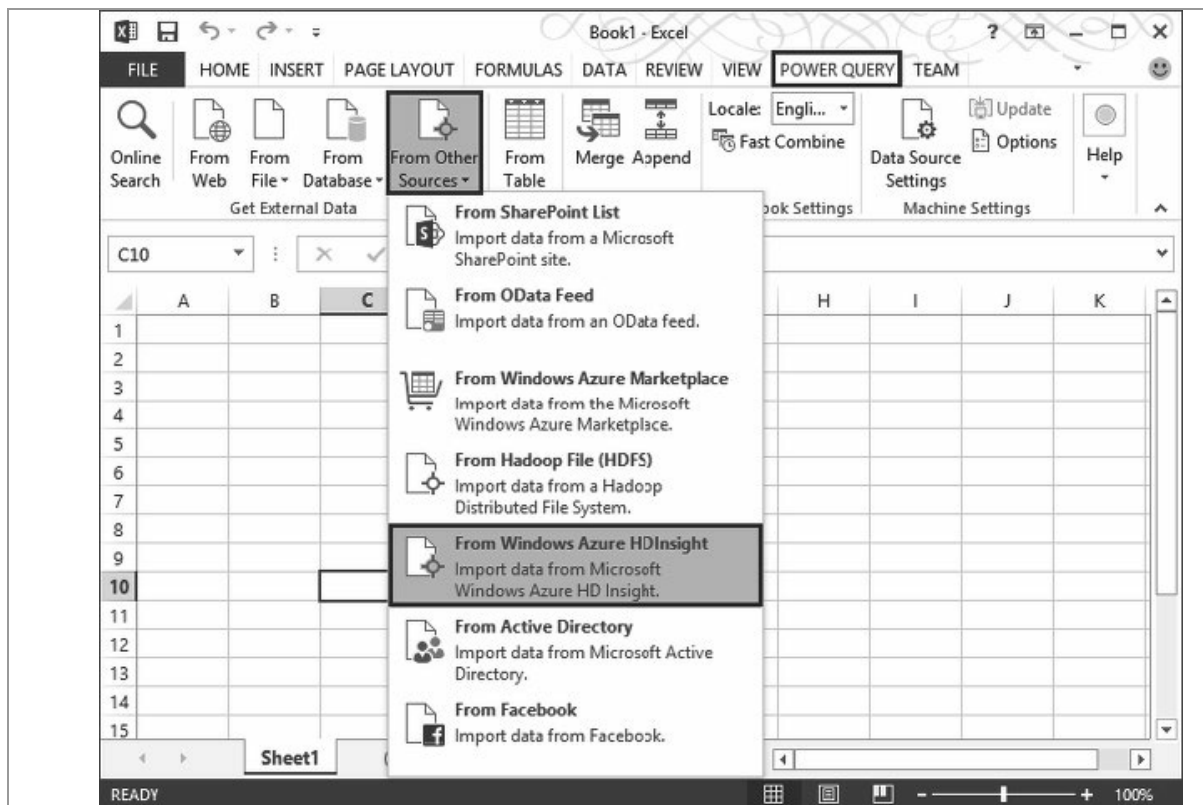
Para completar esta parte del tutorial, debe tener instalado Excel 2010 o 2013. Aquí importaremos la tabla de Hive que se incluye con HDInsight.

Para descargar Microsoft Power Query para Excel

Descargue Microsoft Power Query para Excel en el Centro de descarga de Microsoft e instálelo.

Para importar datos de HDInsight:

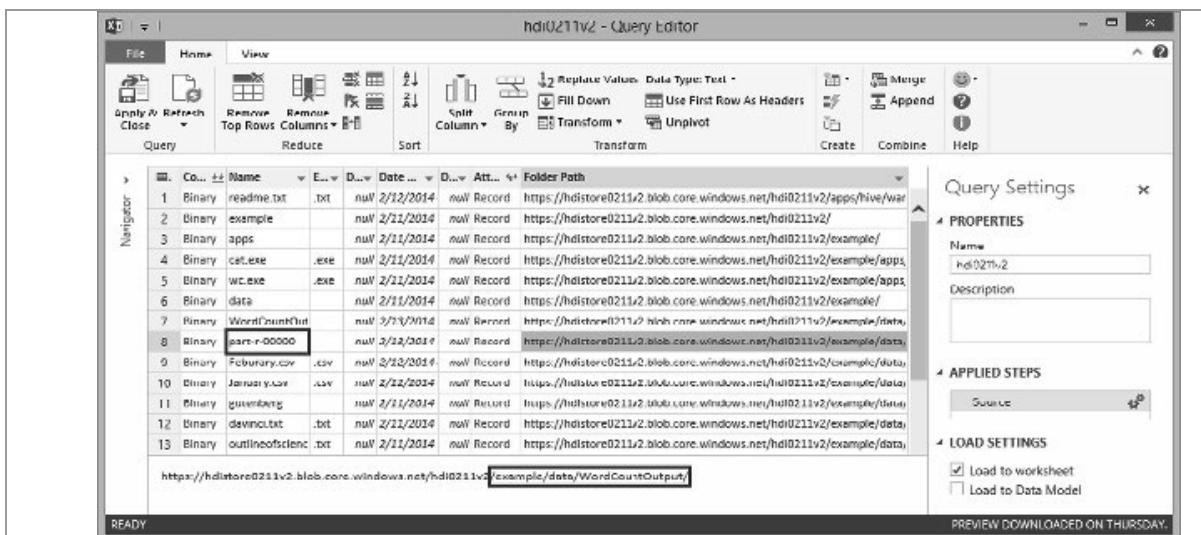
- Abra Excel y cree un libro en blanco. 2. Haga clic en el menú **Power Query**, en **From Other Sources** y, a continuación, en **From Azure HDInsight**.



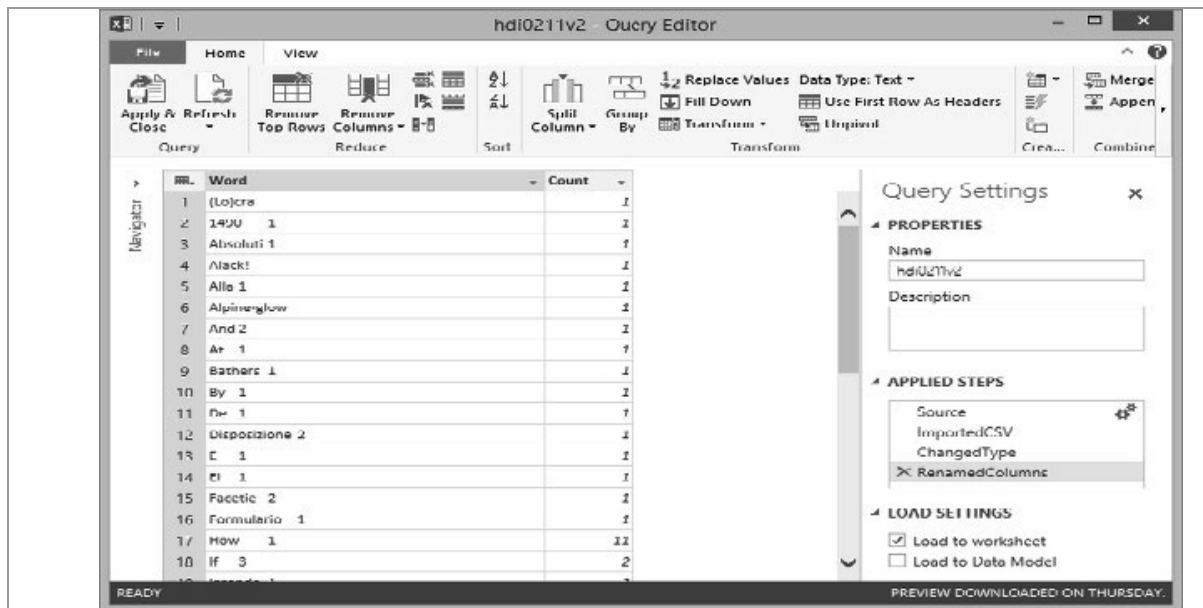
- En **Account Name**, escriba el nombre de la cuenta de

almacenamiento de blobs de Azure asociada con su cluster y, a continuación, haga clic en **Aceptar**. Esta es la cuenta de almacenamiento creada previamente en el tutorial.

- En **Account Key**, escriba la clave de la cuenta de almacenamiento de blobs de Azure y, a continuación, haga clic en **Guardar**.
- En el panel de navegación de la derecha, haga doble clic en el nombre del contenedor de almacenamiento de blobs. De forma predeterminada, el nombre del contenedor es el mismo que el del cluster.
- Busque **part-r-00000** en la columna **Name** (la ruta de acceso es `.../example/data/WordCountOutput`) y, a continuación, haga clic en **Binary** a la izquierda de **part-r-00000**.



- Haga clic con el botón secundario en **Column1.1** y seleccione **Rename**.
- Cambie el nombre a **Word**. 9. Repita el proceso para cambiar el nombre de **Column1.2** a **Count**.



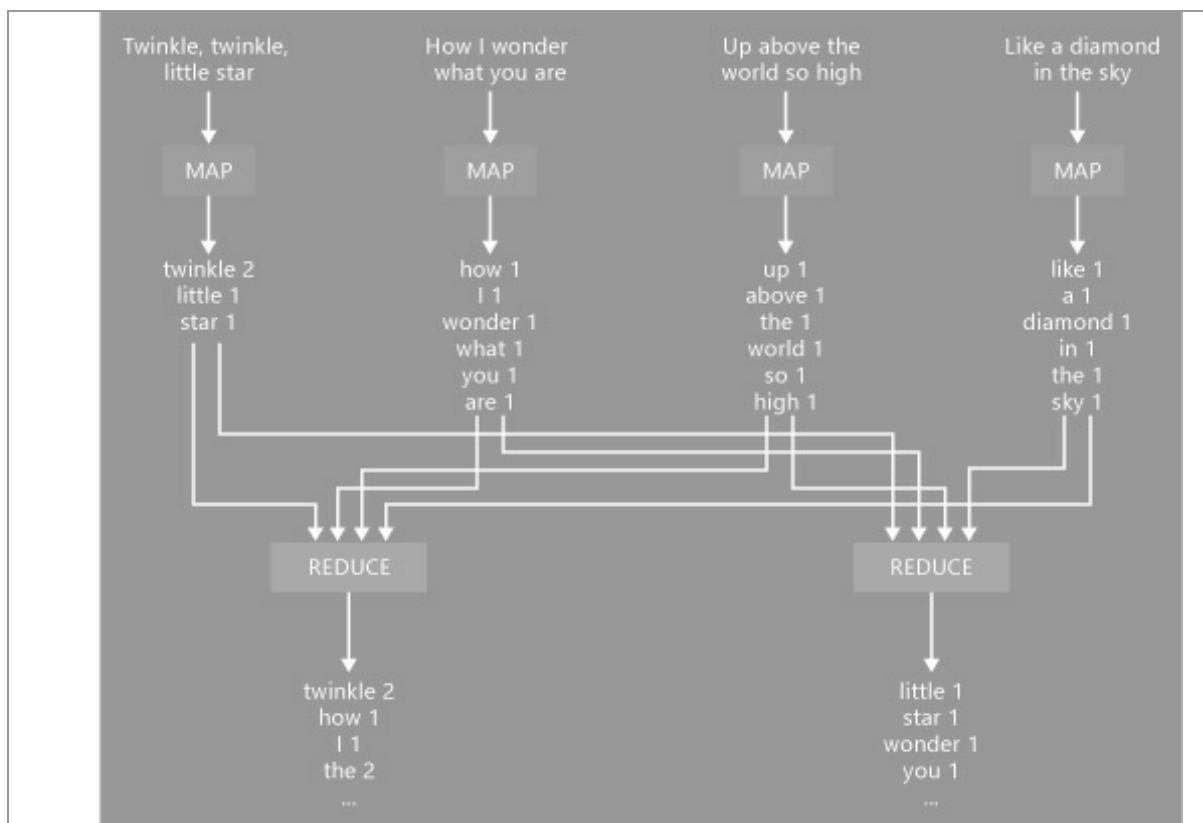
- Haga clic en **Apply & Close** en la esquina superior izquierda. La consulta importa la tabla de Hive en Excel.

USO DE MAPREDUCE CON HDINSIGHT

MapReduce de Hadoop es un marco de software para escribir aplicaciones que procesan enormes cantidades de datos. En este tutorial, usará Azure PowerShell desde la estación de trabajo para enviar un programa de MapReduce que cuenta las ocurrencias de palabras en un texto para un cluster de HDInsight. El programa de recuento de palabras está escrito en Java y el programa incluye el cluster de HDInsight.

Escenario

En el siguiente diagrama se ilustra cómo funciona MapReduce para el escenario de recuento de palabras:



La salida del trabajo de MapReduce es un conjunto de pares clave-valor. La clave es una cadena que especifica una palabra y el valor es un entero que especifica el número total de ocurrencias de esa palabra en el

texto. Esto se hace en dos etapas:

- El asignador utiliza cada línea del texto de entrada como una entrada y la desglosa en palabras. Emite un par clave-valor cada vez que se detecta una palabra, seguido por un 1. La salida se ordenará antes de enviarla al reductor.
- Posteriormente, el reductor suma estos recuentos individuales de cada palabra y emite un solo par clave-valor que contiene la palabra seguido de la suma de sus ocurrencias.

Para ejecutar un trabajo de MapReduce se requieren los siguientes elementos:

- Un programa de MapReduce. En este tutorial, usará la muestra de recuento de palabras incluida con los clusters de HDInsight, por lo que no tendrá que escribir una propia. Se encuentra en */example/jars/hadoop-examples.jar*. El nombre del archivo es *hadoop-mapreduce-examples.jar* en la versión 3.0 de clusters de HDInsight. Para obtener instrucciones acerca de cómo escribir su propio trabajo de MapReduce, consulte Desarrollo de programas MapReduce de Java para HDInsight.
- Un archivo de entrada. Se usará */example/data/gutenberg/davinci.txt* como este tipo de archivo. Para obtener información acerca de cómo cargar archivos, consulte Carga de datos en HDInsight.
- Una carpeta de archivo de salida. Se usará */example/data/WordCountOutput* como la carpeta mencionada. El sistema creará la carpeta en caso de que esta no exista. El trabajo de MapReduce producirá un error si la carpeta existe. Si desea ejecutar el trabajo de MapReduce por segunda vez, asegúrese de eliminar la carpeta de salida o especifique otra carpeta de salida.

Ejecución de la muestra con Azure PowerShell

- Abra Azure PowerShell. Para obtener instrucciones acerca de cómo abrir la ventana de la consola de Azure PowerShell, consulte Instalación y configuración de Azure PowerShell.
- Ajuste las dos variables en los comandos siguientes y, a

continuación, ejecútelos:

```
$subscriptionName = "<SubscriptionName>" # Nombre de la suscripción a Azure  
$clusterName = "<ClusterName>" # Nombre del  
cluster de HDInsight
```

- Ejecute el siguiente comando y proporcione información de la cuenta de Azure:

```
Add-AzureAccount
```

- Ejecute los siguientes comandos para crear una definición del trabajo de MapReduce:

```
# Defina el trabajo de MapReduce  
$wordCountJobDefinition = New-  
AzureHDInsightMapReduceJobDefinition -JarFile  
"wasb:///example/jars/hadoop-examples.jar" -ClassName  
"wordcount" -Arguments  
"wasb:///example/data/gutenberg/davinci.txt",  
"wasb:///example/data/WordCountOutput"
```

El archivo `hadoop-examples.jar` se incluye con la distribución del cluster de HDInsight. Existen dos argumentos para el trabajo de MapReduce. El primero es el nombre del archivo de origen y, el segundo, la ruta de acceso del archivo de salida. El archivo de origen se incluye con la distribución del cluster de HDInsight y la ruta de acceso del archivo de salida se creará en tiempo de ejecución.

- Ejecute el siguiente comando para enviar el trabajo de MapReduce:

```
# Envíe el trabajo  
Select-AzureSubscription $subscriptionName  
$wordCountJob = Start-AzureHDInsightJob -Cluster  
$clusterName -JobDefinition $wordCountJobDefinition |  
Wait-AzureHDInsightJob -WaitTimeoutInSeconds 3600
```

Además de la definición de trabajo de MapReduce, también debe proporcionar el nombre del cluster de HDInsight en el que desea ejecutar el trabajo de MapReduce y las credenciales. `Start-AzureHDInsightJob` es una llamada no sincronizada. Para comprobar la finalización del trabajo, use el cmdlet `Wait-AzureHDInsightJob`.

- Ejecute el siguiente comando para comprobar la finalización del trabajo de MapReduce:

```
Wait-AzureHDInsightJob -Job $swordCountJob -
WaitTimeoutInSeconds 3600
```

- Ejecute el siguiente comando para comprobar posibles errores al ejecutar el trabajo de MapReduce:

```
# Obtenga la salida del trabajo
Get-AzureHDInsightJobOutput -Cluster $clusterName -
JobId
$swordCountJob.JobId -StandardError
```

Para recuperar los resultados del trabajo de MapReduce:

- Abra **Azure PowerShell**.
- Ejecute el siguiente comando para cambiar al directorio a c:\root:

```
cd \
```

El directorio predeterminado de Azure Powershell es *C:\Windows\System32\WindowsPowerShell\v1.0*. De forma predeterminada, no tiene permiso de escritura en esta carpeta. Debe cambiar el directorio al directorio raíz *C:* o a una carpeta en la que tiene permiso de escritura.

- Establezca tres variables en los comandos siguientes y, a continuación, ejecútelos:

```
$subscriptionName = "<SubscriptionName>" # Nombre
de la suscripción a Azure
$storageAccountName = "<StorageAccountName>" # Nombre
de la cuenta de almacenamiento de Azure $containerName =
"<ContainerName>" #
Nombre del contenedor de almacenamiento de blobs
```

La cuenta de almacenamiento de Azure es la misma que se creó anteriormente en el tutorial. Esta se usa para hospedar el contenedor de blobs utilizado como sistema de archivos predeterminado del cluster de HDInsight. El nombre del contenedor de almacenamiento de blobs suele coincidir con el del cluster de HDInsight, a menos que se especifique un nombre distinto durante el aprovisionamiento del cluster.

- Ejecute los siguientes comandos para crear un objeto de contexto de almacenamiento de Azure:

```
# Seleccione la suscripción actual
Select-AzureSubscription $subscriptionName
```

```
# Crear el objeto de contexto de la cuenta de almacenamiento
$storageAccountKey = Get-AzureStorageKey
$storageAccountName | %{ $_.Primary }
$storageContext = New-AzureStorageContext -
StorageAccountName $storageAccountName -
StorageAccountKey $storageAccountKey
```

Select-AzureSubscription se usa para establecer la suscripción actual en caso de tener varias y no usar la suscripción predeterminada.

- Ejecute el siguiente comando para descargar el resultado del trabajo de MapReduce del contenedor de blobs a la estación de trabajo:

```
# Download the job output to the workstation
Get-AzureStorageBlobContent -Container $ContainerName -
Blob example/data/WordCountOutput/part-r-00000 -Context $storage
Context -Force
```

La carpeta */example/data/WordCountOutput* es la carpeta de salida especificada cuando se ejecuta el trabajo de MapReduce. *part-r-00000* es el nombre de archivo predeterminado para la salida de trabajos de MapReduce. El archivo se descargará en la misma estructura de carpetas de la carpeta local. Por ejemplo, en la captura de pantalla siguiente, la carpeta actual es la carpeta raíz C. El archivo se descargará en la carpeta *C:\example\data\WordCountOutput*.

- Ejecute el siguiente comando para imprimir el archivo de salida del trabajo de MapReduce:

```
cat ./example/data/WordCountOutput/part-r-00000 | findstr "there"
```

El trabajo de MapReduce genera un archivo denominado *part-r-00000* con las palabras y los recuentos. El script usa el comando *findstr* para enumerar todas las palabras que contienen *"there"*.

Tenga en cuenta que los archivos de salida de un trabajo de MapReduce son inmutables. Por lo tanto, si vuelve a ejecutar esta muestra tendrá que cambiar el nombre del archivo de salida.

El código Java para el programa de recuento de palabras de MapReduce

El siguiente es el código fuente del programa MapReduce de Java de recuento de palabras:

```
package org.apache.hadoop.examples;
import java.io.IOException;
import java.util.StringTokenizer;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import
org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import
org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.util.GenericOptionsParser;
public class WordCount {
    public static class TokenizerMapper
        extends Mapper<Object, Text, Text, IntWritable>{
        private final static IntWritable one = new
IntWritable (1) ;
        private Text word = new Text();
        public void map(Object key, Text value, Context
context
            ) throws IOException,
InterruptedException {
            StringTokenizer itr = new
StringTokeni zer(value.toString());
            while (itr.hasMoreTokens()) {
```



```

        word.set(itr.nextToken());
        context.write(word, one);
    }
}
}

public static class IntSumReducer
    extends
Reducer<Text,IntWritable,Text,IntWritable> {
    private IntWritable result = new IntWritable ();
    public void reduce(Text key, Iterable<IntWritable>
values,
        Context context
        ) throws IOException,
InterruptedException {
        int sum = 0;
        for (IntWritable val : values) {
            sum += val.get();
        }
        result.set(sum);
        context.write(key, result);
    }
}

public static void main(String[] args) throws
Exception {
    Configuration conf = new Configuration();
    String[] otherArgs = new GenericOptionsParser(conf,
args).getRemainingArgs();
    if (otherArgs.length != 2) {
        System.err.println("Usage: wordcount <in> <out>");
        System.exit(2);
    }
    Job job = new Job(conf, "word count");
    job.setJarByClass(WordCount.class);

```

```
        job.setMapperClass(TokenizerMapper.class);
        job.setCombinerClass(IntSumReducer.class);
        job.setReducerClass(IntSumReducer.class);
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);
        FileInputFormat.addInputPath(job, new
Path(otherArgs[0]) );
        FileOutputFormat.setOutputPath(job, new
Path(otherArgs[1]));
        System.exit(job.waitForCompletion(true) ? 0 : 1) ;
    }
}
```

CARGA DE DATOS EN HDINSIGHT

HDInsight de Azure proporciona un sistema de archivos distribuidos Hadoop (HDFS) completo a través del servicio de almacenamiento de blobs de Azure. Se ha diseñado como extensión HDFS para proporcionar una experiencia sin igual a los clientes al habilitar un conjunto completo de componentes del ecosistema Hadoop para poder operar directamente en los datos que administra. Tanto el almacenamiento de blobs de Azure como HDFS son sistemas de archivos diferentes que se han optimizado para el almacenamiento de datos y el cálculo en ellos.

Los clusters de HDInsight de Azure se implementan normalmente para ejecutar trabajos de MapReduce y se anulan una vez terminados. El mantenimiento de datos en los clusters de HDFS después de haber completado los cálculos supondría un alto coste para el almacenamiento de estos datos. El almacenamiento de blobs de Azure tiene una excelente disponibilidad, es altamente escalable, cuenta con una gran capacidad, un bajo coste y la opción de almacenamiento que se puede compartir para los datos que se van a procesar usando HDInsight. Almacenar los datos en un blob permite que los clusters de HDInsight que se usan para los cálculos se lancen de forma segura y sin perder los datos.

Se puede tener acceso al almacenamiento de blobs de Azure mediante AzCopy, Azure PowerShell, Biblioteca de cliente de almacenamiento de Azure para .NET o a través de las herramientas del explorador. Estas son algunas de las herramientas disponibles:

- Explorador de almacenamiento de Azure

- Cloud Storage Studio 2

- CloudXplorer

- Azure Explorer

- Azure Explorer PRO

Requisitos previos

Tenga en cuenta los siguientes requisitos antes de empezar este artículo:

Un cluster de HDInsight de Azure.

Carga de datos en el almacenamiento de blobs usando AzCopy

AzCopy es una utilidad de línea de comandos que se ha diseñado para simplificar la tarea de transferir datos a una cuenta de almacenamiento de Azure y desde ella. Puede usarla como herramienta independiente o incorporarla a una aplicación que ya existe. [Descarga de AzCopy](#).

La sintaxis de AzCopy es la siguiente:

```
AzCopy [filePattern [filePattern...]] [Options]
```

Carga de datos en el almacenamiento de blobs usando Azure PowerShell

Azure PowerShell es un potente entorno de scripting que puede usar para controlar y automatizar la implementación y la administración de sus cargas de trabajo en Azure. Puede usar Azure PowerShell para cargar datos en el almacenamiento de blobs, a fin de que los trabajos de MapReduce puedan procesar los datos. Para obtener información acerca de cómo configurar su estación de trabajo para que ejecute Azure PowerShell, consulte [Instalación y configuración de Azure PowerShell](#).

Para cargar un archivo local en el almacenamiento de blobs

- Abra la ventana de la consola de Azure PowerShell como se indica en [Instalación y configuración de Azure PowerShell](#).
- Configure los valores de las cinco primeras variables del script siguiente:

```
$subscriptionName = ""
```

```
$storageAccountName = ""
```

```
$containerName = ""
```

```
$fileName = ""
```

```
$blobName = ""
```

```
# Obtener la clave de la cuenta de almacenamiento
```

```
Select-AzureSubscription $subscriptionName
```

```
$storageaccountkey = get-azurestoragekey
```

```
$storageAccountName | %{$_.Primary}
```

```
# Crear el objeto de contexto de almacenamiento
```

```
$destContext = New-AzureStorageContext -
```

```
StorageAccountName $storageAccountName -
```

```
StorageAccountKey $storageaccountkey
```

```
# Copiar el archivo desde la estación de trabajo local al contenedor de blobs
```

```
Set-AzureStorageBlobContent -File $fileName -Container
```

```
$containerName -Blob $blobName -context $destContext
```

- Pegue el script en la ventana de la consola de Azure PowerShell para ejecutarlo.

Los contenedores del almacenamiento de blobs almacenan los datos como pares de clave/valor y no hay jerarquía de directorios. No obstante, el carácter "/" se puede usar en el nombre de la clave para que parezca que el archivo está almacenado dentro de una estructura de directorios. Por ejemplo, la clave de un blob puede ser input/log1.txt. No hay directorios "input", pero dada la presencia del carácter "/" en el nombre de la clave, parece la ruta de un archivo. En el script anterior, puede dar una estructura de carpeta al archivo configurando la variable \$blobname. Por ejemplo, \$blobname="myfolder/myfile.txt".

Al usar las herramientas de Azure Explorer, puede que vea algunos archivos de 0 bytes. Estos archivos tienen dos propósitos:

- o En caso de que haya carpetas vacías, sirven como marcador de la existencia de la carpeta. El almacenamiento de blobs es lo suficientemente inteligente como para saber que si hay un blob que se llama foo/bar, es porque hay una carpeta llamada foo. Pero si quiere tener una carpeta vacía llamada foo, entonces la única forma de

indicarlo es teniendo este archivo especial de 0 bytes dentro.

o Contienen metadatos especiales que necesita el sistema de archivos Hadoop, concretamente los permisos y los propietarios de las carpetas.

Carga de datos en el almacenamiento de blobs usando el explorador de almacenamiento de Azure

El explorador de almacenamiento de Azure es una herramienta útil para inspeccionar y modificar los datos de su almacenamiento de Azure. Se trata de una herramienta gratuita que se puede descargar de <http://azurestorageexplorer.codeplex.com/>.

Antes de usar la herramienta, debe saber el nombre y la clave de la cuenta de almacenamiento de Azure. Para obtener instrucciones acerca de cómo conseguir esta información, consulte la sección “Visualización, copia y generación de claves de acceso al almacenamiento” de Administración de cuentas de almacenamiento.

- Ejecute el explorador de almacenamiento de Azure.



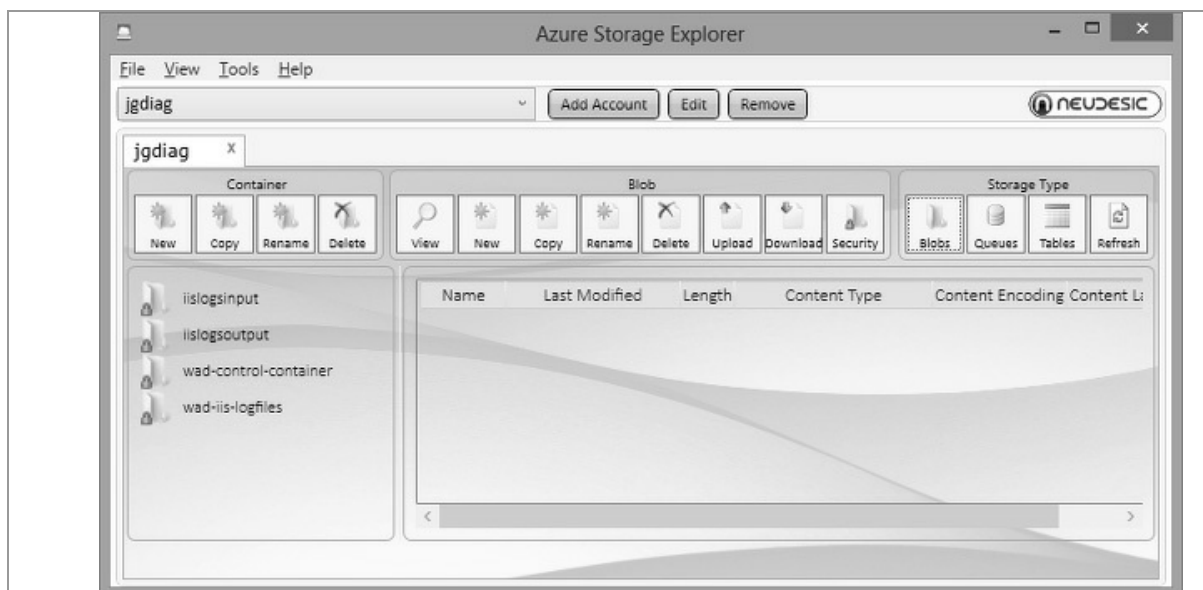
- Haga clic en **Add Account**. Una vez que agregue la cuenta al explorador de almacenamiento de Azure, no tendrá que volver a realizar este paso.



- Escriba el nombre de la cuenta de almacenamiento y la clave de la cuenta de almacenamiento y, a continuación, haga clic en **Add Storage Account**.

Puede agregar varias cuentas de almacenamiento y cada una de ellas aparecerá en una pestaña.

- Debajo de Storage Type, haga clic en **Blobs**.



- En **Container**, haga clic en el contenedor que esté asociado a su cluster de HDInsight. Cuando cree un cluster de HDInsight, debe especificar un contenedor. Si no, el proceso de creación del cluster creará uno automáticamente.

- Debajo de **Blob**, haga clic en **Upload**.
- Especifique el archivo que vaya a cargar y, a continuación, haga

clic en **Open**.

Carga de datos en el almacenamiento de blobs usando la línea de comandos de Hadoop

Para usar la línea de comandos de Hadoop, primero debe conectarse al cluster usando un escritorio remoto.

- Inicie sesión en el Portal de administración.
- Haga clic en **HDINSIGHT**. Aparecerá una lista de los clusters de Hadoop implementados.
- Haga clic en el cluster de HDInsight en el que desee cargar los datos.
- Haga clic en **CONFIGURATION** en la parte superior de la página.
- Haga clic en **ENABLE REMOTE** si todavía no ha habilitado el escritorio remoto y siga las instrucciones. Si no, continúe con el paso siguiente.
- Haga clic en **CONNECT** en la parte inferior de la página.
- Haga clic en **Open**.
- Escriba sus credenciales y, a continuación, haga clic en **OK**.
- Haga clic en **Yes**.
- En el escritorio, haga clic en **Hadoop Command Line**.
- El siguiente ejemplo indica cómo copiar el archivo davinci.txt desde el sistema de archivos local del nodo principal de HDInsight en el directorio /example/data.

```
hadoop dfs -copyFromLocal C:\temp\davinci.txt /example/data/davinci.txt
```

Como el sistema de archivos predeterminado está en el almacenamiento de blobs de Azure, /example/datadavinci.txt está en realidad en el almacenamiento de blobs de Azure. También puede referirse al archivo como:

wasb:///example/data/davinci.txt O bien,

wasbs : //@.blob.core.Windows .net/example/data/davinci .txt Se necesita el nombre completo del dominio para usar wasbs.

- Use el comando siguiente para incluir los archivos cargados:

```
hadoop dfs -lsr /example/data
```

Importación de datos a HDFS desde Base de datos SQL o SQL Server usando Sqoop

Sqoop es una herramienta diseñada para transferir datos entre Hadoop y las bases de datos relacionales. Puede usarla para importar datos desde un sistema de administración de bases de datos relacionales (RDBMS) como SQL, MySQL u Oracle en el sistema de archivos distribuidos Hadoop (HDFS), transformar los datos de Hadoop con MapReduce o Hive y, a continuación, exportar los datos en RDBMS. Para obtener más información, consulte el manual del usuario de Sqoop (en inglés).

Antes de importar los datos, debe saber el nombre de la base de datos SQL de Azure, el nombre de la cuenta de la base de datos, la contraseña de la cuenta y el nombre de la base de datos.

De forma predeterminada, una base de datos SQL de Azure permite realizar conexiones desde servicios de Azure como HDinsight. Si la configuración del firewall está deshabilitada, debe habilitarla en el Portal de administración de Azure. Para obtener instrucciones acerca de la creación de una base de datos SQL y la configuración de las reglas de firewall, consulte [Creación y configuración de una base de datos SQL](#).

El procedimiento siguiente usa PowerShell para enviar un trabajo de Sqoop.

Para importar datos a HDinsight usando Sqoop y PowerShell

- Abra la ventana de la consola de Azure PowerShell como se indica en [Instalación y configuración de Azure PowerShell](#).
- Configure los valores de las ocho primeras variables del script siguiente:

```
$subscriptionName = ""
```

```
$clusterName = ""
```

```

$SqlServerName = "
$SqlDatabaseUserName = ""
$SqlDatabasePassword = ""
$SqlDatabaseDatabaseName =
$TableName = ""

$HdfsOutputDir = "" # Esta es la ruta de HDFS para el archivo de salida, por
ejemplo "/lineltemData".

```

```

Select-AzureSubscription $subscriptionName
$SqoopDef = New-AzureHDInsightSqoopJobDefinition -
Command "import --connect
jdbc:sqlserver://$SqlServerName.database.windows
.net;user=$SqlDatabaseUserName@$SqlServerName;pa
ssword=$SqlDatabasePassword;database=$SqlDatabaseDatabas
eName --table
$TableName --target-dir $HdfsOutputDir -m 1"

```

```

$SqoopJob = Start-AzureHDInsightJob -Cluster
$ClusterName -JobDefinition $SqoopDef #-Debug -Verbose
Wait-AzureHDInsightJob -WaitTimeoutInSeconds 3600 -Job
$SqoopJob

```

```

Write-Host "Standard Error" -BackgroundColor Green
Get-AzureHDInsightJobOutput -Cluster $ClusterName -JobId
$SqoopJob.JobId -StandardError
Write-Host "Standard Output" -BackgroundColor Green
Get-AzureHDInsightJobOutput -Cluster $ClusterName -JobId
$SqoopJob.JobId -StandardOutput

```

- Pegue el script en la ventana de la consola de Azure PowerShell para ejecutarlo. Consulte [Introducción a HDInsight](#) para ver un ejemplo de PowerShell para recuperar el archivo de datos del almacenamiento de blobs de Azure.

ADMINISTRACIÓN DE HDINSIGHT CON POWERSHELL

Azure PowerShell es un potente entorno de scripting que puede usar para controlar y automatizar la implementación y la administración de sus cargas de trabajo en Azure. En este artículo, aprenderá a administrar clusters de HDInsight con una consola local de Azure PowerShell mediante el uso de Windows PowerShell.

Requisitos previos

Antes de empezar este artículo, debe tener lo siguiente:

Una suscripción de Azure. Azure es una plataforma basada en suscripción. Los cmdlets de HDInsight PowerShell realizan las tareas con su suscripción.

Una estación de trabajo con Azure PowerShell.

Aprovisionamiento de un cluster de HDInsight

HDInsight utiliza contenedores de almacenamiento de blobs de Azure como sistemas de archivos predeterminados. Es preciso tener una cuenta de almacenamiento de Azure y un contenedor de almacenamiento antes de crear un cluster de HDInsight.

Para crear una cuenta de Almacenamiento de Azure

Después de importar el archivo publishsettings, puede usar el siguiente comando para crear una cuenta de almacenamiento:

```
# Cree una cuenta de almacenamiento de Azure.  
$storageAccountName = "<StorageAccountName>"  
$location = "<Microsoft data center>" # Por  
ejemplo, "West US"
```

```
New-AzureStorageAccount -StorageAccountName
```

```
$storageAccountName -Location $location
```

Si ya tiene una cuenta de almacenamiento pero no sabe su nombre ni su clave, puede usar los comandos siguientes para recuperar dicha información:

```
# Incluya las cuentas de almacenamiento de la suscripción actual.
```

```
Get-AzureStorageAccount
```

```
# Enumere las claves de una cuenta de almacenamiento.
```

```
Get-AzureStorageKey <StorageAccountName>
```

Para crear un contenedor de almacenamiento de Azure

PowerShell no puede crear un contenedor de blobs durante el proceso de aprovisionamiento de HDInsight. Puede crear uno con el siguiente script:

```
$storageAccountName = "<StorageAccountName>"
```

```
$storageAccountKey = Get-AzureStorageKey
```

```
$storageAccountName | %{ $_.Primary }
```

```
$containerName="<ContainerName>"
```

```
# Cree un objeto de contexto de almacenamiento.
```

```
$destContext = New-AzureStorageContext -
```

```
StorageAccountName $storageAccountName
```

```
StorageAccountKey $storageAccountKey
```

```
# Cree un contenedor de almacenamiento de blobs.
```

```
New-AzureStorageContainer -Name $containerName -Context
```

```
$destContext
```

Para aprovisionar un duster

Una vez que tenga preparados la cuenta de almacenamiento y el contenedor de blobs, podrá proceder con la creación de un cluster.

```
$storageAccountName = "<StorageAccountName>"
```

```
$containerName = "<ContainerName>"
```

```
$clusterName = "<HDInsightClusterName>"
```

```
$location = "<MicrosoftDataCenter>"
```

```
$clusterNodes = <ClusterSizeInNodes>
```

```
# Obtener la clave de la cuenta de almacenamiento
```

```
$storageAccountKey = Get-AzureStorageKey
```

```
$storageAccountName | %{ $_.Primary }
```

```
# Cree un nuevo cluster de HDInsight.
```

```
New-AzureHDInsightCluster -Name $clusterName -Location
```

```
$location -DefaultStorageAccountName
```

```
"$storageAccountName.blob.core.windows.net" -
```

```
DefaultStorageAccountKey $storageAccountKey -
```

```
DefaultStorageContainerName $containerName -
```

```
ClusterSizeInNodes $clusterNodes
```

En la siguiente captura de pantalla se muestra la ejecución del script:

```

Windows PowerShell
PS C:\> $subscriptionName = "Azure-3702"
PS C:\> $clusterName = "HDI1017"
PS C:\> $location = "West US"
PS C:\> $storageAccountName = "hdistorewu"
PS C:\> $containerName = "hdi1017"
PS C:\> $clusterNodes = 1
PS C:\> # Get the storage account key
PS C:\> Select-AzureSubscription $subscriptionName
PS C:\> $storageAccountKey = Get-AzureStorageKey $storageAccountName | %{ $ .Primary }
PS C:\> # Create a Blob storage container
PS C:\> $destContext = New-AzureStorageContext -StorageAccountName $storageAccountName -StorageAccountKey $storageAccountKey
PS C:\> New-AzureStorageContainer -Name $containerName -Context $destContext

Blob End Point: https://hdistorewu.blob.core.windows.net/

Name                PublicAccess        LastModified
----                -
hdi1017              Off                  10/16/2013 2:51:15 AM +00:00

PS C:\> # Create a new HDInsight cluster
PS C:\> New-AzureHDInsightCluster -Subscription $subscriptionName -Name $clusterName -Location $location -DefaultStorageAccountName "$storageAccountName.blob.core.windows.net" -DefaultStorageAccountKey $storageAccountKey -DefaultStorageContainerName $containerName -ClusterSizeInNodes $clusterNodes

cmdlet New-AzureHDInsightCluster at command pipeline position 1
Supply values for the following parameters:
(Type !? for Help.)
Credential

Name                : HDI1017
HttpUserName        : admin
HttpPassword        : Password1
Version             : 2.0.1.0.257042
VersionStatus       : Compatible
ConnectionUrl       : https://HDI1017.azurehdinsight.net
State               : Running
CreateDate          : 10/16/2013 2:51:30 AM
UserName            : admin
Location            : West US
ClusterSizeInNodes : 1
DefaultStorageAccount : hdistorewu.blob.core.windows.net
SubscriptionId      : 65a1016d-0f67-45d2-d52e
StorageAccounts     : {}

PS C:\>

```

Enumeración y visualización de clusters

Use los comandos siguientes para enumerar y mostrar los detalles del cluster:

Para enumerar todos los clusters en la suscripción actual

Get-AzureHDInsightCluster

Para mostrar los detalles del cluster específico en la suscripción actual

Get-AzureHDInsightCluster -Name <ClusterName>

Eliminación de un cluster

Use el comando siguiente para eliminar un cluster:

Remove-AzureHDInsightCluster -Name <ClusterName>

Concesión/Revocación del acceso a los servicios de

HTTP

Los clusters de HDInsight tienen los siguientes servicios web HTTP (todos estos servicios tienen extremos RESTful):

ODBC

JDBC

Ambari

Oozie

Templeton

De manera predeterminada, estos servicios se conceden para el acceso. Puede revocar/conceder el acceso. A continuación se ofrece una muestra:

```
Revoke-AzureHDInsightHttpServicesAccess -Name hdiv2-  
Location "East US"
```

En la muestra *hdiv2* es el nombre de un cluster de HDInsight.

Envío de trabajos de MapReduce

La distribución de clusters de HDInsight se incluye con algunas muestras de MapReduce. Una de las muestras ilustra el recuento de la frecuencia de las palabras en los archivos de código fuente.

Para enviar un trabajo de MapReduce

El siguiente script de PowerShell envía el trabajo de muestra de recuento de palabras:

```
$clusterName = "<HDInsightClusterName>"  
  
# Defina el trabajo de MapReduce  
$wordCountJobDefinition = New-  
AzureHDInsightMapReduceJobDefinition -JarFile  
"wasb:///example/jars/hadoop-examples.jar" -ClassName  
"wordcount" -Arguments  
"wasb:///example/data/gutenberg/davinci.txt",  
"wasb:///example/data/WordCountOutput"
```

```
# Ejecute el trabajo y muestre el error estándar
$wordCountJobDefinition | Start-AzureHDInsightJob -
Cluster $clusterName | Wait-AzureHDInsightJob -
WaitTimeoutInSeconds 3600 | %{ Get-
AzureHDInsightJobOutput -Cluster $clusterName -JobId
$_.JobId -StandardError}
```

Para obtener información acerca del prefijo WASB, consulte [Uso del almacenamiento de blobs de Azure para HDInsight][hdinsight-storage],

Para descargar la salida del trabajo de MapReduce

El siguiente script de PowerShell recupera la salida del trabajo de MapReduce desde el último procedimiento:

```
$storageAccountName = "<StorageAccountName>"
$containerName = "<ContainerName>"

# Crear el objeto de contexto de la cuenta de almacenamiento
$storageAccountKey = Get-AzureStorageKey
$storageAccountName | %{ $_.Primary }
$storageContext = New-AzureStorageContext -
StorageAccountName $storageAccountName -
StorageAccountKey $storageAccountKey
# Descargar la salida en el equipo local
Get-AzureStorageBlobContent -Container $ContainerName -
Blob example/data/WordCountOutput/part-r-OOOOO -Context
$storageContext -Force

# Mostrar la salida
cat ./example/data/WordCountOutput/part-r-OOOOO |
findstr "there"
```

Envío de trabajos de Hive

La distribución de clusters de HDInsight se incluye con una tabla de muestra de Hive llamada *hivesampletable*. Puede usar una consulta “show tables;” de HiveQL para enumerar las tablas de Hive en un cluster.

Para enviar un trabajo de Hive

El siguiente script envía un trabajo de Hive para enumerar las tablas de Hive:

```
$clusterName = "<HDInsightClusterName>"
```

```
# Consulta de HiveQL
```

```
$querystring =
```

```
    SHOW TABLES;
```

```
    SELECT * FROM hivesampletable
```

```
    WHERE Country='United Kingdom'
```

```
    LIMIT 10;
```

```
@
```

```
Use-AzureHDInsightCluster -Name $clusterName
```

```
Invoke-Hive $querystring
```

El trabajo de Hive mostrará primero las tablas de Hive creadas en el cluster y los datos devueltos por *hivesampletable*.

INTRODUCCIÓN AL EMULADOR DE HDINSIGHT

El emulador de HDInsight proporciona un entorno de desarrollo local para HDInsight de Azure. Si ya conoce Hadoop, puede comenzar con el emulador usando HDFS. No obstante, en HDInsight, el sistema de archivos predeterminado es el almacenamiento de blobs de Azure (WASB, alias Azure Storage - Blobs), por lo que, finalmente, querrá desarrollar sus trabajos con WASB. Puede comenzar a desarrollar contra WASB usando el emulador de almacenamiento de Azure; probablemente solo desee usar un pequeño subconjunto de los datos (no se necesitan cambios en la configuración en el emulador de HDInsight, solo un nombre de cuenta de almacenamiento diferente). Posteriormente, prueba sus trabajos localmente contra el almacenamiento de Azure, nuevamente usando solamente un subconjunto de los datos (requiere un cambio en la configuración del emulador de HDInsight). Finalmente, está listo para cambiar la parte de cálculo de su trabajo HDInsight y ejecutar un trabajo contra los datos de producción.

Requisitos previos

Antes de comenzar este tutorial, debe cumplir los siguientes requisitos previos:

El emulador de HDInsight requiere una versión de Windows de 64 bits. Se debe cumplir uno de los siguientes requisitos:

Windows 7 Service Pack 1

Windows Server 2008 R2 Service Pack 1

Windows 8

Windows Server 2012

Instale y configure Azure PowerShell.

Instalación del emulador de HDInsight

El emulador de Microsoft HDInsight se puede instalar a través del instalador de plataforma web de Microsoft.

El emulador de HDInsight solo es compatible actualmente con un sistema operativo en idioma inglés.

[WACOM.NOTE] Si ya ha tenido instalado Microsoft HDInsight Developer Preview, deberá primero desinstalar los siguientes dos componentes desde el Panel de control o Programas y características.

HDInsight Developer Preview

Hortonworks Data Platform Developer Preview

Para instalar el emulador de HDInsight

- Abra Internet Explorer y diríjase a la página de instalación del emulador de Microsoft HDInsight para Azure.
- Haga clic en **Install Now**.
- Haga clic en **Run** cuando se le pregunte por la instalación de HDINSIGHT.exe en la parte inferior de la página.
- Haga clic en el botón **Yes** en la ventana **User Account Control** que aparece para completar la instalación. Verá la ventana Instalador de plataforma web 4.6.
- Haga clic en **Install** en la parte inferior de la página.
- Haga clic en **I Accept** para aceptar los términos de la licencia.
- Verifique que el Instalador de plataforma web muestre **the following Products were successfully installed** y, a continuación, haga clic en **Finish**.
- Haga clic en **Exit** para cerrar la ventana del Instalador de plataforma web 4.6.

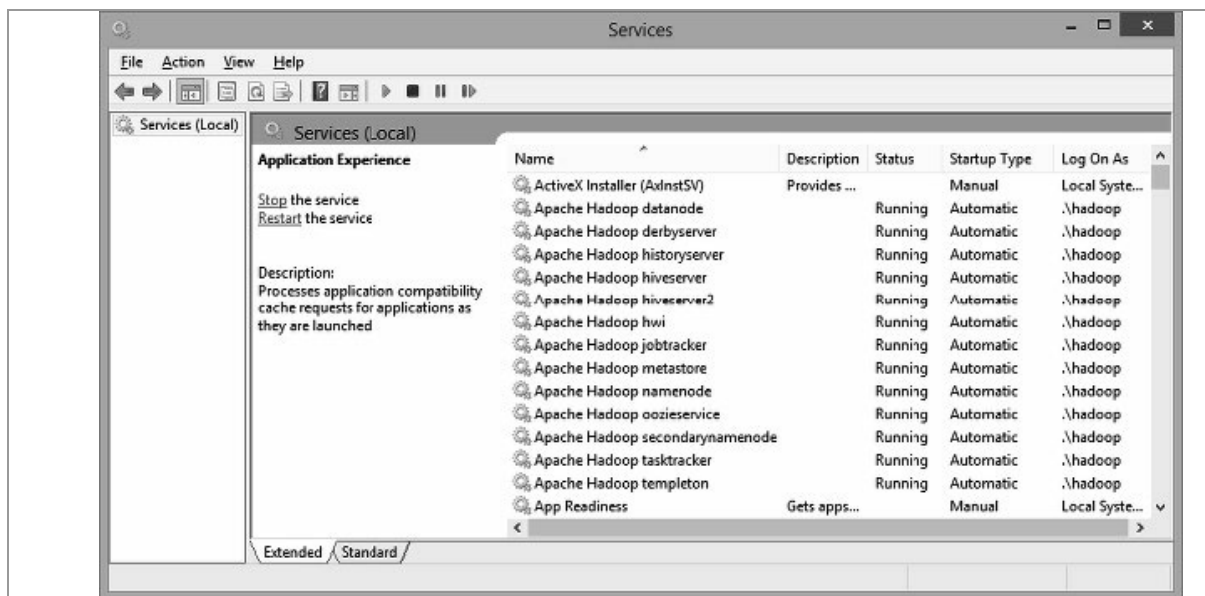
La instalación debería haber instalado tres iconos en el escritorio. Los tres iconos están vinculados como se muestra a continuación:

Hadoop Command Line: el símbolo del sistema de Hadoop desde el que se ejecutan trabajos de MapReduce, Pig y Hive en el emulador de HDInsight.

Hadoop Name Node Status: el NameNode mantiene el directorio de árbol para todos los archivos en HDFS. Realiza también un seguimiento de dónde se mantienen los datos de todos los archivos en un cluster de Hadoop. Los clientes se comunican con el NameNode para resolver los nodos de datos de todos los archivos que están almacenados.

Hadoop MapReduce Status: El seguimiento de trabajo que asigna tareas de MapReduce a los nodos en un cluster.

La instalación debería haber instalado también varios servicios locales. La siguiente es una captura de pantalla de la ventana de servicios:



Para obtener información acerca de los problemas conocidos con la instalación y ejecución de HDInsight Server, consulte las Notas de la versión del emulador de HDInsight. El registro de instalación se encuentra en

C:\HadoopFeaturePackSetup\HadoopFeaturePackSetupTools\gettingSg.install.log.

Ejecución de un trabajo de MapReduce de recuento de palabras

Ahora el emulador de HDInsight está configurado en su estación de trabajo. Puede ejecutar un trabajo de MapReduce para probar la instalación. Primero, cargará algunos archivos de texto en HDFS y, posteriormente, ejecutará un trabajo de MapReduce de recuento de

palabras para contar las frecuencias de palabras de esos archivos.

El programa MapReduce de recuento de palabras se empaquetó en *hadoop-examples.jar*. El archivo jar está ubicado en la carpeta *C:\Hadoop\hadoop-1.1.0- SNAPSHOT*.

La sintaxis del comando jar es:

```
hadoop jar <jar> [mainClass] args...
```

Utilizará también algunos comandos fs.

El trabajo de MapReduce de recuento de palabras utiliza dos argumentos: una carpeta de entrada y una carpeta de salida. Usará *hdfs://localhost/user/HDIUser* como carpeta de entrada y *hdfs://localhost/user/HDIUser/WordCount_Output* como directorio de salida. La carpeta de salida no puede estar en una carpeta existente, de lo contrario, el trabajo de MapReduce producirá un error. Si desea ejecutar el trabajo de MapReduce por segunda vez, debe especificar una carpeta de salida diferente o eliminar la carpeta de salida existente.

Para ejecutar el trabajo de MapReduce de recuento de palabras

- En el escritorio, haga doble clic en **Hadoop Command Line** para abrir la ventana de línea de comandos de Hadoop. La carpeta actual debe ser:

```
c:\Hadoop\hadoop-1.1.0-SNAPSHOT>
```

De lo contrario, ejecute el siguiente comando:

```
cd %hadoop_home%
```

- Ejecute el siguiente comando de Hadoop para crear una carpeta de HDFS para el almacenamiento de los archivos de entrada y salida:

```
hadoop fs -mkdir /user/HDIUser
```

- Ejecute el siguiente comando de Hadoop para copiar algunos archivos locales en HDFS:

```
hadoop fs -copyFromLocal *.txt /user/HDIUser/
```

- Ejecute el siguiente comando para ver en una lista de los archivos de la carpeta /user/HDIUser:

```
hadoop fs -ls /user/HDIUser
```

Debería ver los siguientes archivos:

```
c:\Hadoop\hadoop-1.1.0-SNAPSHOT>hadoop fs -ls
```

/user/HDIUser

Found 8 ítems

```
-rw-r--r--      1 username supergroup      16372 2013-10-
30 12:07 /user/HDIUser/CHANGES.branch-1-win.txt
-rw-r--r--      1 username supergroup      463978 2013-10-
30 12:07 /user/HDIUser/CHANGES.txt
-rw-r--r--      1 username supergroup       6631 2013-10-
30 12:07 /user/HDIUser/Jira-Analysis.txt
-rw-r--r--      1 username supergroup      13610 2013-10-
30 12:07 /user/HDIUser/LICENSE.txt
-rw-r--r--      1 username supergroup       1663 2013-10-
30 12:07 /user/HDIUser/Monarch-CHANGES.txt
-rw-r--r--      1 username supergroup       103 2013-10-
30 12:07 /user/HDIUser/NOTICE.txt
-rw-r--r--      1 username supergroup      2295 2013-10-
30 12:07 /user/HDIUser/README.Monarch.txt
-rw-r--r--      1 username supergroup      1397 2013-10-
30 12:07 /user/HDIUser/README.txt
```

- Ejecute el siguiente comando para ejecutar el trabajo de MapReduce de recuento de palabras:

```
hadoop jar hadoop-examples.jar wordcount /user/HDIUser/*.txt
/user/HDIUser/WordCount_Output
```

- Ejecute el siguiente comando para mostrar en una lista las palabras que tienen “ventanas” desde el archivo de salida:

```
hadoop fs -cat /user/HDIUser/WordCount_Output/part-r-
00000 | findstr “Windows”
```

La salida debe ser:

```
c:\Hadoop\hadoop-1.1.0-SNAPSHOT>hadoop fs -cat
/user/HDIUser/WordCount_Output/pa
rt-r-00000 | findstr “Windows”
Windows 12
windows+java6.      1
```

Ejecución de los ejemplos de introducción

La instalación del emulador de HDInsight proporciona algunos ejemplos para que los nuevos usuarios aprendan rápidamente los servicios basados en Apache Hadoop en Windows. Estos ejemplos cubren algunas tareas que generalmente se necesitan cuando se procesa un conjunto de datos grande. El repaso de los ejemplos puede familiarizarlo con los conceptos asociados con el modelo de programación de MapReduce y su ecosistema.

Los ejemplos se organizan en torno a los escenarios de datos del registro de IIS W3C de procesamiento. Se proporciona una herramienta de generación de datos para crear e importar los conjuntos de datos de diferentes tamaños a HDFS o WASB (almacenamiento de blobs de Azure). Luego, los trabajos de MapReduce, Pig o Hive se pueden ejecutar en las páginas de datos generadas por el script de PowerShell. Tenga en cuenta que los scripts de Pig y Hive usados se compilan ambos para programas de MapReduce. Los usuarios pueden ejecutar una serie de trabajos para que observen, por su cuenta, los efectos de usar estas diferentes tecnologías y los efectos del tamaño de los datos en la ejecución de las tareas de procesamiento.

Los escenarios de datos del registro de IIS w3c

El escenario de w3c genera e importa datos de registro de IIS W3C en tres tamaños a HDFS o WASB: 1 MB, 500 MB y 2 GB. Proporciona tres tipos de trabajos y los implementa en C#, Java, Pig y Hive.

totalhits: calcula la cantidad total de solicitudes de una página determinada.

avgtime: calcula el tiempo promedio utilizado (en segundos) para una solicitud por página.

errors: calcula la cantidad de errores por página y por hora para las solicitudes cuyo estado fue 404 o 500.

Estos ejemplos y su documentación no proporcionan un estudio en

profundidad ni una implementación completa de las tecnologías clave de Hadoop. El cluster usado tiene un solo nodo y, por lo tanto, no se puede observar el efecto de agregar más nodos en esta versión.

Carga de los datos de ejemplo del registro de w3c

La generación e importación de los datos a HDFS se hace usando el script `importdata.psl` de PowerShell.

Para importar los datos de ejemplo del registro de w3c:

- Abra la línea de comandos de Hadoop desde el escritorio.
- Ejecute el siguiente comando para cambiar el directorio a `C:\Hadoop\GettingStarted`:

```
cd \Hadoop\GettingStarted
```

- Ejecute el siguiente comando para generar e importar datos a HDFS:

```
powershell -File importdata.psl w3c -ExecutionPolicy unrestricted
```

- Ejecute el siguiente comando desde la línea de comandos de Hadoop para mostrar los archivos importados en el HDFS:

```
hadoop fs -lsr /w3c
```

La salida debe ser similar a la siguiente:

```
c:\Hadoop\GettingStarted\w3c>hadoop fs -lsr /w3c
drwxr-xr-x    -   username    supergroup    0 2013-10-
30 13:29 /w3c/input
drwxr-xr-x    -   username    supergroup    0 2013-10-
30 13:29 /w3c/input/large
-rw-r--r--    1   username    supergroup    543692369 2013-10-
30 13:29 /w3c/input/large/data_w3c_large.txt
drwxr-xr-x    -   username    supergroup    0 2013-10-
30 13:28 /w3c/input/medium
-rw-r--r--    1   username    supergroup    272394671 2013-10-
30 13:28 /w3c/input/medium/data_w3c_medium.txt
```



```
drwxr-xr-x    -   username    supergroup    0 2013-10-
30 13:28 /w3c/input/small
-rw-r--r--    1   username    supergroup    1058328 2013-10-
30 13:28 /w3c/input/small/data_w3c_small.txt
```

- Ejecute el siguiente comando para mostrar uno de los archivos de datos en la ventana de la consola:

```
hadoop fs -cat /w3c/input/small/data_w3c_small.txt
```

Ahora, el archivo de datos se ha creado e importado en HDFS. Puede ejecutar diferentes trabajos de Hadoop.

Ejecución de trabajos de MapReduce de Java

MapReduce es el motor de cálculo básico de Hadoop. De manera predeterminada, está implementado en Java, pero también hay ejemplos que aprovechan el streaming de .NET y Hadoop que usan C#. La sintaxis para ejecutar un trabajo de MapReduce es:

```
hadoop jar <jarFileName>.jar <className> <inputFiles>
<outputFolder>
```

El archivo jar y los archivos de origen se encuentran en la carpeta C:\Hadoop\GettingStarted\Java.

Para ejecutar un trabajo de MapReduce y así calcular las visitas de la página web

- Abra la línea de comandos de Hadoop.
- Ejecute el siguiente comando para cambiar el directorio a **C:\Hadoop\GettingStarted**:

```
cd \Hadoop\GettingStarted
```

- Ejecute el siguiente comando para quitar el directorio de salida en caso de que exista la carpeta. El trabajo de MapReduce producirá un error si la carpeta de salida ya existe.

```
hadoop fs -rmr /w3c/output
```

- Ejecute el siguiente comando:

```
hadoop jar .\Java\w3c_scenarios.jar
```

“microsoft.hadoop.w3c.TotalHitsForPage”

“/w3c/input/small/data_w3c_small.txt” “/w3c/output”

En la siguiente tabla se describen los elementos del comando:

Parámetro

Nota:

w3c_scenarios.jar

El archivo jar está ubicado en la carpeta C:\Hadoop\GettingStarted\Java.

microsoft.hadoop.w3c.TotalHitsForPage

El tipo se puede sustituir por una de las siguientes opciones:

- o microsoft.hadoop.w3c.AverageTimeTaken
- o microsoft.hadoop.w3c.ErrorsByPage

/w3c/input/small/data_w3c_small.txt

El archivo de entrada se puede sustituir por lo siguiente:

- o /w3c/input/medium/data_w3c_medium.txt
- o /w3c/input/large/data_w3c_large.txt

/w3c/output

Este es el nombre de la carpeta de salida.

- Ejecute el siguiente comando para ver el archivo de salida:

```
hadoop fs -cat /w3c/output/part-00000
```

La salida debe ser similar a:

```
c:\Hadoop\GettingStarted\Java>hadoop fs -cat
```

```
/w3c/output/part-00000
```

```
/Default.aspx      3409
```

```
/Info.aspx         1115
```

```
/UserService       1130
```

De esta manera, la página Default.aspx recibe 3409 visitas y así sucesivamente.

Ejecución de trabajos de Hive

Para los analistas con profundos conocimientos de SQL el motor de consulta de Hive les será familiar. Proporciona una interfaz similar a SQL y un modelo de datos relacionales para HDFS. Hive usa un lenguaje llamado HlveQL (o HQL), que es un dialecto de SQL.

Para ejecutar un trabajo de Hive

Abra la línea de comandos de Hadoop.

Cambie el directorio a la carpeta **C:\Hadoop\GettingStarted**.

Ejecute el siguiente comando para quitar la carpeta **/w3c/hive/input** en caso de que exista. El trabajo de Hive producirá un error si la carpeta existe.

```
hadoop fs -rmr /w3c/hive/input
```

- Ejecute el siguiente comando para crear la carpeta **/w3c/hive/input** y copie el archivo de datos de la estación de trabajo en HDFS:

```
hadoop fs -mkdir /w3c/hive/input
```

```
hadoop fs -cp /w3c/input/small/data_w3c_small.txt
```

```
/w3c/hive/input
```

- Ejecute el siguiente comando para ejecutar el archivo de script **w3ccreate.hql**. El script crea una tabla de Hive y carga datos en la tabla de Hive:

```
C:\Hadoop\hive-0.9.0\bin\hive.cmd -f
```

```
./Hive/w3c/w3ccreate.hql -hiveconf
```

```
“input = /w3c/hive/input/data_w3c_small.txt”
```

El script de HiveQL es:

```
DROP TABLE W3C;
```

```
CREATE TABLE W3C(
```

```
    logdate string,
```

```
    logtime string,
```

```
    c_ip string,
```

```
    cs_username string,
```

```
    s_ip string,
```

```
    s_port string,
```

```
cs_method string,  
cs_uri_stem string,  
cs_uri_query string,  
sc_status int,  
sc_bytes int,  
cs_bytes int,  
time_taken int,  
cs_agent string,  
cs_Referrer string)
```

ROW FORMAT delimited

FIELDS TERMINATED BY ' ';

LOAD DATA INPATH '\${hiveconf:input}' OVERWRITE INTO TABLE W3C;

La salida debe ser similar a la siguiente:

```
c:\Hadoop\GettingStarted>C:\Hadoop\hive-  
0.9.0\bin\hive.cmd -f ./Hive/w3c/w3ccreate.hql -  
hiveconf "input=/w3c/hive/input/data_w3c_small.txt"  
Hive history file=c:\hadoop\hive-  
0.9.0\logs\history\hive_job_log_username_201310311452_10_53491002.txt  
Logging initialized using configuration in  
file:/C:/Hadoop/hive-0.9.0/conf/hive-log4j.properties  
OK  
Time taken: 0.616 seconds  
OK  
Time taken: 0.139 seconds  
Loading data to table default.w3c  
Moved to trash:  
hdfs://localhost:8020/apps/hive/warehouse/w3c  
OK  
Time taken: 0.573 seconds
```

- Ejecute el siguiente comando para ejecutar el archivo de script de HiveQL

w3ctotalhitsbypate.hql.

```
C:\Hadoop\hive-0.9.0\bin\hive.cmd -f
./Hive/w3c/w3ctotalhitsbypage.hql
```

En la siguiente tabla se describen los elementos del comando:

Archivo

Descripción

C: \Hadoop\hive-0.9.0\bin\hive.cmd

El script de comando de Hive.

C: \Hadoop\Getting Started\Hive\w3c\w3ctotalhitsbypage.hql

Puede sustituir el archivo de script de Hive por una de las siguientes opciones:

- o C: \Hadoop\Getting Started\Hive\w3c\w3caveragetimetaken.hql
- o C: \Hadoop\Getting Started\Hive\w3c\w

El script de HiveQL w3ctotalhitsbypage.hql es:

```
SELECT filtered.cs_uri_stem,COUNT(*)
FROM (
    SELECT logdate,cs_uri_stem from w3c WHERE logdate NOT
RLIKE '.*#.*'
    ) filtered
GROUP BY (filtered.cs_uri_stem);
```

El final de la salida debe ser similar a la siguiente:

```
MapReduce Total cumulative CPU time: 3 seconds 47 msec
Ended Job = job_201310291309_0006
MapReduce Jobs Launched:
Job 0: Map: 1 Reduce: 1 Cumulative CPU: 3.047 sec
HDFS Read: 1058546 HDFS Write: 53 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 47 msec
OK
/Default.aspx 3409
/Info.aspx 1115
```

/UserService 1130

Time taken: 34.68 seconds

Tenga en cuenta que como primer paso en cada uno de los trabajos, se creará una tabla y los datos se cargarán en la tabla desde el archivo que se creó anteriormente. Puede examinar el archivo que se creó buscando bajo el nodo /Hive en HDFS con el siguiente comando:

```
hadoop fs -lsr /apps/hive/
```

Ejecución de trabajos de Pig

El procesamiento de Pig usa un lenguaje de flujo de datos llamado Pig Latín. Las abstracciones de Pig Latín proporcionan estructuras de datos más enriquecidas que MapReduce y realizan para Hadoop lo que SQL realiza para los sistemas RDBMS.

Para ejecutar trabajos de Pig

- Abra la línea de comandos de Hadoop.
- Cambie el directorio a la carpeta C:\Hadoop\GettingStarted.
- Ejecute el siguiente comando para enviar un trabajo de Pig:

```
C:\Hadoop\pig-0.9.3-SNAPSHOT\bin\pig.cmd -f
```

```
“.\Pig\w3c\TotalHitsForPage.pig” -p
```

```
“input=/w3c/input/small/data_w3c_small.txt”
```

En la siguiente tabla se muestran los elementos del comando:

Archivo

Descripción

C:\Hadoop\pig-0.9.3-SNAPSHOT\bin\pig.cmd

El script de comando de Pig

C:\Hadoop\Getting Started\Pig\w3c\TotalHitsForPage.pig

Puede sustituir el archivo de script de Pig Latin por una de las siguientes opciones:

- o C:\Hadoop\GettingStarted\Pig\w3c\AverageTimeTaken.pig
- o C:\Hadoop\GettingStarted\Pig\w3c\ErrorsByPage.pig

/w3c/input/small/data_w3c_small.txt

Puede sustituir el parámetro por un archivo más grande:

- o `/w3c/input/medium/data_w3c_medium.txt`
- o `/w3c/input/large/data_w3c_large.txt`

La salida debe ser similar a la siguiente:

`(/Info.aspx,1115)`

`(/UserService,1130)`

`(/Default.aspx,3409)`

Tenga en cuenta que debido a que los Scripts de Pig se compilan en trabajos de MapReduce y posiblemente en más de uno de tales trabajos, los usuarios pueden ver varios trabajos de MapReduce que se ejecutan en el curso del procesamiento de un trabajo de Pig.

Recompilación de los ejemplos

Los ejemplos contienen actualmente todos los binarios necesarios, por lo que no se requiere una compilación. Si desea realizar cambios en los ejemplos de Java o .NET, puede volver a compilarlos usando `msbulld` o el script de PowerShell que se incluye.

Para recompilar los ejemplos

- Abra la línea de comandos de Hadoop.
- Ejecute el siguiente comando:

```
powershell -F buildsamples.psl
```

Almacenamiento de blobs de Azure

HDInsight de Azure usa el almacenamiento de blobs de Azure como sistema de archivos predeterminado.

Se puede configurar un cluster local en el emulador de HDInsight para usar el almacenamiento de blobs de Azure en vez del almacenamiento local. Esta sección cubre los siguientes temas:

Conexión con el emulador de almacenamiento

Conexión con el almacenamiento de blobs de Azure

Configuración de un almacenamiento de blobs de Azure como el

sistema de archivos predeterminado para el emulador de HDInsight

Conexión con el emulador de almacenamiento

El emulador de almacenamiento de Azure viene con el SDK de Azure para .NET. El emulador de almacenamiento no se inicia automáticamente. Debe iniciarlo manualmente. El nombre de la aplicación es Emulador de almacenamiento de Azure. Para iniciar/detener los emuladores, haga clic con el botón secundario en el icono azul de Azure en la bandeja del sistema de Windows y, a continuación, haga clic en Show Storage Emulator UI.

Es posible que reciba el siguiente mensaje de error al iniciar el emulador de almacenamiento:

El proceso no puede obtener acceso al archivo porque está siendo utilizado en otro proceso.

Esto se debe a que uno de los servicios Hive de Hadoop también usa el puerto 10000. Para solucionar el problema, use el siguiente procedimiento:

- Detenga los dos servicios Hive de Hadoop con Services.msc: Apache Hadoop hiveserver y Apache Hadoop Hiveserver2.
- Inicie el emulador de almacenamiento de blobs.
- Reinicie los dos servicios Hive de Hadoop.

La sintaxis para tener acceso al emulador de almacenamiento es:

```
wasb://<ContainerName>@storageemulator
```

Por ejemplo:

```
hadoop fs -ls wasb://myContainer@storageemulator
```

Si recibe el siguiente mensaje de error:

```
ls: No FileSystem for scheme: wasb
```

Se debe a que todavía usa la versión Developer Preview. Siga las instrucciones que se encuentran en la sección Instalación del emulador de HDInsight de este artículo para desinstalar la versión preliminar del desarrollador y luego vuelva a instalar la aplicación.

Conexión con el almacenamiento de blobs de Azure

Para crear un contenedor

- Inicie sesión en el Portal de administración.
- Haga clic en **ALMACENAMIENTO** a la izquierda. Debe ver una lista de cuentas de almacenamiento debajo de su suscripción.
- En la lista, haga clic en la cuenta de almacenamiento donde desea crear el contenedor.
- Haga clic en **CONTAINERS** en la parte superior de la página.
- Haga clic en **ADD** en la parte inferior de la página.
- Escriba **NAME** y seleccione **ACCESS**. Puede usar cualquiera de los tres niveles de acceso. El valor predeterminado es **Private**.
- Haga clic en **OK** para guardar los cambios. Podrá ver que el contenedor nuevo se muestra en el portal.

Para que pueda tener acceso a una cuenta de almacenamiento de Azure, debe agregar el nombre y la clave de la cuenta al archivo de configuración.

Para configurar la conexión a una cuenta de almacenamiento de Azure

- Abra **C:\Hadoop\hadoop-1.1.0-SNAPSHOT\conf\core-site.xml** en el Bloc de notas.
- Agregue la siguiente etiqueta `<property>` junto a las demás etiquetas `<property>`:

```
<property>  
  
<name>fs.azure.account.key.<StorageAccountName>.blob.cor  
e.Windows.net</name>  
  <value>xStorageAccountKey</value>  
</property>
```

Debe sustituir `<StorageAccountName>` y `<StorageAccountKey>` por los valores que coinciden con la información de su cuenta de almacenamiento.

- Guarde el cambio. No es necesario reiniciar los servicios de Hadoop.

Use la siguiente sintaxis para tener acceso a la cuenta de almacenamiento:

```
wasb://<ContainerName>@<StorageAccountName>.blob.core.windows.net/
```

Por ejemplo:

```
hadoop fs -ls
```

```
wasb://myContainer@myStorage.blob.core.Windows.net/
```

Uso de un contenedor de almacenamiento de blobs de Azure como sistema de archivos predeterminado

También es posible usar un contenedor de almacenamiento de blobs de Azure como sistema de archivos predeterminado, como ocurre con HDInsight de Azure.

Para configurar el sistema de archivos predeterminado con un contenedor de almacenamiento de blobs de Azure:

- Abra C:\Hadoop\hadoop-1.1.0-SNAPSHOT\conf\core-site.xml en el Bloc de notas.
- Busque la siguiente etiqueta <property>:

```
<property>  
  <name>fs.default.name</name>  
  <!-- cluster variant -->  
  <value>hdfs://localhost:8020</value>  
  <description>The name of the default file system.
```

Either the literal string “local” or a host:port for
NDFS.</description>

```
  <final>true</final>  
</property>
```

- Reemplácela por las siguientes dos etiquetas <property>:

```
<property>  
  <name>fs.default.name</name>
```

```
<!-- cluster variant -->
<!--<value>hdfs://localhost:8020</value>-->
```

```
<value>wasb://<ContainerName>@<StorageAccountName>.blob.
core.Windows.net</value>
```

```
<description>The name of the default file system.
```

```
Either the literal string “local” or a host:port for
NDFS.</description>
```

```
<final>true</final>
```

```
</property>
```

```
<property>
```

```
<name>dfs.namenode.rpc-address</name>
```

```
<value>hdfs://localhost:8020</value>
```

```
<description>A base for other temporary
directories.</description>
```

```
</property>
```

Debe sustituir <StorageAccountName> y <StorageAccountKey> por los valores que coinciden con la información de su cuenta de almacenamiento.

- Guarde los cambios.
- Abra la línea de comandos de Hadoop en su escritorio en modo elevado (Ejecutar como administrador).
- Ejecute los siguientes comandos para reiniciar los servicios de Hadoop:

```
C:\Hadoop\stop-onebox.cmd
```

```
C:\Hadoop\start-onebox.cmd
```

- Ejecute el siguiente comando para probar la conexión con el sistema de archivos predeterminado:

```
hadoop fs -ls /
```

Los siguientes comandos muestran el contenido en la misma carpeta:

```
hadoop fs -ls wasb:///
```

```
hadoop fs -ls  
wasb://<ContainerName>@<StorageAccountName>.blob.core.wi  
ndows.net/
```

```
hadoop fs -ls  
wasbs://<ContainerName>@<StorageAccountName>.blob.core.w  
indows.net/
```

Para tener acceso a HDFS, use el siguiente comando:

```
hadoop fs -ls hdfs://localhost:8020/
```

Ejecución de HDInsight PowerShell

Algunos de los cmdlets de HDInsight PowerShell son compatibles con el emulador de HDInsight. Estos cmdlets incluyen:

Cmdlets de definición del trabajo de HDInsight

New-AzureHDInsightSqoopJobDefinition

New-AzureHDInsightStreamingMapReduceJobDefinition

New-AzureHDInsightPigJobDefinition

New-AzureHDInsightHiveJobDefinition

New-AzureHDInsightMapReduceJobDefinition

Start-AzureHDInsightlob

Get-AzureHDInsightlob

Wait-AzureHDInsightlob

El siguiente es un ejemplo para enviar un trabajo de Hadoop:

```
$creds = Get-Credential (hadoop as username, password  
can be anything)
```

```
$hdinsightJob = <JobDefinition>
```

```
Start-AzureHDInsightJob -Cluster http://localhost:50111  
-Credential $creds -JobDefinition $hdinsightJob
```

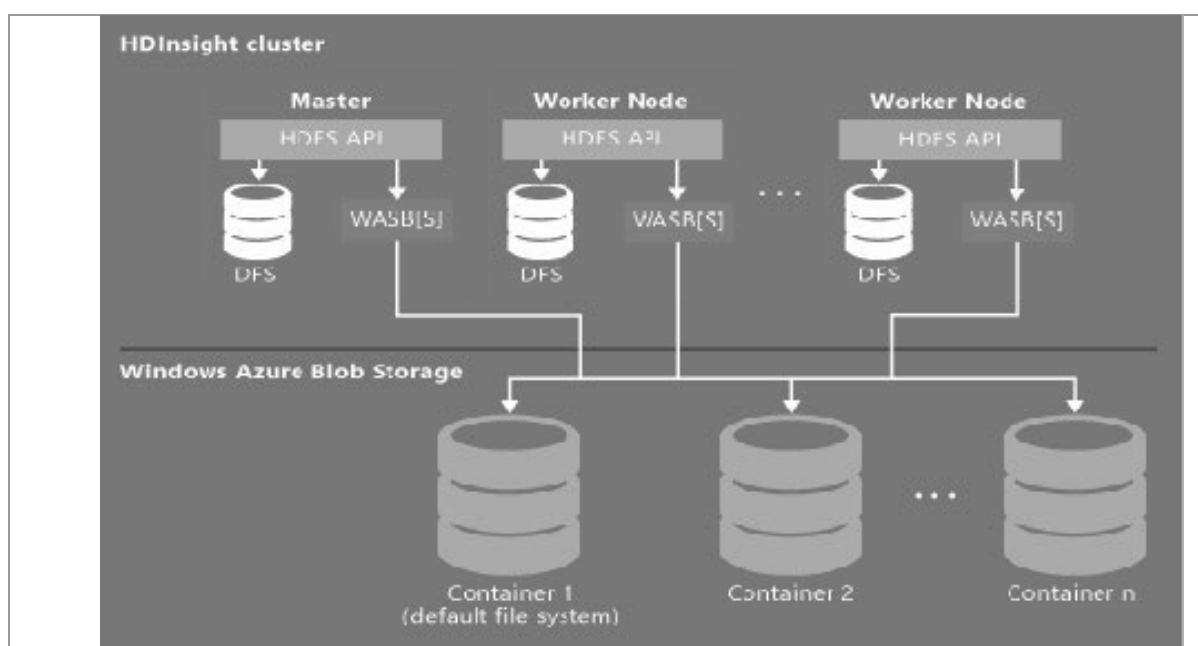
Recibirá una indicación cuando llame a Get-Credential. Debe usar **hadoop** como nombre de usuario. La contraseña puede ser cualquier cadena. El nombre del cluster siempre es **http://localhost:50111**.

USO DEL ALMACENAMIENTO DE BLOBS DE AZURE CON HDINSIGHT

HDInsight de Azure admite el sistema de archivos distribuidos Hadoop (HDFS) y el almacenamiento de blobs de Azure para almacenar datos. El almacenamiento de blobs es una solución eficaz y de uso general de Azure. El almacenamiento de blobs proporciona una interfaz HDFS con todas las características para ofrecer una experiencia sin igual que habilita un conjunto completo de componentes en el ecosistema de Hadoop para operar (de forma predeterminada) directamente en los datos. El almacenamiento de blobs no es solo una solución de bajo coste; almacenar datos en un almacenamiento de blobs permite eliminar de forma segura de los clusters de HDInsight usados para los cálculos sin perder los datos del usuario.

Arquitectura de almacenamiento de HDInsight

El diagrama siguiente proporciona una panorámica de la arquitectura de almacenamiento de HDInsight:



HDInsight brinda acceso al sistema de archivos distribuidos que se adjunta localmente a los nodos de ejecución. Se puede acceder a este

sistema de archivos usando el URI completo. Por ejemplo:

```
hdfs://<namenodehost>/<path>
```

Además, HDInsight ofrece la capacidad de acceder a los datos almacenados en el almacenamiento de blobs. La sintaxis para acceder al almacenamiento de blobs es la siguiente:

```
wasb [s] ://@.blob.core.Windows.net/
```

Hadoop admite una noción del sistema de archivos predeterminado. El sistema de archivos predeterminado implica un esquema y una autoridad predeterminados y también se puede usar para solucionar rutas relativas. Durante el proceso de aprovisionamiento de HDInsight, se designan una cuenta de almacenamiento de Azure y un contenedor de almacenamiento de blobs específico de dicha cuenta como sistema de archivos predeterminado.

Además de esta cuenta de almacenamiento, puede agregar más cuentas de almacenamiento desde la misma suscripción de Azure o desde otras diferentes durante el proceso de aprovisionamiento. Para obtener instrucciones acerca de cómo agregar más cuentas de almacenamiento, consulte [Aprovisionamiento de clusters de HDInsight](#).

Los contenedores de las cuentas de almacenamiento que están conectados a un cluster: como el nombre y la clave de la cuenta se almacenan en `core-slte.xml`, tendrá acceso total a los blobs de dichos contenedores.

Los contenedores o los blobs públicos de las cuentas de almacenamiento que NO están conectados a un cluster: solo tendrá permiso de lectura en los blobs de los contenedores.

Los contenedores privados de las cuentas de almacenamiento que NO están conectados a un cluster: no puede acceder a los blobs de los contenedores.

Los contenedores del almacenamiento de blobs almacenan los datos como pares de clave/valor y no hay jerarquía de directorios. No obstante, el carácter "/" se puede usar en el nombre de la clave para que parezca que el archivo está almacenado dentro de una estructura de directorios. Por ejemplo, la clave de un blob puede ser `input/logl.txt`. No hay directorios `input`, pero dada la presencia del carácter "/" en el nombre de la clave, parece la ruta de un archivo.

Ventajas del almacenamiento de blobs de Azure

El coste de rendimiento implícito por no tener ubicados juntos la ejecución y el almacenamiento se ve mitigado por el modo en que los clusters de cálculo se aprovisionan cerca de los recursos de la cuenta de almacenamiento, dentro del centro de datos de Azure, donde la red de alta velocidad consigue que los nodos de ejecución sean muy eficientes en el acceso a los datos del almacenamiento de blobs.

Hay varias ventajas asociadas al almacenamiento de datos en blobs en lugar de utilizar HDFS:

Reutilización y uso compartido de los datos: los datos de HDFS se ubican dentro del cluster de cálculo. Solamente las aplicaciones que tengan acceso al cluster de cálculo podrán usar los datos usando la API HDFS. Se puede acceder a los datos del almacenamiento de blobs a través de las API HDFS o las API REST de almacenamiento de blobs. Por lo tanto, se puede usar un conjunto mayor de aplicaciones (incluyendo otros clusters de HDInsight) y herramientas para producir y consumir los datos.

Archivado de datos: almacenar los datos en un almacenamiento de blobs hace que los clusters de HDInsight que se usan para los cálculos se eliminen de forma segura sin perder los datos del usuario.

Coste del almacenamiento de datos: almacenar datos en DFS es más caro a largo plazo que almacenarlos en el almacenamiento de blobs, ya que el coste de un cluster de cálculo es superior al de un contenedor de almacenamiento de blobs. Además, como no hay que volver a cargar los datos para cada generación de cluster de cálculo, también se ahorra en costes de carga de datos.

Escalado horizontal elástico: aunque HDFS proporciona un sistema de archivos escalable en horizontal, la escala se determina en función del número de nodos que aprovisiona para su cluster. Cambiar la escala puede ser un proceso más complicado que basarse en las capacidades de escalado elástico del almacenamiento de blobs que tiene automáticamente.

Replicación geográfica: sus contenedores de almacenamiento de blobs se pueden replicar geográficamente mediante el Portal de Azure. Mientras que esto le aporta recuperación geográfica y redundancia de datos, una conmutación por error en la ubicación replicada geográficamente afectaría gravemente a su rendimiento y podría incurrir

en costes adicionales. Por lo tanto, nuestra recomendación es que elija la replicación geográfica de forma inteligente y únicamente si merece la pena pagar el coste adicional por el valor de los datos.

Determinados trabajos y paquetes de MapReduce podrían crear resultados intermedios que realmente no desea almacenar en el contenedor de almacenamiento de blobs. En tal caso, puede seguir optando por almacenar los datos en el HDFS local. De hecho, HDInsight usa DFS para varios de estos resultados intermedios en los trabajos de Flive y otros procesos.

Preparación de un contenedor para el almacenamiento de blobs

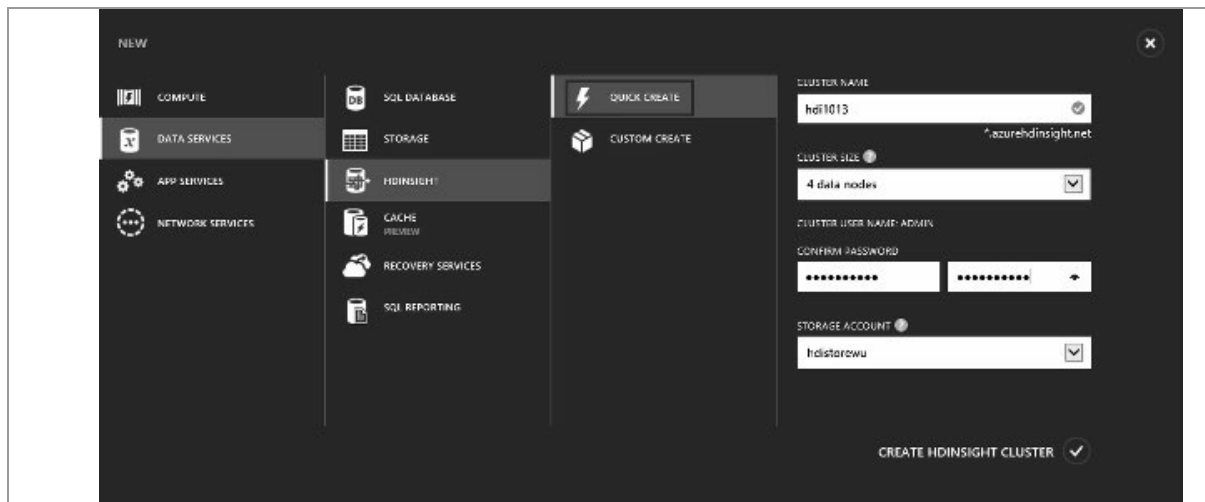
Para usar los blobs, primero debe crear una cuenta de almacenamiento de Azure. Como parte de este proceso, debe especificar un centro de datos de Azure que almacenará los objetos que cree con esta cuenta. Tanto el cluster como la cuenta de almacenamiento deben hospedarse en el mismo centro de datos (además, la base de datos SQL de Hive metastore y la base de datos SQL de Oozie metastore deben estar ubicadas en el mismo centro de datos). Cualquiera que sea su ubicación, todos los blobs que cree pertenecerán a algún contenedor de su cuenta de almacenamiento. Este contenedor puede ser un contenedor de almacenamiento de blobs existente creado fuera de HDInsight, o bien un contenedor que se haya creado para un cluster de HDInsight.

Creación de un contenedor de blobs para HDInsight usando el Portal de administración

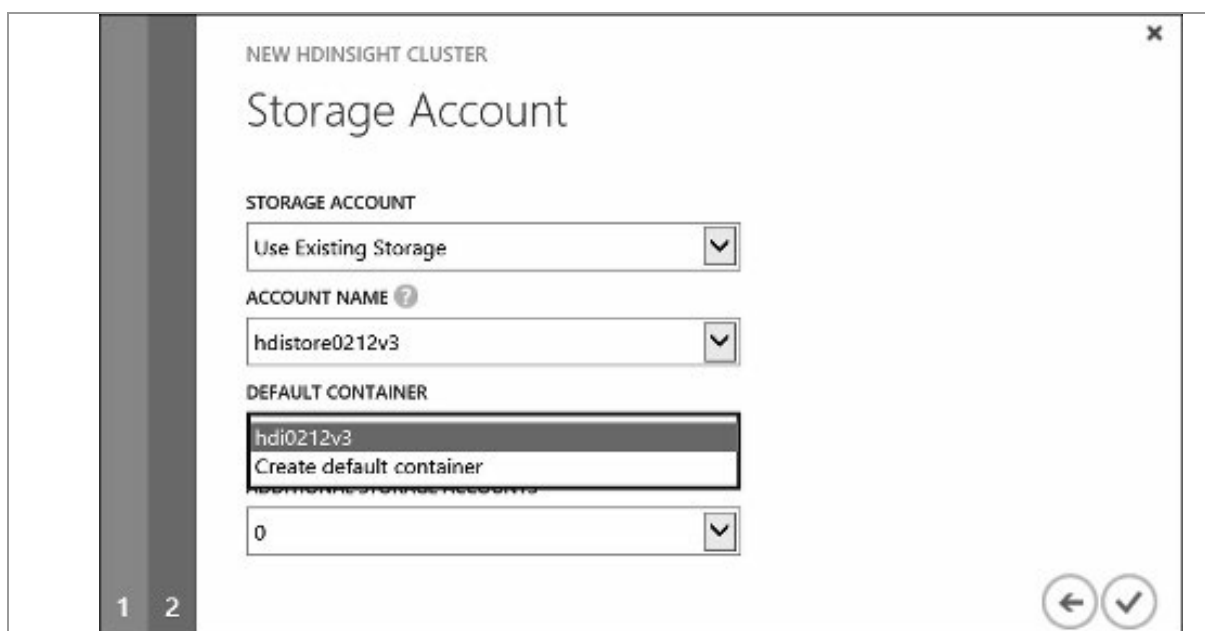
Al aprovisionar un cluster de HDInsight desde el Portal de administración de Azure, tiene dos opciones: quick create y custom create. La opción de creación rápida (quick create) requiere la creación previa de la cuenta de almacenamiento de Azure. Para obtener instrucciones, consulte [Creación de una cuenta de almacenamiento](#).

Al usar la opción de creación rápida, puede elegir una cuenta de almacenamiento existente. El proceso de aprovisionamiento crea un nuevo contenedor con el mismo nombre que el del cluster de HDInsight. Este

contenedor se usa como sistema de archivos predeterminado.



Al usar la creación personalizada (custom create), puede elegir un contenedor de almacenamiento de blobs existente o crear uno predeterminado. El contenedor predeterminado tiene el mismo nombre que el del cluster de HDInsight.



Creación de un contenedor usando Azure PowerShell

Azure PowerShell se puede usar para crear contenedores de blobs. A continuación se muestra un ejemplo de script de PowerShell:

```
$subscriptionName = "  
$storageAccountName =
```

```
$containerName=""
```

```
Add-AzureAccount # La conexión es buena durante 12 horas.
```

```
Select-AzureSubscription $subscriptionName # Obligatorio solamente si tiene varias suscripciones.
```

```
# Cree un objeto de contexto de almacenamiento.
```

```
$storageAccountkey = get-azurestoragekey
```

```
$storageAccountName | %{$_.Primary}
```

```
$destContext = New-AzureStorageContext -
```

```
StorageAccountName $storageAccountName -
```

```
StorageAccountKey $storageAccountKey
```

```
# Cree un contenedor de almacenamiento de blobs.
```

```
New-AzureStorageContainer -Name $containerName -Context
```

```
$destContext
```

Archivos de dirección en almacenamiento de blobs

El esquema URI para obtener acceso a los archivos del almacenamiento de blobs es:

```
wasb [s] ://@.blob.core.Windows.net/
```

El esquema de URI proporciona tanto acceso no cifrado con el prefijo `wasb:` como acceso SSL cifrado con `wasbs:`. Se recomienda usar `wasbs` siempre que sea posible, incluso al obtener acceso a los datos que residen en el mismo centro de datos de Azure.

El valor `id` identifica el nombre del contenedor de almacenamiento de blobs. El valor `sa` identifica el nombre de la cuenta de almacenamiento de Azure. Se necesita el nombre completo de dominio (FQDN).

Si no se ha especificado ningún valor de los anteriores, se usará el sistema de archivos predeterminado. Para los archivos del sistema de archivos predeterminado, puede usar una ruta relativa o absoluta. Por ejemplo, se puede hacer referencia al archivo `hadoop-mapreduce-examples.jar` que viene con los clusters de HDInsight usando uno de los

siguientes:

```
wasb://mycontainer@myaccount.blob.core.Windows.net/example/jars/hadoop-mapreduce-examples.jar  
wasb:///example/jars/hadoop-mapreduce-examples.jar  
/example/jars/hadoop-mapreduce-examples.jar
```

La ruta es el nombre de la ruta HDFS del archivo o el directorio. Como los contenedores de almacenamiento de blobs son solamente un almacén de pares clave-valor, no hay un sistema de archivos jerárquico real. Una “/” dentro de la clave de blob se interpreta como separador de directorios. Por ejemplo, el nombre del blob para hadoop-mapreduce-examples.jar es:

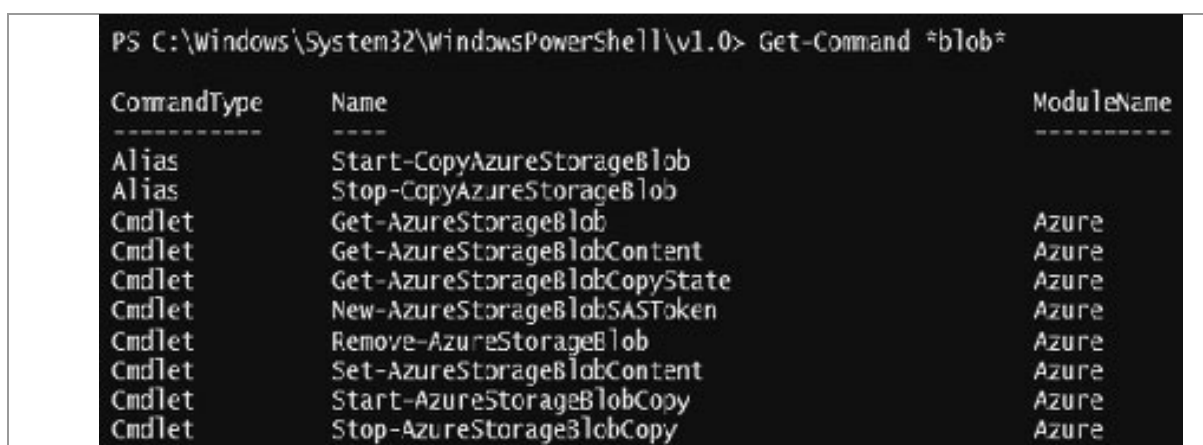
```
example/jars/hadoop-mapreduce-examples.jar
```

Acceso a un blob usando Azure PowerShell

Consulte [Instalación y configuración de Azure PowerShell](#) para obtener información acerca de la instalación y la configuración de Azure PowerShell en su estación de trabajo. Puede usar la ventana de la consola de Azure PowerShell o PowerShellJSE para ejecutar los cmdlets de PowerShell.

Utilice el comando siguiente para incluir los cmdlets relacionados con el blob:

```
Get-Command *blob*
```



```
PS C:\Windows\System32\WindowsPowerShell\v1.0> Get-Command *blob*
```

CommandType	Name	ModuleName
Alias	Start-CopyAzureStorageBlob	
Alias	Stop-CopyAzureStorageBlob	
Cmdlet	Get-AzureStorageBlob	Azure
Cmdlet	Get-AzureStorageBlobContent	Azure
Cmdlet	Get-AzureStorageBlobCopyState	Azure
Cmdlet	New-AzureStorageBlobSASToken	Azure
Cmdlet	Remove-AzureStorageBlob	Azure
Cmdlet	Set-AzureStorageBlobContent	Azure
Cmdlet	Start-AzureStorageBlobCopy	Azure
Cmdlet	Stop-AzureStorageBlobCopy	Azure

Ejemplo de PowerShell para cargar un archivo

Consulte [Carga de datos en HDInsight](#).

Ejemplo de PowerShell para descargar un archivo

El script siguiente descarga un blob en bloques a la carpeta actual. Antes de ejecutar el script, cambie el directorio a una carpeta en la que tenga permiso de escritura.

```
$storageAccountName = "" # La cuenta de almacenamiento
usada para el sistema de archivos predeterminado
especificado en el momento del aprovisionamiento.
$containerName = "" # El contenedor del sistema de
archivos predeterminado tiene el mismo nombre que el
cluster.
$blob = "example/data/sample.log" # El nombre del blob que se va a descargar.

# Use Add-AzureAccount si no se ha conectado a su
suscripción de Azure.
#Add-AzureAccount # La conexión es buena durante 12 horas.

# Use estos dos comandos si tiene varias suscripciones.
#$subscriptionName = ""
#Select-AzureSubscription $subscriptionName

Write-Host "Create a context object ..." -
ForegroundColor Green
$storageAccountKey = Get-AzureStorageKey
$storageAccountName | %{ $_.Primary }
$storageContext = New-AzureStorageContext -
StorageAccountName $storageAccountName -
StorageAccountKey $storageAccountKey

Write-Host "Download the blob ..." -ForegroundColor
Green
Get-AzureStorageBlobContent -Container $ContainerName -
Blob $blob -Context $storageContext -Force
Write-Host "List the downloaded file ..." -
```

ForegroundColor Green

cat “./\$blob”

Ejemplo de PowerShell para eliminar un archivo

\$storageAccountName = “” # La cuenta de almacenamiento

usada para el sistema de archivos predeterminado

especificado en el momento del aprovisionamiento.

\$containerName = “” # El contenedor del sistema de

archivos predeterminado tiene el mismo nombre que el

cluster.

\$blob = “example/data/sample.log” # El nombre del blob

que se va a descargar.

Use Add-AzureAccount si no se ha conectado a su
suscripción de Azure.

#Add-AzureAccount # La conexión es buena durante 12 horas.

Use estos dos comandos si tiene varias suscripciones.

#\$subscriptionName = “”

#Select-AzureSubscription \$subscriptionName

Write-Host “Create a context object ... “ -

ForegroundColor Green

\$storageAccountKey = Get-AzureStorageKey

\$storageAccountName | %{ \$_.Primary }

\$storageContext = New-AzureStorageContext -

StorageAccountName \$storageAccountName -

StorageAccountKey \$storageAccountKey

Write-Host “Delete the blob ...” -ForegroundColor Green

Remove-AzureStorageBlob -Container \$containerName -

Context \$storageContext -blob \$blob

Ejemplo de PowerShell para incluir archivos en una carpeta

```
$storageAccountName = "" # La cuenta de almacenamiento  
usada para el sistema de archivos predeterminado  
especificado en el momento del aprovisionamiento.
```

```
$containerName = "" # El contenedor del sistema de  
archivos predeterminado tiene el mismo nombre que el  
cluster.
```

```
$blobPrefix = "example/data/"
```

```
# Use Add-AzureAccount si no se ha conectado a su  
suscripción de Azure.
```

```
#Add-AzureAccount # La conexión es buena durante 12 horas.
```

```
# Use estos dos comandos si tiene varias suscripciones.
```

```
#$subscriptionName = ""
```

```
#Select-AzureSubscription $subscriptionName
```

```
Write-Host "Create a context object ... " -
```

```
ForegroundColor Green
```

```
$storageAccountKey = Get-AzureStorageKey
```

```
$storageAccountName | %{ $_.Primary }
```

```
$storageContext = New-AzureStorageContext -
```

```
StorageAccountName $storageAccountName -
```

```
StorageAccountKey $storageAccountKey
```

```
Write-Host "List the files in $blobPrefix ..."
```

```
Get-AzureStorageBlob -Container $containerName -Context
```

```
$storageContext -prefix $blobPrefix
```

CAPÍTULO 6

HIVE, PIG, OOZIE, MAPREDUCE Y EXCEL EN HDINSIGHT

UTILIZANDO HIVE CON HDINSIGHT

Apache Hive ofrece el modo de ejecutar un trabajo de MapReduce mediante un lenguaje de scripting de tipo SQL, llamado *HiveQL*, que se puede aplicar al resumen, la consulta y el análisis de grandes volúmenes de datos. En este capítulo, se utilizará HiveQL para consultar los datos de un archivo de registro log4j de Apache y elaborar las estadísticas básicas.

Requisitos previos

Debe aprovisionar un cluster de HDInsight. Para saber cómo hacerlo con el Portal de Azure, consulte *Introducción a HDInsight*. Para obtener instrucciones acerca de otras formas de creación de dichos clusters, consulte *Aprovisionamiento de clusters de HDInsight*.

Debe tener instalado **Azure PowerShell** en su estación de trabajo.

Uso de Hive

Las bases de datos son adecuadas para administrar conjuntos pequeños de datos que admiten consultas de baja latencia. No obstante, si se trata de grandes conjuntos de datos que contienen terabytes de datos, las bases de datos SQL tradicionales no son por lo general la solución ideal. Cada vez con más frecuencia los administradores de las bases de datos se enfrentan a estos grandes conjuntos de datos; y compran hardware más voluminoso a medida que crece la carga de la base de datos y disminuye el rendimiento.

Hive soluciona los problemas asociados a las grandes cantidades de datos permitiendo a los usuarios escalar horizontalmente cuando realizan consultas en grandes conjuntos de datos. Hive consulta los datos en paralelo entre varios nodos que usan MapReduce distribuyendo la base de datos entre un número cada vez mayor de hosts a medida que la carga aumenta.

Hive y HiveQL ofrecen además una alternativa a la escritura de trabajos de MapReduce en Java al realizar consultas en los datos. Proporciona un sencillo contenedor de tipo SQL que permite que las consultas se escriban en HiveQL. A continuación, HDInsight las compila en MapReduce y las ejecuta en el cluster.

Hive además permite que los programadores familiarizados con el marco de MapReduce conecten mapeadores y reductores personalizados para realizar un análisis más sofisticado que podría no ser compatible con las capacidades Integradas del lenguaje de HiveQL.

Hive está destinado al procesamiento de lotes de grandes cantidades de datos Inmutables (como blogs). No es adecuado para aplicaciones a transacciones que necesitan tiempos de respuesta rápidos, como los sistemas de administración de bases de datos. Hive se ha optimizado para la escalabilidad (se pueden añadir dinámicamente más máquinas al cluster de Hadoop), la extensibilidad (dentro del marco de MapReduce y con otras Interfaces de programación) y la tolerancia a errores. La latencia no es una consideración clave del diseño.

Normalmente, las aplicaciones guardan errores, excepciones y otros problemas de código en un archivo de registro, para que los administradores puedan usar los datos de estos archivos de registro a fin de revisar los problemas que pueden surgir y para generar métricas relevantes a los errores u otros problemas, como los de rendimiento. Estos archivos de registro normalmente adquieren grandes tamaños

y contienen una gran cantidad de datos que deben procesarse y extraerse para ofrecer inteligencia en la aplicación.

Por lo tanto, los archivos de registro son ejemplos de grandes cantidades de datos. HDInsight proporciona un sistema de almacén de datos de Hive que facilita un sencillo resumen de los datos, consultas ad hoc y el análisis de estos grandes conjuntos de datos en sistemas de archivos compatibles con Hadoop, como el almacenamiento de blobs de

Azure.

Carga de archivos de datos al almacenamiento de blobs

HDInsight utiliza el contenedor de almacenamiento de blobs de Azure como sistema de archivos predeterminado.

En este capítulo, usará un archivo log4j de muestra distribuido con el cluster de HDInsight que se almacena en `\example\data\sample.log`. Cada registro del archivo consta de una línea de campos que contiene uno llamado [LOG LEVEL] que muestra el tipo y la gravedad. Por ejemplo:

```
2012-02-03 20:26:41 SampleClass3 [ERROR] verbose detail
for id 1527353937
```

Para acceder a los archivos, use la sintaxis siguiente:

```
wasb://@.blob.core.Windows.net
```

Por ejemplo:

```
wasb://mycontainer@mystorage.blob.core.Windows.net/example/data/sample.log
```

Reemplace `mycontainer` por el nombre del contenedor y `mystorage` por el nombre de la cuenta de almacenamiento de blobs.

Como el archivo se almacena en el sistema de archivos predeterminado, también puede acceder al archivo usando lo siguiente:

```
wasb:///example/data/sample.log
/example/data/sample.log
```

Para generar sus propios archivos log4j, use la utilidad de registro Apache Log4j. Para obtener información acerca de cómo cargar datos al almacenamiento de blobs de Azure, consulte [Carga de datos en HDInsight](#).

Ejecución de las consultas de Hive usando PowerShell

En la sección anterior, ha cargado un archivo log4j llamado `sample.log` al contenedor del sistema de archivos predeterminado. Y en esta, ejecutará HiveQL para crear una tabla de Hive, cargar datos en ella y, finalmente,

realizar consultas en los datos para determinar cuántos registros de error había.

Este capítulo proporciona las instrucciones para usar los cmdlets de Azure PowerShell para ejecutar una consulta de Hive. Antes de comenzar con esta sección, debe configurar el entorno local y la conexión a Azure como se indica en la sección de Requisitos previos al inicio de este tema.

Las consultas de Hive se pueden ejecutar en PowerShell usando el cmdlet Start-AzureHDInsightJob o Invoke-Hive.

Para ejecutar las consultas de Hive usando Start-AzureHDInsightJob

- Abra una ventana de la consola de Azure PowerShell. Las instrucciones se encuentran en Instalación y configuración de Azure PowerShell.
- Ejecute el comando siguiente para conectarse a su suscripción de Azure:

```
Add-AzureAccount
```

Se le pedirá que introduzca las credenciales de su cuenta de Azure.

- Configure las variables en el script siguiente y ejecútelo:

```
# Proporcione el nombre de la suscripción de Azure, la cuenta de almacenamiento de Azure y el contenedor que se utilizan para el sistema de archivos predeterminado de HDInsight.
```

```
$subscriptionName = ""
```

```
# Proporcione el nombre del cluster de HDInsight donde desea ejecutar el trabajo de Hive.
```

```
$clusterName = ""
```

- Ejecute el script siguiente para definir las consultas de HiveQL:

```
# Consultas de HiveQL.
```

```
# Use la opción de tabla interna.
```

```
$queryString = "DROP TABLE log4jLogs;" +
```

```
    "CREATE TABLE log4jLogs(t1 string, t2 string, t3 string, t4 string, t5 string, t6 string, t7 string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ' ';"
```

```

+
        "LOAD DATA INPATH
        ' wasb://$containerName@$storageAccountName.blob.core . windows.net/example/data/sample.log OVERWRITE INTO TABLE log4jLogs,-" +
        "SELECT t4 AS sev, COUNT(*) AS cnt FROM log4j Logs
        WHERE t4 = ' [ERROR] ' GROUP BY t4;"

# Use la opción de tabla externa.
$queryString = "DROP TABLE log4jLogs,-" +
        "CREATE EXTERNAL TABLE log4jLogs(t1 String, t2 string, t3
        string, t4 string, t5 string, t6 string, t7 String) ROW FORMAT DELIMITED FIELDS
        TERMINATED BY ' '
        "SELECT t4 AS sev, COUNT(*) AS cnt FROM
        STORED AS TEXTFILE LOCATION
        'wasb://$containerName@$storageAccountName.blob.core.win
        dows.net/example/data/';" +
        log4jLogs WHERE t4 = '[ERROR]' GROUP BY t4;"

```

El comando `LOAD DATA` de HiveQL hará que se mueva el archivo de datos a la carpeta `\hive\warehouse`. El comando `DROP TABLE` eliminará la tabla y el archivo de datos. Si utiliza la opción de tabla interna y desea volver a ejecutar el script, debe cargar el archivo `.log` de muestra de nuevo. Si desea conservar el archivo de datos, debe usar el comando `CREATE EXTERNAL TABLE` como se indica en el script.

También puede usar la tabla externa para cuando el archivo de datos se ubique en un contenedor o una cuenta de almacenamiento diferentes.

Use `DROP TABLE` primero en caso de que ejecute el script de nuevo y que la tabla de `log4jlogs` ya exista.

- Ejecute el script siguiente para crear una definición de trabajo de Hive:

```

# Cree una definición de trabajo de Hive.
$hiveJobDefinition = New-AzureHDInsightHiveJobDefinition
-Query $queryString

```

También usará el conmutador `-File` para especificar un archivo de script de HiveQL en HDFS.

- Ejecute el script siguiente para enviar un trabajo de Hive:

Envíe el trabajo al cluster.

```
Select-AzureSubscription $subscriptionName
```

```
$hiveJob = Start-AzureHDInsightJob -Cluster $clusterName
```

```
-JobDefinition $hiveJobDefinition
```

- Ejecute el script siguiente para esperar a que el trabajo de Hive termine:

Espere a que el trabajo de Hive se termine.

```
Wait-AzureHDInsightJob -Job $hiveJob -
```

```
WaitTimeoutInSeconds 3600
```

- Ejecute el script siguiente para imprimir la salida estándar:

Imprima el error estándar y la salida estándar del trabajo de Hive.

```
Get-AzureHDInsightJobOutput -Cluster $clusterName -JobId
```

```
$hiveJob.JobId -StandardOutput
```

```

Windows PowerShell
PS C:\> ### Create a Hive job definition
PS C:\> $hiveJobDefinition = New-AzureHDInsightHiveJobDefinition -Query $querystring -JobName "HiveJob-log4j"
PS C:\>
PS C:\> ### Submit the job to the cluster
PS C:\> $hiveJob = Start-AzureHDInsightJob -Credentials $creds -Cluster $clusterName -JobDefinition $hiveJobDefinition
PS C:\>
PS C:\> ### Wait for the Hive job to complete
PS C:\> $hiveJob | Wait-AzureHDInsightJob -Credentials $creds -WaitTimeoutInSeconds 3600

StatusDirectory : 613a77aa-20ed-4372-920a-79c121e4d985
ExitCode         : 0
Name             : HiveJob-log4j
Query            : DROP TABLE log4jLogs;CREATE EXTERNAL TABLE log4jLogs(t1 string, t2 string, t3 string, t4 string, t5
                  string, t6 string, t7 string) ROW FORMAT DELIMITED FIELDS TERMINATED BY '|' STORED AS TEXTFILE
                  LOCATION 'wasb://hd1002@hdistorewu.blob.core.windows.net/example/data/log4j/';SELECT t4 AS sev,
                  COUNT(*) AS cnt FROM log4jLogs WHERE t4 = '[ERROR]' GROUP BY t4;
State            : Completed
SubmissionTime   : 10/7/2013 6:21:13 PM
Cluster          : hd1002
PercentComplete :
JobId            : job_201310021625_0053

PS C:\>
PS C:\> ### Print the standard error and the standard output of the Hive job.
PS C:\> Get-AzureHDInsightJobOutput -Cluster $clusterName -Subscription $subscriptionName -JobId $hiveJob.JobId -StandardOutput
[ERROR] 3
PS C:\>

```

El resultado es:

```
[ERROR] 3
```

Para enviar las consultas de Hive usando Invoke-Hive:

- Abra una ventana de la consola de Azure PowerShell.
- Ejecute el comando siguiente para conectarse a su suscripción de Azure:

```
Add-AzureAccount
```

Se le pedirá que introduzca las credenciales de su cuenta de Azure.

- Configure la variable y, a continuación, ejecútela:

```
$clusterName = ""
```

- Ejecute el script siguiente para Invocar las consultas de HlveQL:

```
Use-AzureHDInsightCluster $clusterName
```

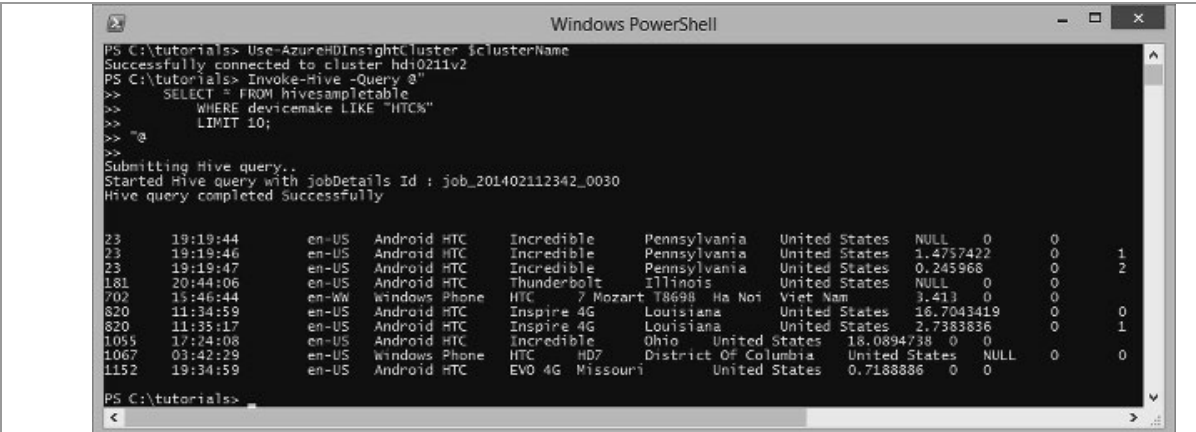
```
$response = Invoke-Hive -Query@"
```

```
    SELECT * FROM hivesampletable
        WHERE devicemake LIKE "HTC%"
        LIMIT 10;
```

```
@"
```

```
Write-Host $response
```

La salida es la siguiente:



```
PS C:\tutorials> Use-AzureHDInsightCluster $clusterName
Successfully connected to cluster hdi0211v2
PS C:\tutorials> Invoke-Hive -Query @"
>>    SELECT * FROM hivesampletable
>>        WHERE devicemake LIKE "HTC%"
>>        LIMIT 10;
>> @"
>>
Submitting Hive query..
Started Hive query with jobDetails Id : job_201402112342_0030
Hive query completed successfully

23  19:19:44  en-US  Android HTC      Incredible  Pennsylvania  United States  NULL  0  0
23  19:19:46  en-US  Android HTC      Incredible  Pennsylvania  United States  1.4757422  0  1
23  19:19:47  en-US  Android HTC      Incredible  Pennsylvania  United States  0.245966  0  2
181 20:44:06  en-US  Android HTC      Thunderbolt Illinois  United States  NULL  0  0
702 13:46:44  en-WW  Windows Phone HTC      7 Mozart T8698 Ha Noi Viet Nam  3.413  0  0
820 11:34:59  en-US  Android HTC      Inspire 4G Louisiana  United States  16.7043419  0  0
820 11:35:17  en-US  Android HTC      Inspire 4G Louisiana  United States  2.7383836  0  1
1055 17:24:08  en-US  Android HTC      Incredible Ohio  United States  18.0894738  0  0
1067 03:42:29  en-US  Windows Phone HTC      HD7 District Of Columbia  United States  NULL  0  0
1152 19:34:59  en-US  Android HTC      EVO 4G Missouri  United States  0.7188886  0  0

PS C:\tutorials>
```

Para consultas de HlveQL más extensas, se recomienda usar PowerShell Here-Strings o el archivo de script de HlveQL. Los ejemplos siguientes indican cómo usar el cmdlet Invoke-Hive para ejecutar un archivo de script de HlveQL. El archivo de script de HlveQL debe cargarse en WASB.

```
Invoke-Hive -File "wasb://@//query.hql"
```

UTILIZANDO PIG CON HDINSIGHT

Apache *Pig* proporciona un lenguaje de scripting para ejecutar trabajos de *MapReduce* como alternativa a la escritura de código Java. En este tutorial, usará PowerShell para ejecutar algunas instrucciones de Pig Latin a fin de analizar un archivo de registro log4j de Apache y ejecutar varias consultas en los datos para generar una salida. En este tutorial se muestran las ventajas de Pig y cómo se puede usar para simplificar los trabajos de MapReduce.

El lenguaje de scripting de Pig se llama *Pig Latin*. Las instrucciones de Pig Latin siguen este flujo general:

Carga: datos leídos que se van a manipular desde el sistema de archivos.

Transformación: manipulación de los datos.

Volcado o almacenamiento: datos de salida en la pantalla o almacenamiento

para su procesamiento.

Requisitos previos

Tenga en cuenta los siguientes requisitos antes de empezar este artículo:

Un cluster de HDInsight de Azure..

Instale y configure Azure PowerShell.

Uso de Pig

Las bases de datos son ideales para pequeños conjuntos de datos y consultas de baja latencia. Sin embargo, cuando nos enfrentamos a grandes conjuntos que contienen terabytes de datos, las bases de datos SQL tradicionales no son la solución ideal. Desde siempre, los administradores de bases de datos han tenido que adquirir hardware de grandes dimensiones a medida que la carga de las bases de datos se incrementaba y el rendimiento disminuía.

Por lo general, todas las aplicaciones guardan errores, excepciones y otros problemas codificados en un archivo de registro para que los administradores puedan revisarlos o generar determinadas métricas a partir de los datos de dicho archivo de registro. Estos archivos de registro normalmente adquieren grandes tamaños y contienen una gran cantidad de datos que deben procesarse y extraerse.

Por lo tanto, los archivos de registro son buenos ejemplos de datos de gran tamaño. Trabajar con datos de gran tamaño es difícil si se usan bases de datos relacionales y paquetes de estadísticas/visualización. Debido a las grandes cantidades de datos y el cálculo de los mismos, a menudo es necesario tener software que se ejecute en paralelo en decenas, cientos o incluso miles de servidores para calcular estos datos en un plazo razonable. Hadoop proporciona un marco de MapReduce para escribir aplicaciones que procesan grandes cantidades de datos estructurados y sin estructurar en paralelo en grandes clusters de máquinas de forma muy fiable y con tolerancia a errores.

Usar Pig reduce el tiempo necesario para escribir programas de asignador y reductor. Es decir, no es necesario usar Java ni código reutilizable. Además, tiene la flexibilidad de combinar el código Java con Pig. Se pueden escribir muchos algoritmos complejos en menos de cinco líneas de código Pig legible para el ojo humano.

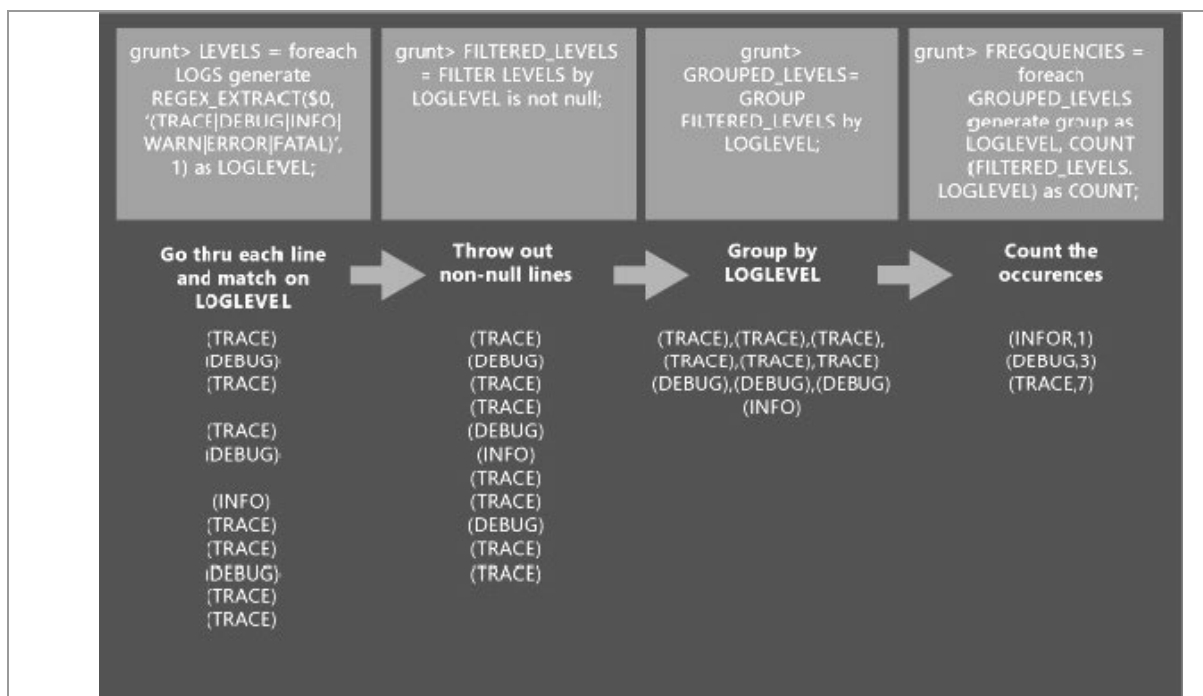
La representación visual de lo que conseguirá en este artículo se muestra en las dos ilustraciones siguientes. Estas ilustraciones muestran un ejemplo representativo del conjunto de datos para ilustrar el flujo y la transformación de los datos a medida que revisa las líneas de código Pig del script. La primera ilustración muestra un ejemplo del archivo log4j:

```
(2012-02-05 19:23:50 SampleClass5 [TRACE] verbose detail for id 313393809)
(2012-02-05 19:23:50 SampleClass6 [DEBUG] detail for id 536603383)
(2012-02-05 19:23:50 SampleClass9 [TRACE] verbose detail for id 564842645)

(2012-02-05 19:23:50 SampleClass8 [TRACE] verbose detail for id 1929822199)
(2012-02-05 19:23:50 SampleClass5 [DEBUG] detail for id 1599724386)

(2012-02-05 19:23:50 SampleClass0 [INFO] everything normal for id 2047808796)
(2012-02-05 19:23:50 SampleClass2 [TRACE] verbose detail for id 1774407365)
(2012-02-05 19:23:50 SampleClass2 [TRACE] verbose detail for id 2099982986)
(2012-02-05 19:23:50 SampleClass4 [DEBUG] detail for id 180683124)
(2012-02-05 19:23:50 SampleClass2 [TRACE] verbose detail for id 1072988373)
(2012-02-05 19:23:50 SampleClass9 [TRACE] verbose detail)
```

La segunda ilustración muestra la transformación de los datos:



Carga de archivos de datos al almacenamiento de blobs

HDInsight utiliza contenedores de almacenamiento de blobs de Azure como sistemas de archivos predeterminados. Para obtener más información, consulte [Uso del almacenamiento de blobs de Azure con HDInsight](#). En este artículo, usará un archivo log4j de muestra distribuido con el cluster de HDInsight que se almacena en `\example\data\sample.log`. Para obtener información acerca de cómo cargar los datos, consulte [Carga de datos en HDInsight](#).

Para acceder a los archivos, use la sintaxis siguiente:

```
wasb[s]://[<containerName>@<storageAccountName>.blob.core.Windows.net]/<pa
```

Por ejemplo:

```
wasb: //mycontainer@mystorage . blob. core . Windows . net/examp  
le/data/sample.log
```

Reemplace *mycontainer* por el nombre del contenedor y *mystorage* por el nombre de la cuenta de almacenamiento de blobs.

Como el archivo se almacena en el sistema de archivos

predeterminado, también puede acceder al archivo usando lo siguiente:

```
wasb:///example/data/sample.log  
/example/data/sample.log
```

Descripción de Pig Latin

En esta sesión repasará algunas de las instrucciones de Pig Latin y los resultados de ejecutar las instrucciones. Y en la siguiente, utilizará PowerShell para ejecutar las instrucciones de Pig.

- Carga de datos desde el sistema de archivos y muestra de los resultados:

```
LOGS = LOAD 'wasb:///example/data/sample.log';  
DUMP LOGS;
```

La salida será similar a la siguiente:

```
(2012-02-05 19:23:50 SampleClass5 [TRACE] verbose  
detail for id 313393809)
```

```
(2012-02-05 19:23:50 SampleClass6 [DEBUG] detail for id  
536603383)
```

```
(2012-02-05 19:23:50 SampleClass9 [TRACE] verbose  
detail for id 564842645)
```

```
(2012-02-05 19:23:50 SampleClass8 [TRACE] verbose  
detail for id 1929822199)
```

```
(2012-02-05 19:23:50 SampleClass5 [DEBUG] detail for id  
1599724386)
```

```
(2012-02-05 19:23:50 SampleClass0 [INFO] everything  
normal for id 2047808796)
```

```
(2012-02-05 19:23:50 SampleClass2 [TRACE] verbose  
detail for id 1774407365)
```

CAPÍTULO 6: HIVE, PIG, OOZIE, MAPREDUCE Y EXCEL EN HDINSIGHT
187

```
(2012-02-05 19:23:50 SampleClass2 [TRACE] verbose  
detail for id 2099982986)
```

```
(2012-02-05 19:23:50 SampleClass4 [DEBUG] detail for id
```

180683124)

(2012-02-05 19:23:50 SampleClass2 [TRACE] verbose
detail for id 1072988373)

(2012-02-05 19:23:50 SampleClass9 [TRACE] verbose
detail)

...

- Revise todas las líneas del archivo de datos para buscar una coincidencia en los seis niveles de registro:

```
LEVELS = foreach LOGS generate REGEX_EXTRACT($0,  
'(TRACE|DEBUG|INFO|WARN|ERROR|FATAL)', 1) as LOGLEVEL;
```

- Filtre las filas que no tengan una coincidencia. Por ejemplo, las filas vacías.

```
FILTEREDLEVELS = FILTER LEVELS by LOGLEVEL is not null; DUMP  
FILTEREDLEVELS;
```

La salida será similar a la siguiente:

(DEBUG)

(TRACE)

(TRACE)

(DEBUG)

(TRACE)

(TRACE)

(DEBUG)

(TRACE)

(TRACE)

(DEBUG)

- Agrupe todos los niveles de registro en su fila correspondiente:

```
GROUPEDLEVELS = GROUP FILTEREDLEVELS by LOGLEVEL;  
DUMP GROUPEDLEVELS;
```

La salida será similar a la siguiente:

(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),

(TRACE),

(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),

(TRACE),
(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),
(TRACE),
(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),
(TRACE),
(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),
(TRACE),
(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),
(TRACE),
(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),
(TRACE),
(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),
(TRACE),
(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),
(TRACE),
(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),(TRACE),
(TRACE),
(TRACE), ...

- En cada grupo, cuente las repeticiones de los niveles de registro. Esta será la frecuencia de cada nivel de registro:

```
FREQUENCIES = foreach GROUPEDELEVELS generate group as  
LOGLEVEL, COUNT(FILTEREDLEVELS.LOGLEVEL) as COUNT;  
DUMP FREQUENCIES;
```

La salida será similar a la siguiente:

```
(INFO,3355)  
(WARN,361)  
(DEBUG,15608)  
(ERROR,181)  
(FATAL,37)  
(TRACE,29950)
```

- Ordene las frecuencias en orden descendente:

```
RESULT = order FREQUENCIES by COUNT dése;
```

DUMP RESULT;

La salida será similar a la siguiente:

(TRACE,29950)

(DEBUG,15608)

(INFO,3355)

(WARN,361)

(ERROR,181)

(FATAL,37)

Ejecución de Pig Latín usando PowerShell

En esta sección se proporcionan las instrucciones para usar los cmdlets de PowerShell. Antes de pasar a esta sección, debe configurar el entorno local y la conexión con Azure.

Para ejecutar Pig Latín usando PowerShell

- Abra una ventana de la consola de Azure PowerShell.
- Ejecute el comando siguiente para conectarse a su suscripción de Azure:

```
Add-AzureAccount
```

Se le pedirá que introduzca las credenciales de su cuenta de Azure.

- Configure la variable del script siguiente y ejecútelo:

```
$clusterName = "<HDInsightClusterName>"
```

- Ejecute los comandos siguientes para definir la cadena de consulta de Pig Latín:

```
# Crear una definición de trabajo de Pig.
```

```
$ 0 = ' $ 0 ' ;
```

```
$QueryString = "LOGS = LOAD
```

```
'wasb:///example/data/sample.log';" +
```

```
    "LEVELS = foreach LOGS generate
```

```
REGEX_EXTRACT($ 0,      ' (TRACE|DEBUG|INFO|WARN|ERROR|FATAL)
```

```
    ,
```

```
    1) as LOGLEVEL;" +
```

```

        "FILTEREDLEVELS = FILTER LEVELS by
LOGLEVEL is not null;" +
        "GROUPEDLEVELS = GROUP FILTEREDLEVELS
by LOGLEVEL;" +
        "FREQUENCIES = foreach GROUPEDLEVELS generate group
as LOGLEVEL,
COUNT(FILTEREDLEVELS.LOGLEVEL) as COUNT;" +
        "RESULT = order FREQUENCIES by COUNT
desc;" +
        "DUMP RESULT;"
$pigJobDefinition = New-AzureHDInsightPigJobDefinition -
Query $QueryString

```

También usará el modificador `-File` para especificar un archivo de script de Pig en HDFS.

- Ejecute el script siguiente para enviar el trabajo de Pig:

```

# Enviar el trabajo de Pig.
Select-AzureSubscription $subscriptionName
$pigJob = Start-AzureHDInsightJob -Cluster $clusterName
-JobDefinition $pigJobDefinition

```

- Ejecute el script siguiente para esperar a que el trabajo de Pig termine:

```

# Esperar hasta que se haya completado el trabajo de
Pig.
Wait-AzureHDInsightJob -Job $pigJob -
WaitTimeoutInSeconds 3600

```

- Ejecute el script siguiente para imprimir la salida del trabajo de Pig:

```

# Imprimir el error estándar y la salida estándar del
trabajo de Pig.
Get-AzureHDInsightJobOutput -Cluster $clusterName -JobId
$pigJob.JobId -StandardOutput

```

```
PS C:\Windows\System32\WindowsPowerShell\v1.0> C:\Users\HDInsight\PowerShell\Pig-UsePig-SubmitJobs.ps1
Submit the Pig job ...
Wait for the Pig job to complete ...

Cluster      : hdi0404
ExitCode     : 0
Name         :
PercentComplete : 100% complete
Query        : LOGS = LOAD 'wasb:///example/data/sample.log';LEVELS = foreach LOGS generate REGEX_EXTRA
              null;GROUPEDELEVELS = GROUP FILTEREDELEVELS by LOGLEVEL;FREQUENCIES = foreach GROUPEDELEVEL
              RESULT;
State        : Completed
StatusDirectory : /tutorials/usepig/status
SubmissionTime : 4/7/2014 3:36:25 PM
JobId        : job_1396629829073_0009

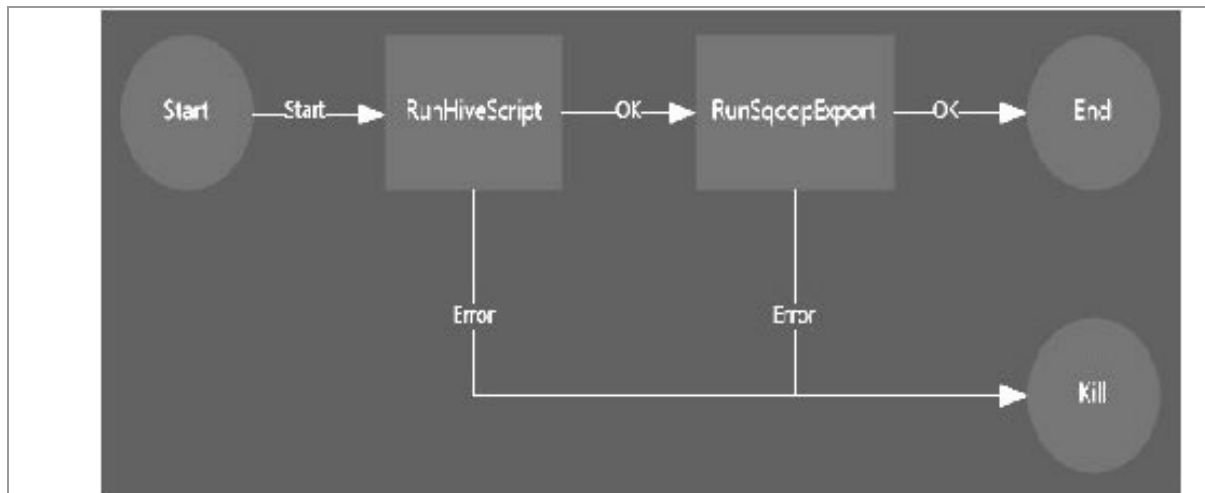
Display the standard output ...
(TRACE,816)
(DEBUG,434)
(INFO,96)
(WARN,11)
(ERROR,6)
(FATAL,2)
```

El trabajo de Pig calcula las frecuencias de los diferentes tipos de registro.

UTILIZANDO OOZIE CON HDINSIGHT

Oozie de Apache es un sistema de coordinación o flujo de trabajo que administra trabajos de Hadoop. Se integra con la pila de Hadoop y es compatible con los trabajos de Hadoop para MapReduce, Pig, Hive y Sqoop de Apache. También puede usarse para programar trabajos específicos de un sistema, como Scripts de shell o programas Java.

El flujo de trabajo que se implementará consta de dos acciones:



- Una acción de Hive ejecuta un script de HiveQL para contar las apariciones de cada tipo de nivel de registro en un archivo de registro log4j. Cada registro log4j consta de una línea de campos que contiene uno llamado [LOG LEVEL] que muestra el tipo y la gravedad. Por ejemplo:

2012-02-03 18:35:34 SampleClass6 [INFO] everything

normal for id 577725851

2012-02-03 18:35:34 SampleClass4 [FATAL] system problem

at id 1991281254

2012-02-03 18:35:34 SampleClass3 [DEBUG] detail for id

1304807656

...

El resultado del script de Hive será parecido al siguiente:

[DEBUG]	434
[ERROR]	3
[FATAL]	1
[INFO]	96
[TRACE]	816
[WARN]	4

Para obtener más información acerca de Hive, consulte [Uso de Hive con HDInsight](#).

- Una acción de Sqoop exporta el resultado de la acción de HiveQL a una tabla en Base de datos SQL de Azure.

Requisitos previos

Antes de empezar este tutorial, debe tener lo siguiente:

Una **estación de trabajo** con Azure PowerShell instalado y configurado. Para ejecutar Scripts de PowerShell, debe ejecutar Azure PowerShell como administrador y establecer la directiva de ejecución en *RemoteSigned*.

Un **cluster de HDInsight**.

Se necesitarán los datos siguientes:

PROPIEDAD DEL CLUSTER

NOMBRE DE VARIABLE DE POWERSHELL

VALOR

DESCRIPCIÓN

Nombre del cluster de HDInsight

\$clusterName

El cluster de HDInsight al que aplicará este tutorial.

Nombre de usuario del cluster de HDInsight

\$clusterUsename

El nombre de usuario del cluster de HDInsight.

Contraseña del usuario del cluster de HDInsight

\$cl usier Pass word

La contraseña de usuario del cluster de HDInsight.

Nombre de la cuenta de almacenamiento de Azure

\$storageAccountName

Cuenta de almacenamiento de Azure disponible para el cluster de HDInsight. Para este tutorial, use la cuenta de almacenamiento predeterminada especificada durante el proceso de aprovisionamiento del cluster.

Nombre del contenedor de blobs|containerName de Azure

containerName

Para este ejemplo, use el contenedor de almacenamiento de blobs de Azure utilizado para el sistema de archivos predeterminado del cluster de HDInsight. De manera predeterminada, tiene el mismo nombre que el del cluster de HDInsight.

Una **Base de datos SQL de Azure**. Debe configurar una regla de firewall para que el servidor de Base de datos SQL permita el acceso desde la estación de trabajo.

PROPIEDAD DE LA BASE DE DATOS SQL

NOMBRE DE VARIABLE DE POWERSHELL

VALOR

DESCRIPCIÓN

Nombre del servidor de base de datos SQL

\$sqlDatabaseServer

El servidor de Base de datos SQL en el que Sqoop exportará los datos.

Nombre de inicio de sesión de la base de datos SQL

\$sqlDatabaseLogin

Nombre de inicio de sesión de la base de datos SQL.

Contraseña de inicio de sesión de la base de datos SQL

\$sqlDatabaseLoginPassword

Contraseña de inicio de sesión de la base de datos SQL.

Nombre de la base de datos SQL

\$sqlDatabaseName

Base de datos SQL de Azure en la que Sqoop exportará los datos.

> [WACOM.NOTE] De forma predeterminada, una base de datos SQL de Azure permite realizar conexiones desde servicios de Azure como HDInsight. Si la configuración del firewall está deshabilitada, debe habilitarla en el Portal de administración de Azure.

Definición del flujo de trabajo de Oozie y el script de HiveQL relacionado

Las definiciones de los flujos de trabajo de Oozie se escriben en hPDL (un lenguaje de definición de proceso XML). El nombre de archivo de flujo de trabajo predeterminado es *workflow.xml*. Guardará el archivo de flujo de trabajo localmente y lo implementará en el cluster de HDInsight con Azure PowerShell posteriormente en este tutorial.

La acción de Hive en el flujo de trabajo llama a un archivo de script de HiveQL. El archivo de script contiene tres instrucciones de HiveQL:

- La instrucción **DROP TABLE** elimina la tabla de Hive log4j en caso de que exista.
- La instrucción **CREATE TABLE** crea una tabla externa de Hive log4j que apunta a la ubicación del archivo de registro log4j. El delimitador de campo es “,”. El delimitador de línea predeterminado es “\n”. La tabla externa de Hive se usa para evitar que el archivo de datos se quite de la ubicación original, en el caso de que desee ejecutar el flujo de trabajo de Oozie varias veces.
- La instrucción **INSERT OVERWRITE** cuenta las apariciones de cada tipo de nivel de registro desde la tabla de Hive log4j y guarda el resultado en una ubicación de blobs de almacenamiento de Azure (WASB).

Hay un problema conocido de la ruta de acceso de Hive. Se producirá este problema cuando envíe un trabajo de Oozie.

Para definir el archivo de script de HiveQL para que lo llame el flujo de trabajo.

- Cree un archivo de texto con el siguiente contenido:

```

DROP TABLE ${hiveTableName};
CREATE EXTERNAL TABLE ${hiveTableName}(t1 string, t2
string, t3 string, t4 string, t5 string, t6 string, t7
string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ‘ ‘
STORED AS TEXTFILE LOCATION ‘${hiveDataFolder}’;
INSERT OVERWRITE DIRECTORY ‘${hiveOutputFolder}’ SELECT
t4 AS sev, COUNT(*) AS cnt FROM ${hiveTableName} WHERE
t4 LIKE ‘[%]’ GROUP BY t4;

```

Existen tres variables que se usan en el script:

\\${hiveTableName}

\\${hiveDataFolder}

\\${hiveOutputFolder}

El archivo de definición de flujo de trabajo (workflow.xml en este tutorial) pasará estos valores al script de HiveQL en tiempo de ejecución.

- Guarde el archivo como C:\Tutorials\UseOozie\useooziewf.hql con la codificación ANSI(ASCII). Use el Bloc de notas si el editor de texto no proporciona la opción. Este archivo de script se implementará en el cluster de HDInsight más tarde en el tutorial.

Para definir un flujo de trabajo

- Cree un archivo de texto con el siguiente contenido:

```

<workflow-app name="useooziewf"
xmlns="uri:oozie:workflow:0.2">
  <start to = "RunHiveScript"/>

  <action name="RunHiveScript">
    <hive xmlns="uri:oozie:hive-action:0.2">
      <job-tracker>${jobTracker}</job-tracker>
      <name-node>${nameNode}</name-node>
      <configuration>
      <property>
        <name>mapred.job.queue.name</name>
        <value>${queueName}</value>

```

```

        </property>
    </configuration>
    <script>${hiveScript}</script>

    <param>hiveTableName=${hiveTableName}</param>

    <param>hiveDataFolder=${hiveDataFolder}</param>

    <param>hiveOutputFolder=${hiveOutputFolder}</param>
    </hive>
    <ok to="RunSqoopExport"/>
CAPÍTULO 6: HIVE, PIG, OOZIE, MAPREDUCE Y EXCEL EN HDINSIGHT
197
    <error to="fail"/>
</action>

<action name="RunSqoopExport">
    <sqoop xmlns="uri:oozie:sqoop-action:0.2">
        <job-tracker>${jobTracker}</job-tracker>
        <name-node>${nameNode}</name-node>
        <configuration>
    <property>

    <name>mapred.compress.map.output</name>
        <value>true</value>
    </property>
    </configuration>
    <arg>export</arg>
    <arg>--connect</arg>
    <arg>${sqlDatabaseConnectionString}</arg>
    <arg>--table</arg>
    <arg>${sqlDatabaseTableName}</arg>

```

```

        <arg>--export-dir</arg>
        <arg>${hiveOutputFolder}</arg>
        <arg>-m</arg>
        <arg>1</arg>
        <arg>--input-fields-terminated-by</arg>
        <arg>" <workflow-app name="useooziewf"
xmlns="uri:oozie:workflow:0.2">
    <start to = "RunHiveScript"/>

    <action name="RunHiveScript">

<hive xmlns="uri:oozie:hive-action:0.2">
    <job-tracker>${jobTracker}</job-tracker>
    <name-node>${nameNode}</name-node>
    <configuration>
        <property>
            <name>mapred.job.queue.name</name>
            <value>${queueName}</value>
        </property>
    </configuration>
    <script>${hiveScript}</script>

<param>hiveTableName=${hiveTableName}</param>

<param>hiveDataFolder=${hiveDataFolder}</param>

<param>hiveOutputFolder=${hiveOutputFolder}</param>
    </hive>
    <ok to="RunSqoopExport"/>
    <error to="fail"/>
</action>

```

```

<action name="RunSqoopExport">
  <sqoop xmlns="uri:oozie:sqoop-action:0.2">
    <job-tracker>${jobTracker}</job-tracker>
    <name-node>${nameNode}</name-node>
    <configuration>
      <property>

<name>mapred.compress.map.output</name>
CAPÍTULO 6: HIVE, PIG, OOZIE, MAPREDUCE Y EXCEL EN HDINSIGHT
199
      <value>true</value>
    </property>
  </configuration>
  <arg>export</arg>
  <arg>--connect</arg>
  <arg>${sqlDatabaseConnectionString}</arg>
  <arg>--table</arg>
  <arg>${sqlDatabaseTableName}</arg>
    <arg>--export-dir</arg>
    <arg>${hiveOutputFolder}</arg>
    <arg>-m</arg>
    <arg>1</arg>
    <arg>--input-fields-terminated-by</arg>
    <arg>"\001"</arg>
  </sqoop>
  <ok to="end"/>
  <error to="fail"/>
</action>

  <kill name="fail">
    <message>Job failed, error
message[${wf:errorMessage(wf:lastErrorNode())}]

```

```

</message>
</kill>

    <end name="end"/>
</workflow-app>
01"</arg>
    </sqoop>
BIG DATA. TÉCNICAS, HERRAMIENTAS Y APLICACIONES
200
<ok to="end"/>
    <error to="fail"/>
</action>

    <kill name="fail">
        <message>Job failed, error
message[${wf:errorMessage(wf:lastErrorNode())}]
</message>
    </kill>
    <end name="end"/>
</workflow-app>

```

Existen dos acciones definidas en el flujo de trabajo. La acción de inicio es *RunHiveScript*. Si la acción ejecuta ok, la acción siguiente es *RunSqoopExport*.

RunHiveScript tiene distintas variables. Pasará los valores cuando envíe el trabajo de Oozie desde la estación de trabajo con Azure PowerShell.

VARIABLES DE FLUJO DE TRABAJO

DESCRIPCIÓN

`${jobTracker}`

Especifique la dirección URL del seguimiento de trabajo de Hadoop. Use `jobtrackerhost:9010` en la versión del cluster de HDInsight 2.0 y 3.0.

`${nameNode}`

Especifique la dirección URL del NameNode de Hadoop. Use la dirección predeterminada de WASB del sistema de archivos. Por ejemplo, `wasb://<containerName>(a),<storageAccountName>.blob.core.windows.net`.

`${queueName}`

Especifica el QueueName al que se enviará el trabajo. Use default.

VARIABLE DE ACCIÓN DE HIVE

DESCRIPCIÓN

`$ {hiveDataFolder}`

El directorio de origen para el comando Create Table de Hive.

`$ {hiveOutputFolder}`

La carpeta de salida para la instrucción INSERT OVERWRITE.

`$ {hiveTableName}`

El nombre de la tabla de Hive que hace referencia a los archivos de datos log4j.

VARIABLE DE ACCIÓN DE SQOOP

DESCRIPCIÓN

`$ {sqlDatabaseConnectionString}`

Cadena de conexión de Base de datos SQL.

`$ {sqlDatabaseTableName}`

La tabla de Base de datos SQL donde se exportarán los datos.

`$ {hiveOutputFolder}`

La carpeta de salida para la instrucción INSERT OVERWRITE de Hive. Esta es la misma carpeta para export-dir de exportación de Sqoop.

- Guarde el archivo como **C:\Tutorials\UseOozie\workflow.xml** con la codificación ANSI(ASCII). Use el Bloc de notas si el editor de texto no proporciona la opción.

Implementación del proyecto de Oozie y preparación del ejemplo

Ejecutará el script de Azure PowerShell para realizar las siguientes acciones:

Copiar el almacenamiento de blobs de Azure del script de HiveQL (useoozie.hql), `wasb:///tutorials/useoozie/useoozie.hql`.

Copiar `workflow.xml` en `wasb:///tutorials/useoozie/workflow.xml`.

Copiar el archivo de datos (`/example/data/sample.log`) en `wasb:///tutorials/useoozie/data/sample.log`.

Crear una tabla de Base de datos SQL para el almacenamiento de datos de exportación de Sqoop. El nombre de tabla es `log4jLogCount`.

Descripción del almacenamiento de HDInsight

HDInsight usa el almacenamiento de blobs de Azure para el almacenamiento de datos. Se llama *WASB* o *Almacenamiento de Azure - Blob*. WASB es la implementación del sistema de archivos distribuido de Hadoop (HDFS) de Microsoft en el almacenamiento de blobs de Azure.

Cuando se aprovisiona un cluster de HDInsight, se designan una cuenta de almacenamiento de Azure y un contenedor de almacenamiento de blobs específico de dicha cuenta como sistema de archivos predeterminado, de la misma forma que en HDFS. Además de esta cuenta de almacenamiento, puede agregar más cuentas de almacenamiento desde la misma suscripción de Azure o desde otras diferentes durante el proceso de aprovisionamiento. Para simplificar el script de PowerShell que se utiliza en este tutorial, todos los archivos se almacenan en el contenedor del sistema de archivos predeterminado, ubicado en `/tutorials/useoozie`. De forma

predeterminada, este contenedor tiene el mismo nombre que el del cluster de HDInsight. La sintaxis de WASB es la siguiente:

```
wasb [s] ://<ContainerName>@<StorageAccountName>.blob.core
.windows.net/<path>/<filename>
```

Para tener acceso a un archivo almacenado en el contenedor del sistema de archivos predeterminado desde HDInsight se puede usar cualquiera de los URI siguientes (use `workflow.xml` como ejemplo):

```
wasb://mycontainer@mystorageaccount.blob.core.Windows.net/tutorials/useoozie/workflow.xml  
wasb:///tutorials/useoozie/workflow.xml  
/tutorials/useoozie/workflow.xml
```

Si desea obtener acceso al archivo directamente desde la cuenta de almacenamiento, el nombre de blob del archivo es:

```
tutorials/useoozie/workflow.xml
```

Descripción de la tabla interna y externa de Hive

Existen determinados aspectos que debe conocer en relación con la tabla Interna y externa de Hive:

El comando `CREATE TABLE` crea una tabla interna, también conocida como tabla administrada. El archivo de datos debe ubicarse en el contenedor predeterminado.

El comando `CREATE TABLE` mueve el archivo de datos a la carpeta `/hlve/warehouse/` en el contenedor predeterminado.

El comando `CREATE EXTERNAL TABLE` crea una tabla externa. El archivo de datos puede estar ubicado fuera del contenedor predeterminado.

El comando `CREATE EXTERNAL TABLE` no mueve el archivo de datos.

El comando `CREATE EXTERNAL TABLE` no permite a las subcarpetas en la carpeta especificada en la cláusula `LOCATION`. Este es el motivo por el que el tutorial hace una copia del archivo `sample.log`.

Para obtener más información, consulte [HDInsight: Hive Internal and External Tables Intro](#).

Para preparar el ejemplo

- Abra Windows **PowerShell ISE** (en la pantalla Inicio de Windows 8, escriba `PowerShellMSE` y, a continuación, haga clic en **Windows PowerShell ISE**. Consulte [Start Windows PowerShell on Windows 8 and Windows](#)).
- En el panel inferior, ejecute el comando siguiente para conectarse a su suscripción de Azure:

```
Add-AzureAccount
```

Se le pedirá que escriba las credenciales de la cuenta de Azure. Este método de agregar una conexión de suscripción expira y, transcurridas 12 horas, tendrá que volver a ejecutar el cmdlet.

- Copie el siguiente script en el panel de scripts y, a continuación, establezca las seis primeras variables.

```
# Variables de WASB
```

```
$storageAccountName = "<StorageAccountName>"
```

```
$containerName = "<BlobStorageContainerName>"
```

```
# Variables de Base de datos SQL
```

```
$sqlDatabaseServer = "<SQLDatabaseServerName>"
```

```
$sqlDatabaseLogin = "<SQLDatabaseLoginName>"
```

```
$sqlDatabaseLoginPassword = "SQLDatabaseLoginPassword">"
```

```
$sqlDatabaseName = "<SQLDatabaseName>"
```

```
$sqlDatabaseTableName = "log4jLogsCount"
```

```
# Archivos de Oozie para el tutorial
```

```
$workflowDefinition =
```

```
"C:\Tutorials\UseOozie\workflow.xml"
```

```
$hiveQLScript = "C:\Tutorials\UseOozie\useooziewf.hql"
```

```
# Carpeta de WASB para el almacenamiento de los  
archivos del tutorial de Oozie.
```

```
$destFolder = "tutorials/useoozie" # No usar la ruta  
de acceso larga aquí
```

```
Para ver más descripciones de las variables, consulte la  
sección Requisitos previos de este tutorial.
```

```
Agregue lo siguiente al script en el panel de scripts:
```

```
# Crear un objeto de contexto de almacenamiento
```

```
$storageaccountkey = get-azurestoragekey
```

```
$storageAccountName | %{$_.Primary}
```

```
$destContext = New-AzureStorageContext -
```

```
StorageAccountName $storageAccountName -  
StorageAccountKey $storageaccountkey
```

```
function uploadOozieFiles()  
{  
    Write-Host "Copy workflow definition and HiveQL  
script file ..." -ForegroundColor Green  
    Set-AzureStorageBlobContent -File  
$workflowDefinition -Container $containerName -Blob  
"$destFolder/workflow.xml" -Context $destContext  
    Set-AzureStorageBlobContent -File $hiveQLScript -  
Container $containerName -Blob  
"$destFolder/useooziewf.hql" -Context $destContext  
}
```

```
function prepareHiveDataFile()  
{  
Write-Host "Make a copy of the sample.log file ...  
" -ForegroundColor Green  
    Start-CopyAzureStorageBlob -SrcContainer  
$containerName -SrcBlob "example/data/sample.log" -  
Context $destContext -DestContainer $containerName -  
destBlob "$destFolder/data/sample.log" -DestContext  
$destContext  
}
```

```
function prepareSQLDatabase()  
{  
    # Cadena de consulta de SQL para la creación de la  
tabla log4jLogsCount  
    $cmdCreateLog4jCountTable = " CREATE TABLE  
[dbo].[$sqlDatabaseTableName](
```

```
        [Level] [nvarchar](10) NOT NULL,  
        [Total] float,  
        CONSTRAINT [PK_$$sqlDatabaseTableName] PRIMARY  
KEY CLUSTERED  
    (  
        [Level] ASC  
    )  
    )”
```

```
    #Crear la tabla log4jLogsCount  
    Write-Host “Create Log4jLogsCount table ...” -  
ForegroundColor Green  
    $conn = New-Object  
System.Data.SqlClient.SqlConnection  
$conn.ConnectionString = “Data  
Source=$sqlDatabaseServer.database.windows.net;Initial  
Catalog=$sqlDatabaseName;User  
ID=$sqlDatabaseLogin;Password=$sqlDatabaseLoginPassword;  
Encrypt=true;Trusted_Connection=false;”  
    $conn.open()  
    $cmd = New-Object System.Data.SqlClient.SqlCommand  
$cmd.connection = $conn  
    $cmd.commandtext = $cmdCreateLog4jCountTable  
    $cmd.executenonquery()  
  
    $conn.close()  
}  
  
# Cargar workflow.xml, coordinator.xml y ooziewf.hql  
uploadOozieFiles;  
  
# Realizar una copia de example/data/sample.log para
```

```
example/data/log4j/sample.log
```

```
prepareHiveDataFile;
```

```
# Crear tabla log4jlogsCount en Base de datos SQL
```

```
prepareSQLDatabase;
```

- Haga clic en **Ejecutar script** o presione **F5** para ejecutar el script. La salida debe ser similar a:

```
Copy workFlow definition and HiveQL script file ...
Container Uri: https://hdistore0212v3.blob.core.windows.net/hdi0212v3
Name                               BlobType  Length  ContentType                LastModified                SnapshotTime
-----
tutorials/useoozie/workflow.xml     BlobBlob  1856    application/octet-stream    2/17/2014 2:37:29 AM +00:00
tutorials/useoozie/useooziwf.hql    BlobBlob  378     application/octet-stream    2/17/2014 2:37:30 AM +00:00
Make a copy of the sample.log file ...
tutorials/useoozie/data/sample.log  BlobBlob  99271   application/octet-stream    2/17/2014 2:37:32 AM +00:00
Create Log4jlogsCount table ...
-1
PS C:\Windows\System32\WindowsPowerShell\v1.0>
```

Ejecución de proyecto de Oozie

Azure PowerShell no proporciona actualmente cmdlets para la definición de trabajos de Oozie. Puede usar el cmdlet de PowerShell Invoke-RestMethod para invocar los servicios web de Oozie. La API de servicios web de Oozie es una API HTTP REST JSON. Para obtener más Información acerca de la API de servicios web de Oozie, consulte la documentación de Oozie 4.0 de Apache (en inglés) para la versión del cluster de HDInsight 3.0 o la documentación de Oozie 3.3.2 de Apache (en inglés) para la versión del cluster de HDInsight 2.1.

Para enviar un trabajo de Oozie

- Abra Windows PowerShell ISE (en la pantalla Inicio de Windows 8, escriba **PowerShellMSE** y, a continuación, haga clic en **Windows PowerShell ISE**).
- Copie el siguiente script en el panel de Scripts y, a continuación, configure las diez primeras variables (omite la sexta, `\$storageUri`).

```
#Variables del cluster de HDInsight
```

```
$clusterName = "<HDInsightClusterName>"
```

```
$clusterUsername = "<HDInsightClusterUsername>"
```

```
$clusterPassword = "<HDInsightClusterUserPassword>"
```

#Variables de almacenamiento de blobs de Azure(WASB)

\$storageAccountName = “<StorageAccountName>”

\$storageContainerName = “<BlobContainerName>”

\$storageUri=”wasb://\$storageContainerName@\$storageAccountName.blob.core.windows.net”

#Variables de Base de datos SQL de Azure

\$sqlDatabaseServer = “<SQLDatabaseServerName>”

\$sqlDatabaseLogin = “<SQLDatabaseLoginName>”

\$sqlDatabaseLoginPassword =

“<SQLDatabaseLoginPassword>”

\$sqlDatabaseName = “<SQLDatabaseName>”

#Variables WF de Oozie

\$oozieWFPath=”\$storageUri/tutorials/useoozie” # El nombre predeterminado es workflow.xml y no necesita especificar el nombre de archivo.

\$waitTimeBetweenOozieJobStatusCheck=10

#Variables de acción de Hive

\$hiveScript =

“\$storageUri/tutorials/useoozie/useooziewf.hql”

\$hiveTableName = “log4jlogs”

\$hiveDataFolder = “\$storageUri/tutorials/useoozie/data”

\$hiveOutputFolder =

“\$storageUri/tutorials/useoozie/output”

#Variables de acción de Sqoop

\$sqlDatabaseConnectionString =

“jdbc:sqlserver://\$sqlDatabaseServer.database.windows.net;

user=\$sqlDatabaseLogin@\$sqlDatabaseServer;password=\$sq

```
lDatabaseLoginPassword;database=$sqlDatabaseName”  
$sqlDatabaseTableName = “log4jLogsCount”
```

```
$passwd = ConvertTo-SecureString $clusterPassword -  
AsPlainText -Force  
$creds = New-Object  
System.Management.Automation.PSCredential  
($clusterUsername, $passwd)
```

- Agregue lo siguiente al script. Esta parte define la carga de Oozie:

```
#Carga de Oozie usada para el envío del servicio web de
```

```
Oozie
```

```
$OoziePayload = @"
```

```
<
```

```
xml version="1.0" encoding="UTF-8"
```

```
>
```

```
<configuration>
```

```
<property>
```

```
  <name>nameNode</name>
```

```
  <value>$storageUri</value>
```

```
</property>
```

```
<property>
```

```
  <name>jobTracker</name>
```

```
  <value>jobtrackerhost:9010</value>
```

```
</property>
```

```
<property>
```

```
  <name>queueName</name>
```

```
  <value>default</value>
```

```
</property>
```



```
<property>
  <name>oozie.use.system.libpath</name>
  <value>>true</value>
</property>
```

```
<property>
<name>hiveScript</name>
  <value>$hiveScript</value>
</property>
```

```
<property>
  <name>hiveTableName</name>
  <value>$hiveTableName</value>
</property>
```

```
  <property>
    <name>hiveDataFolder</name>
    <value>$hiveDataFolder</value>
  </property>
```

```
  <property>
    <name>hiveOutputFolder</name>
    <value>$hiveOutputFolder</value>
  </property>
```

```
  <property>
    <name>sqlDatabaseConnectionString</name>
<value>&quot;,$sqlDatabaseConnectionString&quot;</value>
  </property>
```

```
  <property>
    <name>sqlDatabaseTableName</name>
```

```
        <value>$$SQLDatabaseTableName</value>
</property>
```

```
<property>
    <name>user.name</name>
    <value>admin</value>
</property>
```

```
<property>
    <name>oozie.wf.application.path</name>
    <value>$$oozieWFPath</value>
</property>
```

```
</configuration>
```

```
“@
```

- Agregue lo siguiente al script. En esta parte se comprueba el estado del servicio web de Oozie:

```
Write-Host “Checking Oozie server status...” -
ForegroundColor Green
$clusterUriStatus =
“https://$clusterName.azurehdinsight.net:443/oozie/v2/ad
min/status”
$response = Invoke-RestMethod -Method Get -Uri
$clusterUriStatus -Credential $creds -OutVariable
$OozieServerStatus

$jsonResponse = ConvertFrom-Json (ConvertTo-Json -
InputObject $response)
$oozieServerSatus = $jsonResponse[0].(“systemMode”)
Write-Host “Oozie server status is
$oozieServerSatus...”
```

- Agregue lo siguiente al script. En esta parte se crea e inicia un

trabajo de Oozie:

```
# Crear trabajo de Oozie
```

```
Write-Host "Sending the following Payload to the  
cluster:" -ForegroundColor Green
```

```
Write-Host "`n-----`n$OoziePayload`n-----"
```

```
$clusterUriCreateJob =
```

```
"https://$clusterName.azurehdinsight.net:443/oozie/v2/jo  
bs"
```

```
$response = Invoke-RestMethod -Method Post -Uri
```

```
$clusterUriCreateJob -Credential $creds -Body
```

```
$OoziePayload -ContentType "application/xml" -
```

```
OutVariable $OozieJobName #-debug
```

```
$jsonResponse = ConvertFrom-Json (ConvertTo-Json -  
InputObject $response)
```

```
$oozieJobId = $jsonResponse[0].("id")
```

```
Write-Host "Oozie job id is $oozieJobId..."
```

```
# Iniciar trabajo de Oozie
```

```
Write-Host "Starting the Oozie job $oozieJobId..." -
```

```
ForegroundColor Green
```

```
$clusterUriStartJob =
```

```
"https://$clusterName.azurehdinsight.net:443/oozie/v2/jo
```

```
b/" + $oozieJobId + "
```

```
action=start"
```

```
$response = Invoke-RestMethod -Method Put -Uri
```

```
$clusterUriStartJob -Credential $creds | Format-Table -
```

```
HideTableHeaders #-debug
```

- Agregue lo siguiente al script. En esta parte se comprueba el estado del trabajo de Oozie:

```
# Obtener el estado del trabajo
```

```
Write-Host "Sleeping for
```

```
$waitTimeBetweenOozieJobStatusCheck seconds until the  
job metadata is populated in the Oozie metastore...” -
```

```
ForegroundColor Green
```

```
Start-Sleep -Seconds
```

```
$waitTimeBetweenOozieJobStatusCheck
```

```
Write-Host “Getting job status and waiting for the job  
to complete...” -ForegroundColor Green
```

```
$clusterUriGetJobStatus =
```

```
“https://$clusterName.azurehdinsight.net:443/oozie/v2/jo  
b/” + $oozieJobId + “  
show=info”
```

```
$response = Invoke-RestMethod -Method Get -Uri
```

```
$clusterUriGetJobStatus -Credential $creds
```

```
$jsonResponse = ConvertFrom-Json (ConvertTo-Json -  
InputObject $response)
```

```
$JobStatus = $jsonResponse[0].(“status”)
```

```
while($JobStatus -notmatch “SUCCEEDED|KILLED”)  
{
```

```
    Write-Host “$(Get-Date -format ‘G’): $oozieJobId is  
in $JobStatus state...waiting
```

```
$waitTimeBetweenOozieJobStatusCheck seconds for the job  
to complete...”
```

```
Start-Sleep -Seconds
```

```
$waitTimeBetweenOozieJobStatusCheck
```

```
    $response = Invoke-RestMethod -Method Get -Uri
```

```
$clusterUriGetJobStatus -Credential $creds
```

```
    $jsonResponse = ConvertFrom-Json (ConvertTo-Json -  
InputObject $response)
```

```
    $JobStatus = $jsonResponse[0].(“status”)
```

```
}
```

Write-Host “\$(Get-Date -format ‘G’): \$oozieJobId is in
\$JobStatus state!” -ForegroundColor Green

- Si el cluster de HDinsight es la versión 2.1, reemplace “https://\$clusterName.azurehdinsight.net:443/oozie/v2/” por “https://\$clusterName.azurehdinsight.net:443/oozie/v1/”. La versión del cluster de HDinsight 2.1 no es compatible con la versión 2 de los servicios web.
- Haga clic en **Ejecutar script** o presione **F5** para ejecutar el script. La salida debe ser similar a:

```
Oozie job id is 0000009-140211234238000-oozie-hdp-W...
Starting the Oozie job 0000009-140211234238000-oozie-hdp-W...
Sleeping for 10 seconds until the job metadata is populated in the Oozie metastore...
Getting job status and waiting for the job to complete...
2/16/2014 9:02:27 PM: 0000009-140211234238000-oozie-hdp-W is in RUNNING state...waiting 10 seconds for the job to complete...
2/16/2014 9:02:38 PM: 0000009-140211234238000-oozie-hdp-W is in RUNNING state...waiting 10 seconds for the job to complete...
2/16/2014 9:02:49 PM: 0000009-140211234238000-oozie-hdp-W is in RUNNING state...waiting 10 seconds for the job to complete...
2/16/2014 9:02:59 PM: 0000009-140211234238000-oozie-hdp-W is in RUNNING state...waiting 10 seconds for the job to complete...
2/16/2014 9:03:10 PM: 0000009-140211234238000-oozie-hdp-W is in RUNNING state...waiting 10 seconds for the job to complete...
2/16/2014 9:03:21 PM: 0000009-140211234238000-oozie-hdp-W is in RUNNING state...waiting 10 seconds for the job to complete...
2/16/2014 9:03:32 PM: 0000009-140211234238000-oozie-hdp-W is in RUNNING state...waiting 10 seconds for the job to complete...
2/16/2014 9:03:42 PM: 0000009-140211234238000-oozie-hdp-W is in RUNNING state...waiting 10 seconds for the job to complete...
2/16/2014 9:03:53 PM: 0000009-140211234238000-oozie-hdp-W is in RUNNING state...waiting 10 seconds for the job to complete...
2/16/2014 9:04:04 PM: 0000009-140211234238000-oozie-hdp-W is in RUNNING state...waiting 10 seconds for the job to complete...
2/16/2014 9:04:14 PM: 0000009-140211234238000-oozie-hdp-W is in SUCCEEDED state!
Done!
PS C:\Windows\System32\WindowsPowerShell\v1.0>
```

- Conéctese a la Base de datos SQL para ver los datos exportados.

Para comprobar el registro de errores del trabajo

Para solucionar los problemas de un flujo de trabajo, puede encontrar el archivo de registro de Oozie en C:\apps\dlst\oozie-3.3.2.1.3.2.0-05\oozie-win-dlstro\logs\Oozie.log o en C:\apps\dlst\oozie-4.0.0.2.0.7.0-1528\oozie-win-dlstro\logs\Oozie.log desde el nodo principal del cluster. Para obtener Información acerca de RDP, consulte Administración de los clusters de HDInsight mediante el Portal de administración.

Para volver a ejecutar el tutorial

Para volver a ejecutar el flujo de trabajo, debe realizar lo siguiente:

Eliminar el archivo de salida del script de Hive.

Eliminar los datos en la tabla log4jLogsCount.

Aquí tiene un script de PowerShell de ejemplo que puede usar:

```
$storageAccountName = “<WindowsAzureStorageAccountName>”
```

```
$containerName = “<ContainerName>”
```

```
#Variables de Base de datos SQL
```

```
$sqlDatabaseServer = "<SQLDatabaseServerName>"
$sqlDatabaseLogin = "<SQLDatabaseLoginName>"
$sqlDatabaseLoginPassword = "<SQLDatabaseLoginPassword>"
$sqlDatabaseName = "<SQLDatabaseName>"
$sqlDatabaseTableName = "log4jLogsCount"
```

Write-host "Delete the Hive script output file ..." -

```
ForegroundColor Green
```

```
$storageaccountkey = get-azurestoragekey
```

```
$storageAccountName | %{$_.Primary}
```

```
$destContext = New-AzureStorageContext -
```

```
StorageAccountName $storageAccountName -
```

```
StorageAccountKey $storageaccountkey
```

```
Remove-AzureStorageBlob -Context $destContext -Blob
```

```
"tutorials/useoozie/output/000000_0" -Container
```

```
$containerName
```

Write-host "Delete all the records from the

log4jLogsCount table ..." -ForegroundColor Green

```
$conn = New-Object System.Data.SqlClient.SqlConnection
```

```
$conn.ConnectionString = "Data
```

```
Source=$sqlDatabaseServer.database.windows.net;Initial
```

```
Catalog=$sqlDatabaseName;User
```

```
ID=$sqlDatabaseLogin;Password=$sqlDatabaseLoginPassword;
```

```
Encrypt=true;Trusted_Connection=false;"
```

```
$conn.open()
```

```
$cmd = New-Object System.Data.SqlClient.SqlCommand
```

```
$cmd.connection = $conn
```

```
$cmd.commandtext = "delete from $sqlDatabaseTableName"
```

```
$cmd.executenonquery()
```

```
$conn.close()
```

DESARROLLO DE PROGRAMAS MAPREDUCE DE JAVA PARA HDINSIGHT

El ejemplo que se presenta le guiará por un escenario integral, desde el desarrollo y la prueba de un trabajo de MapReduce para el recuento de palabras en el emulador de HDInsight hasta su implementación y ejecución en HDInsight de Azure.

Requisitos previos

- Instale el emulador de HDInsight de Azure.
- Instale Azure PowerShell en el equipo emulador.
- Obtenga una suscripción de Azure.

Desarrollo de un programa de MapReduce para el recuento de palabras en Java

El recuento de palabras es una aplicación sencilla que cuenta las repeticiones de cada palabra en un conjunto de entrada especificado.

Para escribir el trabajo de MapReduce para el recuento de palabras en Java

- Abra el Bloc de notas.
- Copie y pegue el siguiente programa en el Bloc de notas.

```
package org.apache.hadoop.examples;  
import java.io.IOException;  
import java.util.StringTokenizer;  
import org.apache.hadoop.conf.Configuration;  
import org.apache.hadoop.fs.Path;  
import org.apache.hadoop.io.IntWritable;  
import org.apache.hadoop.io.Text;
```

```

import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import
org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import
org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.util.GenericOptionsParser;

public class WordCount {

    public static class TokenizerMapper
        extends Mapper<Object, Text, Text, IntWritable>{

        private final static IntWritable one = new
IntWritable(1);
        private Text word = new Text();
        public void map(Object key, Text value, Context
context
                                ) throws IOException,
InterruptedException {
            StringTokenizer itr = new
StringTokenizer(value.toString());
            while (itr.hasMoreTokens()) {
word.set(itr.nextToken());
                context.write(word, one);
            }
        }
    }

    public static class IntSumReducer
        extends
Reducer<Text,IntWritable,Text,IntWritable> {

```



```

private IntWritable result = new IntWritable();

public void reduce(Text key, Iterable<IntWritable>
values,
                                Context context
                                ) throws IOException,
InterruptedException {
    int sum = 0;
    for (IntWritable val : values) {
        sum += val.get();
    }
    result.set(sum);
    context.write(key, result);
} }

```

```

public static void main(String[] args) throws
Exception {
    Configuration conf = new Configuration();
    String[] otherArgs = new GenericOptionsParser(conf,
args).getRemainingArgs();
    if (otherArgs.length != 2) {
        System.err.println("Usage: wordcount <in>
<out>");
        System.exit(2);
    }
    Job job = new Job(conf, "word count");
    job.setJarByClass(WordCount.class);
    job.setMapperClass(TokenizerMapper.class);
    job.setCombinerClass(IntSumReducer.class);
    job.setReducerClass(IntSumReducer.class);
    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(IntWritable.class);
}

```

```

        FileInputFormat.addInputPath(job, new
Path(otherArgs[0]));
        FileOutputFormat.setOutputPath(job, new
Path(otherArgs[1]));
        System.exit(job.waitForCompletion(true) ? 0 : 1);
    } }

```

Tenga en cuenta que el nombre del paquete es **org.apache.hadoop.examples** y el nombre de clase es **WordCount**. Utilizará los nombres cuando envíe el trabajo de MapReduce.

- Guarde el archivo como **c:\Tutorials\WordCount\WordCount.java**.

Cree la estructura de carpeta si no existe.

El emulador de HDInsight incluye el compilador javac.

Para compilar el programa de MapReduce

- Abra el símbolo del sistema.
- Cambie el directorio a **c:\Tutorials\WordCount**. Esta es la carpeta para el programa de MapReduce para el recuento de palabras.
- Ejecute el comando siguiente para comprobar la existencia de los dos archivos jar:

```
dir %hadoop_home%\hadoop-core-1.1.0-SNAPSHOT.jar
```

```
dir %hadoop_home%\lib\commons-cli-1.2.jar
```

- Ejecute el siguiente comando para compilar el programa:

```
C:\Hadoop\java\bin\javac -classpath
```

```
%hadoop_home%\hadoop-core-1.1.0-
```

```
SNAPSHOT.jar;%hadoop_home%\lib\commons-cli-1.2.jar
```

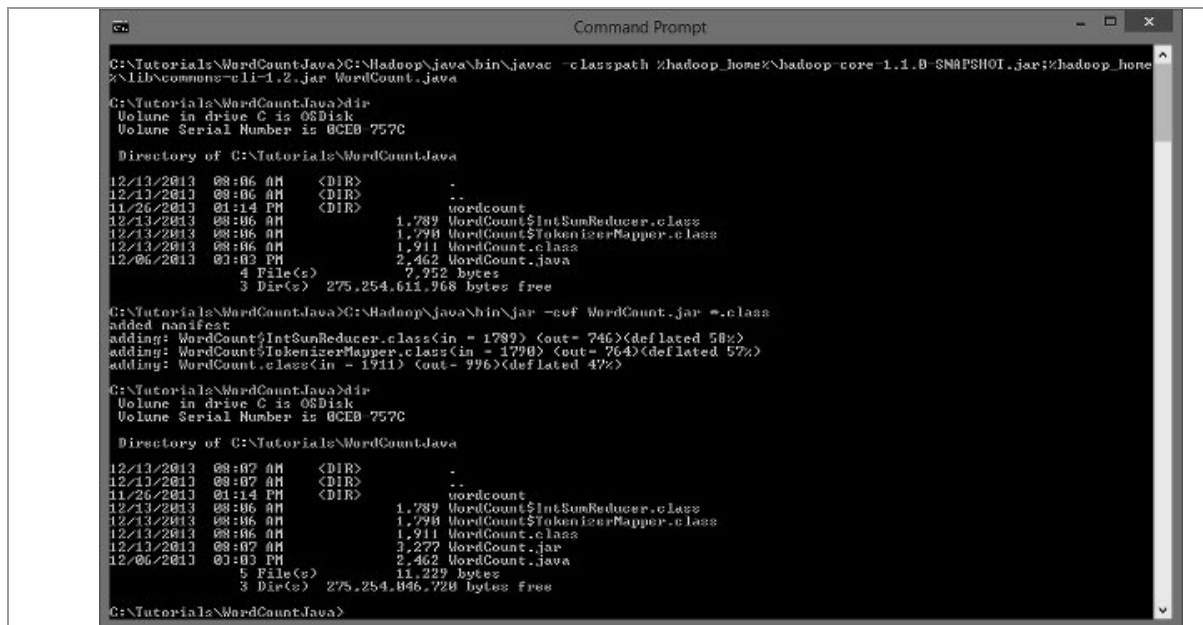
```
WordCount.java
```

javac está ubicado en la carpeta C:\Hadoop\java\bin. El último parámetro es el programa Java que se encuentra en la carpeta actual. El compilador crea tres archivos de clase en la carpeta actual.

- Ejecute el siguiente comando para crear un archivo jar:

```
C:\Hadoop\java\bin\jar -cvf WordCount.jar *.class
```

El comando crea un archivo WordCount.jar en la carpeta actual.



```
C:\Tutorials\WordCountJava>C:\Hadoop\java\bin\javac -classpath %hadoop_home%\hadoop-core-1.1.0-SNAPSHOT.jar;%hadoop_home%\lib\commons-cli-1.2.jar WordCount.java

C:\Tutorials\WordCountJava>dir
Volume in drive C is OSDisk
Volume Serial Number is 0CEB 257C

Directory of C:\Tutorials\WordCountJava

12/12/2013  08:06 AM  <DIR>          .
12/12/2013  08:06 AM  <DIR>          ..
11/26/2013  01:14 PM  <DIR>          wordcount
12/13/2013  08:06 AM             1,789 WordCount$IntSumReducer.class
12/13/2013  08:06 AM             1,798 WordCount$TokenizerMapper.class
12/13/2013  08:06 AM             1,911 WordCount.class
12/06/2013  03:02 PM             2,462 WordCount.java
               4 File(s)              7,952 bytes
               3 Dir(s)          275,254,611,968 bytes free

C:\Tutorials\WordCountJava>C:\Hadoop\java\bin\jar -cf WordCount.jar *.class
added manifest
adding: WordCount$IntSumReducer.class(in = 1789) (out= 746)(deflated 58%)
adding: WordCount$TokenizerMapper.class(in = 1798) (out= 764)(deflated 57%)
adding: WordCount.class(in = 1911) (out= 796)(deflated 47%)

C:\Tutorials\WordCountJava>dir
Volume in drive C is OSDisk
Volume Serial Number is 0CEB 257C

Directory of C:\Tutorials\WordCountJava

12/12/2013  08:07 AM  <DIR>          .
12/12/2013  08:07 AM  <DIR>          ..
11/26/2013  01:14 PM  <DIR>          wordcount
12/13/2013  08:06 AM             1,789 WordCount$IntSumReducer.class
12/13/2013  08:06 AM             1,798 WordCount$TokenizerMapper.class
12/13/2013  08:06 AM             1,911 WordCount.class
12/12/2013  08:07 AM             2,272 WordCount.jar
12/06/2013  03:03 PM             2,462 WordCount.java
               5 File(s)              11,229 bytes
               3 Dir(s)          275,254,846,728 bytes free

C:\Tutorials\WordCountJava>
```

Prueba del programa en el emulador

La prueba del trabajo de MapReduce en el emulador incluye los procedimientos siguientes:

- Cargar los archivos de datos en el HDFS en el emulador
- Enviar un trabajo de mapreduce para el recuento de palabras
- Comprobar el estado del trabajo
- Recuperar los resultados del trabajo

De manera predeterminada, el emulador de HDInsight utiliza HDFS como sistema de archivos predeterminado. Opcionalmente, puede configurar el emulador de HDInsight para utilizar el almacenamiento de blobs de Azure.

En este tutorial, utilizará el comando copyFromLocal de HDFS para cargar los archivos de datos a HDFS. La siguiente sección muestra cómo cargar archivos con Azure PowerShell en el almacenamiento de blobs de Azure.

Este tutorial utiliza la siguiente estructura de carpetas de HDFS:

Carpeta

Nota:

`/WordCount`

La carpeta raíz para el proyecto para el recuento de palabras.

`/WordCount/Apps`

La carpeta para los ejecutables de los programas asignador y reductor.

`/WordCount/Input`

La carpeta de los archivos de origen de MapReduce.

`/WordCount/Output`

La carpeta de los archivos de resultados de MapReduce.

`/WordCount/MRStatusOutput`

La carpeta de resultados del trabajo.

Este tutorial utiliza los archivos `.txt` ubicados en el directorio `%hadoop_home%` como archivos de datos.

Los comandos HDFS de Hadoop distinguen mayúsculas de minúsculas.

Para copiar los archivos de datos al HDFS del emulador

- Abra la línea de comandos de Hadoop desde el escritorio. La línea de comandos de Hadoop es instalada por el instalador del emulador.
- Desde la ventana de la línea de comandos de Hadoop, ejecute el comando siguiente para crear un directorio para los archivos de entrada:

```
hadoop fs -mkdir /WordCount/Input
```

La ruta de acceso utilizada aquí es la ruta de acceso relativa. Es equivalente a la siguiente:

```
hadoop fs -mkdir hdfs://localhost:8020/WordCount/Input
```

- Ejecute el comando siguiente para copiar algunos archivos de texto en la carpeta de entrada en HDFS:

```
hadoop fs -copyFromLocal %hadoop_home%\*.txt
```

```
/WordCount/Input
```

El trabajo de MapReduce contará las palabras de estos archivos.

- Ejecute el comando siguiente para enumerar y comprobar los

archivos cargados:

hadoop fs -ls /WordCount/Input

Verá unos ocho archivos .txt.

Para ejecutar el trabajo de MapReduce mediante la línea de comandos de Hadoop

- Abra la línea de comandos de Hadoop desde el escritorio.
- Ejecute el comando siguiente para eliminar la estructura de carpetas /WordCount/Output de HDFS. /WordCount/Output es la carpeta de resultados del trabajo de MapReduce para el recuento de palabras. El trabajo de MapReduce producirá un error si la carpeta ya existe. Este paso es necesario si es la segunda vez que ejecuta el trabajo.

hadoop fs -rmr /WordCount/Output

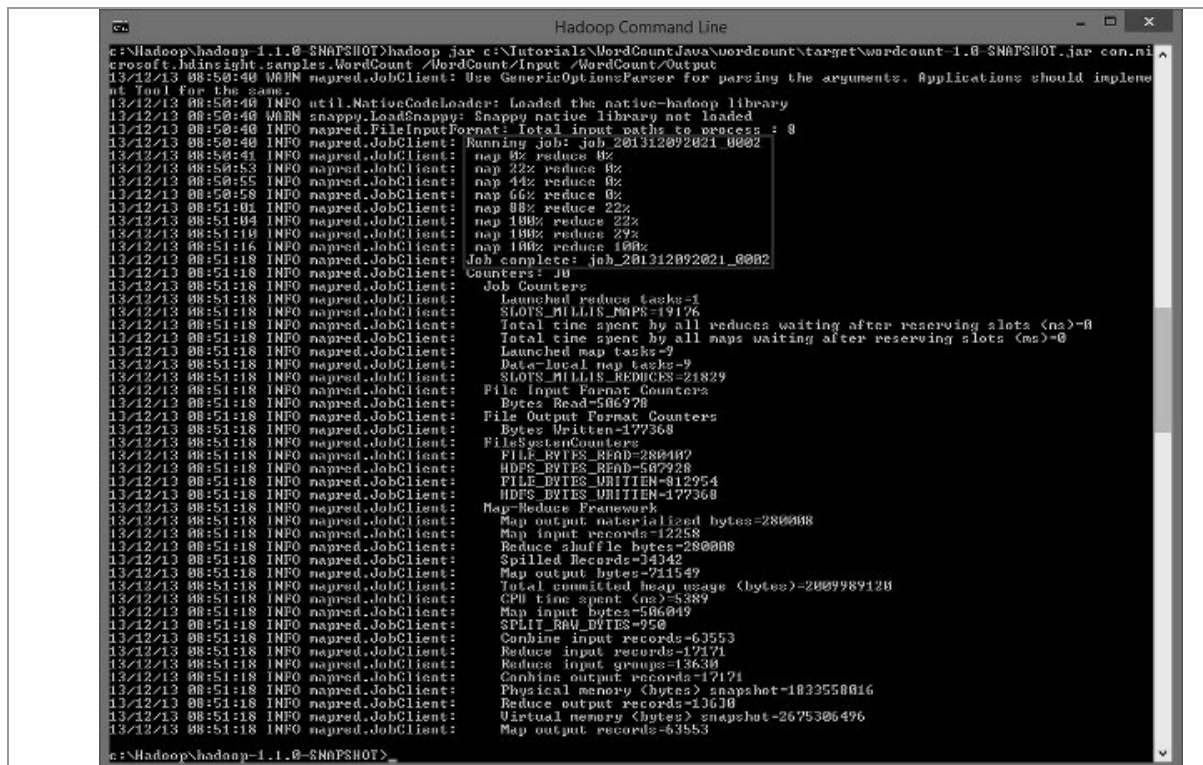
- Ejecute el siguiente comando:

hadoop jar

c:\Tutorials\WordCount Java\wordcount\target\wordcount-

1.0-SNAPSHOT.jar org.apache.hadoop.examples.WordCount

/WordCount/Input /WordCount/Output



```
c:\Hadoop\hadoop-1.1.0-SNAPSHOT>hadoop jar c:\Tutorials\WordCount Java\wordcount\target\wordcount-1.0-SNAPSHOT.jar con.ni
crossft.hbinsight.samples.WordCount /WordCount/Input /WordCount/Output
13/12/13 08:50:40 WARN mapped.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement
Tool for the same.
13/12/13 08:50:40 INFO util.NativeCodeLoader: Loaded the native-hadoop library
13/12/13 08:50:55 INFO mapped.JobClient: map 44% reduce 0%
13/12/13 08:50:40 INFO mapped.FileInputFormat: Total input paths to process : 8
13/12/13 08:50:40 INFO mapped.JobClient: Running job: job_201312092021_0002
13/12/13 08:50:41 INFO mapped.JobClient: map 0% reduce 0%
13/12/13 08:50:53 INFO mapped.JobClient: map 22% reduce 0%
13/12/13 08:50:55 INFO mapped.JobClient: map 44% reduce 0%
13/12/13 08:50:58 INFO mapped.JobClient: map 66% reduce 0%
13/12/13 08:51:01 INFO mapped.JobClient: map 88% reduce 22%
13/12/13 08:51:04 INFO mapped.JobClient: map 100% reduce 22%
13/12/13 08:51:10 INFO mapped.JobClient: map 100% reduce 29%
13/12/13 08:51:16 INFO mapped.JobClient: map 100% reduce 100%
13/12/13 08:51:18 INFO mapped.JobClient: Job samples: job_201312092021_0002
13/12/13 08:51:18 INFO mapped.JobClient: Counters: 10
13/12/13 08:51:18 INFO mapped.JobClient: Job Counters
13/12/13 08:51:18 INFO mapped.JobClient: Launched reduce tasks=1
13/12/13 08:51:18 INFO mapped.JobClient: SLOTS_MILLIS_MAPS=19126
13/12/13 08:51:18 INFO mapped.JobClient: Total time spent by all reduces waiting after receiving slots (ms)=0
13/12/13 08:51:18 INFO mapped.JobClient: Total time spent by all maps waiting after reserving slots (ms)=0
13/12/13 08:51:18 INFO mapped.JobClient: Launched map tasks=9
13/12/13 08:51:18 INFO mapped.JobClient: Data-local map tasks=9
13/12/13 08:51:18 INFO mapped.JobClient: SLOTS_MILLIS_REDUCES=21829
13/12/13 08:51:18 INFO mapped.JobClient: File Input Format Counters
13/12/13 08:51:18 INFO mapped.JobClient: Bytes Read=586276
13/12/13 08:51:18 INFO mapped.JobClient: File Output Format Counters
13/12/13 08:51:18 INFO mapped.JobClient: Bytes Written=177368
13/12/13 08:51:18 INFO mapped.JobClient: FileSystemCounters
13/12/13 08:51:18 INFO mapped.JobClient: FILE_BYTES_READ=280088
13/12/13 08:51:18 INFO mapped.JobClient: HDFS_BYTES_READ=58728
13/12/13 08:51:18 INFO mapped.JobClient: FILE_BYTES_WRITTEN=812954
13/12/13 08:51:18 INFO mapped.JobClient: HDFS_BYTES_WRITTEN=177368
13/12/13 08:51:18 INFO mapped.JobClient: Map-Reduce Framework
13/12/13 08:51:18 INFO mapped.JobClient: Map output materialized bytes=280088
13/12/13 08:51:18 INFO mapped.JobClient: Map input records=1258
13/12/13 08:51:18 INFO mapped.JobClient: Reduce shuffle bytes=280088
13/12/13 08:51:18 INFO mapped.JobClient: Spilled Records=14342
13/12/13 08:51:18 INFO mapped.JobClient: Map output bytes=711549
13/12/13 08:51:18 INFO mapped.JobClient: Total committed heap usage (bytes)=2007989128
13/12/13 08:51:18 INFO mapped.JobClient: CPU time spent (ms)=5389
13/12/13 08:51:18 INFO mapped.JobClient: Map input bytes=586019
13/12/13 08:51:18 INFO mapped.JobClient: SPLIT_RAW_BYTES=950
13/12/13 08:51:18 INFO mapped.JobClient: Combine input records=63553
13/12/13 08:51:18 INFO mapped.JobClient: Reduce input records=17171
13/12/13 08:51:18 INFO mapped.JobClient: Reduce input groups=13630
13/12/13 08:51:18 INFO mapped.JobClient: Combine output records=17171
13/12/13 08:51:18 INFO mapped.JobClient: Physical memory (bytes) snapshot=1032558016
13/12/13 08:51:18 INFO mapped.JobClient: Reduce output records=13630
13/12/13 08:51:18 INFO mapped.JobClient: Virtual memory (bytes) snapshot=2675306496
13/12/13 08:51:18 INFO mapped.JobClient: Map output records=63553
c:\Hadoop\hadoop-1.1.0-SNAPSHOT>
```

Si el trabajo se completa satisfactoriamente, debería obtener un

resultado similar al de la captura de pantalla siguiente:

En la captura de pantalla, puede ver tanto la asignación como la reducción completadas al 100%. También muestra el identificador del trabajo, job_201312092021_0002. El mismo informe se puede recuperar abriendo el acceso directo **Hadoop MapReduce status** en el escritorio y buscando el identificador del trabajo.

La otra opción para ejecutar un trabajo de MapReduce es utilizar Azure PowerShell.

Para mostrar el resultado de HDFS

- Abra la línea de comandos de Hadoop.
- Ejecute los comandos siguientes para ver el resultado:

```
hadoop fs -ls /WordCount/Output/
```

```
hadoop fs -cat /WordCount/Output/part-00000
```

Puede anexar “|more” al final del comando para obtener la vista de la página. O bien, utilice el comando findstr para encontrar un patrón de cadena:

```
hadoop fs -cat /WordCount/Output/part-00000 | findstr
```

```
“there”
```

Hasta ahora, ha desarrollado un trabajo de MapReduce para el recuento de palabras y lo ha probado correctamente en el emulador. El paso siguiente es implementarlo y ejecutarlo en HDInsight de Azure.

Carga de archivos de datos al almacenamiento de blobs de Azure

HDInsight de Azure utiliza el almacenamiento de blobs de Azure para almacenar datos. Cuando se aprovisiona un cluster de HDInsight, se utiliza un contenedor de almacenamiento de blobs de Azure para almacenar archivos del sistema. Puede utilizar este contenedor predeterminado u otro diferente (ya sea en la misma cuenta de almacenamiento de Azure o en otra cuenta de almacenamiento ubicada en el mismo centro de datos que el cluster) para almacenar los archivos de datos.

En este tutorial, creará un contenedor en una cuenta de

almacenamiento independiente para los archivos de datos y la aplicación de MapReduce. Los archivos de datos son los archivos de texto del directorio %hadoop_home% en su estación de trabajo.

Para crear un almacenamiento de blobs y un contenedor

- Abra Azure PowerShell.
- Configure las variables y, a continuación, ejecute los siguientes comandos:

```
$subscriptionName = "<AzureSubscriptionName>"  
$storageAccountName_Data = "<AzureStorageAccountName>"  
$containerName_Data = "<ContainerName>"  
$location = "<MicrosoftDataCenter>" # Por ejemplo,  
"East US"
```

El valor **\$subscriptionName** está asociado a la suscripción de Azure. Debe asignar un nombre a los valores **\$storageAccountName_Data** y **\$containerName_Data**.

- Ejecute el comando siguiente para crear una cuenta de almacenamiento y un contenedor de almacenamiento de blobs en la cuenta:

```
# Seleccionar suscripción de Azure  
Select-AzureSubscription $subscriptionName  
  
# Crear una cuenta de almacenamiento  
New-AzureStorageAccount -StorageAccountName  
$storageAccountName_Data -location $location  
  
# Cree un contenedor de almacenamiento de blobs.  
$storageAccountKey = Get-AzureStorageKey  
$storageAccountName_Data | %{ $_.Primary }  
$destContext = New-AzureStorageContext -  
StorageAccountName $storageAccountName_Data -  
StorageAccountKey $storageAccountKey  
New-AzureStorageContainer -Name $containerName_Data -
```

Context \$destContext

- Ejecute los comandos siguientes para comprobar la cuenta de almacenamiento y el contenedor:

```
Get-AzureStorageAccount -StorageAccountName
```

```
$storageAccountName_Data
```

```
Get-AzureStorageContainer -Context $destContext
```

Para cargar los archivos de datos

- Abra Azure PowerShell.
- Configure las tres primeras variables y, a continuación, ejecute los siguientes comandos:

```
$subscriptionName = "<AzureSubscriptionName>"
```

```
$storageAccountName_Data = "<AzureStorageAccountName>"
```

```
$containerName_Data = "<ContainerName>"
```

```
$localFolder = "c:\Hadoop\hadoop-1.1.0-SNAPSHOT"
```

```
$destFolder = "WordCount/Input"
```

Los valores de `$storageAccountName_Data` y `$containerName_Data` son iguales a los definidos en el último procedimiento.

Tenga en cuenta que la carpeta de los archivos de origen es `c:\Hadoop\hadoop-1.1.0-SNAPSHOT` y la carpeta de destino es `WordCount/Input`.

- Ejecute los comandos siguientes para obtener una lista de los archivos txt en la carpeta de archivos de origen:

```
# Obtener una lista de los archivos txt
```

```
$filesAll = Get-ChildItem $localFolder
```

```
$filesTxt = $filesAll | where {$_.Extension -eq ".txt"}
```

- Ejecute los comandos siguientes para crear un objeto de contexto de almacenamiento:

```
# Crear un objeto de contexto de almacenamiento
```

```
Select-AzureSubscription $subscriptionName
```

```
$storageaccountkey = get-azurestoragekey
```

```
$storageAccountName_Data | %{$_.Primary}
```



```
$destContext = New-AzureStorageContext -  
StorageAccountName $storageAccountName_Data -  
StorageAccountKey $storageaccountkey
```

- Ejecute los comandos siguientes para copiar los archivos:

```
# Copiar el archivo desde la estación de trabajo local  
al contenedor de blobs  
foreach ($file in $filesTxt){
```

```
    $fileName = "$localFolder\$file"
```

```
    $blobName = "$destFolder/$file"
```

```
    write-host "Copying $fileName to $blobName"
```

```
    Set-AzureStorageBlobContent -File $fileName -  
Container $containerName_Data -Blob $blobName -Context  
$destContext  
}
```

- Ejecute el comando siguiente para incluir los archivos cargados:

```
# Enumerar los archivos cargados en el contenedor de  
almacenamiento de blobs  
Write-Host "The Uploaded data files:" -BackgroundColor  
Green  
Get-AzureStorageBlob -Container $containerName_Data -  
Context $destContext -Prefix $destFolder
```

Debería ver unos ocho archivos de datos de texto.

Para cargar la aplicación para el recuento de palabras

- Abra Azure PowerShell.
- Configure las tres primeras variables y, a continuación, ejecute los siguientes comandos:

```
$subscriptionName = "<AzureSubscriptionName>"
```

```
$storageAccountName_Data = "<AzureStorageAccountName>"
```

```
$containerName_Data = "<ContainerName>"
```

```
$jarFile = "C:\Tutorials\WordCountJava\WordCount.jar"
```

```
$blobFolder = "WordCount/jars"
```

Los valores de **\$storageAccountName_Data** y **\$containerName_Data** son los mismos que ha definido en el último procedimiento, lo que significa que cargará tanto el archivo de datos como la aplicación en el mismo contenedor y en la misma cuenta de almacenamiento.

Tenga en cuenta que la carpeta de destino es **WordCount/jars**.

- Ejecute los comandos siguientes para crear un objeto de contexto de almacenamiento:

```
# Crear un objeto de contexto de almacenamiento
```

```
Select-AzureSubscription $subscriptionName
```

```
$storageaccountkey = get-azurestoragekey
```

```
$storageAccountName_Data | %{$_.Primary}
```

```
$destContext = New-AzureStorageContext -
```

```
StorageAccountName $storageAccountName_Data -
```

```
StorageAccountKey $storageaccountkey
```

- Ejecute los comandos siguientes para copiar las aplicaciones:

```
Set-AzureStorageBlobContent -File $jarFile -Container
```

```
$containerName_Data -Blob "$blobFolder/WordCount.jar" -
```

```
Context $destContext
```

- Ejecute el comando siguiente para incluir los archivos cargados:

```
# Enumerar los archivos cargados en el contenedor de
```

```
almacenamiento de blobs
```

```
Write-Host "The Uploaded application file:" -
```

```
BackgroundColor Green
```

```
Get-AzureStorageBlob -Container $containerName_Data -
```

```
Context $destContext -Prefix $blobFolder
```

Debería ver aquí el archivo jar.

Ejecución del programa de MapReduce en HDInsight

de Azure

El siguiente script de PowerShell ejecuta las tareas siguientes:

- Aprovisionar un cluster de hdinsight
- Crear una cuenta de almacenamiento que se utilizará como el sistema de archivos predeterminado del cluster de hdinsight
- Crear un contenedor de almacenamiento de blobs
- Crear un cluster de hdinsight
- Enviar el trabajo de mapreduce
- Crear una definición de trabajo de mapreduce
- Enviar un trabajo de mapreduce
- Esperar hasta que se haya completado el trabajo
- Mostrar el error estándar
- Mostrar el resultado estándar
- Eliminar el cluster
- Eliminar el cluster de hdinsight
- Eliminar la cuenta de almacenamiento utilizada como sistema de archivos predeterminado del cluster de hdinsight

Para ejecutar el script de PowerShell

- Abra el Bloc de notas.
- Copie y pegue el código siguiente:

```
# La cuenta de almacenamiento y las variables del
cluster de HDInsight
$subscriptionName = "<WindowsAzureSubscriptionName>"
$serviceNameToken = "<ServiceNameTokenString>"
$location = "<MicrosoftDataCenter>"      ### coincidir
con la ubicación de la cuenta de almacenamiento de datos
$clusterNodes = <NumberOfNodesInTheCluster>

$storageAccountName_Data =
```

```

"<TheDataStorageAccountName>"
$containerName_Data =
"<TheDataBlobStorageContainerName>"

$clusterName = $serviceNameToken + "hdicluster"

$storageAccountName_Default = $serviceNameToken +
"hdistore"
$containerName_Default = $serviceNameToken +
"hdicluster"

# Las variables de trabajo de MapReduce
$jarFile =
"wasb://$containerName_Data@$storageAccountName_Data.blob.core.windows.net/WordCount/jars/WordCount.jar"
$className = "org.apache.hadoop.examples.WordCount"
$mrInput =
"wasb://$containerName_Data@$storageAccountName_Data.blob.core.windows.net/WordCount/Input/"
$mrOutput =
"wasb://$containerName_Data@$storageAccountName_Data.blob.core.windows.net/WordCount/Output/"
$mrStatusOutput =
"wasb://$containerName_Data@$storageAccountName_Data.blob.core.windows.net/WordCount/MRStatusOutput/"

# Crear un objeto PSCredential. El nombre de usuario y
la contraseña se codifican aquí de forma rígida. Puede
cambiarlos si lo desea.
$password = ConvertTo-SecureString "Pass@word1" -
AsPlainText -Force
$creds = New-Object

```

```
System.Management.Automation.PSCredential ("Admin",
$password)
```

```
Select-AzureSubscription $subscriptionName
```

```
#=====
```

```
# Crear una cuenta de almacenamiento utilizada como
sistema de archivos predeterminado
```

```
Write-Host "Create a storage account" -ForegroundColor
Green
```

```
New-AzureStorageAccount -StorageAccountName
$storageAccountName_Default -location $location
```

```
#=====
```

```
# Crear un contenedor de almacenamiento de blobs
utilizado como sistema de archivos predeterminado
```

```
Write-Host "Create a Blob storage container" -
ForegroundColor Green
```

```
$storageAccountKey_Default = Get-AzureStorageKey
$storageAccountName_Default | %{ $_.Primary }
```

```
$destContext = New-AzureStorageContext -
StorageAccountName $storageAccountName_Default -
StorageAccountKey $storageAccountKey_Default
```

```
New-AzureStorageContainer -Name $containerName_Default
-Context $destContext
```

```
#=====
```

```
# Crear un cluster de HDInsight
```

```
Write-Host "Create an HDInsight cluster" -
ForegroundColor Green
```

```
$storageAccountKey_Data = Get-AzureStorageKey
$storageAccountName_Data | %{ $_.Primary }
```

```
$config = New-AzureHDInsightClusterConfig -
```

```
ClusterSizeInNodes $clusterNodes |
    Set-AzureHDInsightDefaultStorage -
StorageAccountName
"$storageAccountName_Default.blob.core.windows.net" -
StorageAccountKey $storageAccountKey_Default -
StorageContainerName $containerName_Default |
    Add-AzureHDInsightStorage -StorageAccountName
"$storageAccountName_Data.blob.core.windows.net" -
StorageAccountKey $storageAccountKey_Data

New-AzureHDInsightCluster -Name $clusterName -Location
$location -Credential $creds -Config $config
```

```
#=====
# Crear una definición de trabajo de MapReduce
Write-Host "Create a MapReduce job definition" -
ForegroundColor Green
$mrJobDef = New-AzureHDInsightMapReduceJobDefinition -
JobName mrWordCountJob -JarFile $jarFile -ClassName
$className -Arguments $mrInput, $mrOutput -StatusFolder
/WordCountStatus
```

```
#=====
# Ejecutar el trabajo de MapReduce
Write-Host "Run the MapReduce job" -ForegroundColor
Green
$mrJob = Start-AzureHDInsightJob -Cluster $clusterName
-JobDefinition $mrJobDef
Wait-AzureHDInsightJob -Job $mrJob -
WaitTimeoutInSeconds 3600
Get-AzureHDInsightJobOutput -Cluster $clusterName -
JobId $mrJob.JobId -StandardError
```

```
Get-AzureHDInsightJobOutput -Cluster $clusterName -  
JobId $mrJob.JobId -StandardOutput
```

```
#=====
```

```
# Eliminar el cluster de HDInsight
```

```
Write-Host "Delete the HDInsight cluster" -
```

```
ForegroundColor Green
```

```
Remove-AzureHDInsightCluster -Name $clusterName
```

```
# Eliminar la cuenta de almacenamiento del sistema de  
archivos predeterminado
```

```
Write-Host "Delete the storage account" -
```

```
ForegroundColor Green
```

```
Remove-AzureStorageAccount -StorageAccountName
```

```
$storageAccountName_Default
```

- Establezca las seis primeras variables en el script. **\$serviceNameToken** se utilizará para el nombre del cluster de HDInsight, el nombre de la cuenta de almacenamiento predeterminado y el nombre del contenedor de almacenamiento de blobs predeterminado. Como el nombre del servicio debe tener de 3 a 24 caracteres y el script anexa una cadena de hasta 10 caracteres a los nombres, debe limitar la cadena a 14 caracteres o menos. Y **\$serviceNameToken** debe utilizar minúsculas. **\$storageAccountName_Data** y **\$containerName_Data** constituyen la cuenta de almacenamiento y el contenedor que se utilizan para almacenar los archivos de datos y la aplicación. **\$location** debe coincidir con la ubicación de la cuenta de almacenamiento de datos.

- Revise el resto de las variables.
- Guarde el archivo de script.
- Abra Azure PowerShell.
- Ejecute el comando siguiente para establecer la directiva de ejecución en remotesigned:

```
PowerShell -File <FileName> -ExecutionPolicy
```

RemoteSigned

- Cuando se le solicite, escriba el nombre de usuario y la contraseña para el cluster de HDInsight. Como eliminará el cluster al final del script y no necesitará más el nombre de usuario y la contraseña, estos pueden ser cualquier cadena.

Recuperación del resultado del trabajo de MapReduce

En esta sección se muestra cómo descargar y mostrar los resultados. Para obtener más información sobre cómo mostrar los resultados en Excel, consulte [Conexión de Excel a HDInsight con Microsoft Hive ODBC Driver](#) y [Conexión de Excel a HDInsight con Power Query](#).

Para recuperar los resultados

- Abra la ventana de Azure PowerShell.
- Cambie el directorio a C:\Tutorials\WordCountJava. La carpeta predeterminada de Azure PowerShell es C:\Windows\System32\WindowsPowerShell\v1.0. Los cmdlets que ejecutará descargarán el archivo de resultados a la carpeta actual. No tiene permiso para descargar los archivos en las carpetas del sistema.
- Ejecute los siguientes comandos para establecer los valores:

```
$subscriptionName = "<WindowsAzureSubscriptionName>"
```

```
$storageAccountName_Data =
```

```
"<TheDataStorageAccountName>"
```

```
$containerName_Data =
```

```
"<TheDataBlobStorageContainerName>"
```

```
$blobName = "WordCount/Output/part-r-00000"
```

- Ejecute los siguientes comandos para crear un objeto de contexto de almacenamiento de Azure:

```
Select-AzureSubscription $subscriptionName
```

```
$storageAccountKey = Get-AzureStorageKey
```

```
$storageAccountName_Data | %{ $_.Primary }
```

```
$storageContext = New-AzureStorageContext -
```



```
StorageAccountName $storageAccountName_Data –
```

```
StorageAccountKey $storageAccountKey
```

- Ejecute los comandos siguientes para descargar y mostrar el resultado:

```
Get-AzureStorageBlobContent -Container
```

```
$containerName_Data -Blob $blobName -Context
```

```
$storageContext -Force
```

```
cat “./$blobName” | findstr “there”
```

CONEXIÓN DE EXCEL A HDINSIGHT CON POWER QUERY

Una de las características clave de la solución para datos de gran tamaño de Microsoft es la integración de los componentes de Microsoft Business Intelligence (BI) con los clusters de HDInsight Hadoop. Un ejemplo importante de esta integración es la capacidad de conectar Excel a la cuenta de almacenamiento de Azure que contiene los datos asociados a su cluster de HDInsight usando Microsoft Power Query para Excel. En este artículo se ofrecen las pautas para configurar y usar Power Query desde Excel para consultar los datos asociados a un cluster de HDInsight.

Requisitos previos

Antes de empezar este apartado, debe disponer de lo siguiente:

Un cluster de HDInsight.

Un equipo que ejecute Windows 8, Windows 7, Windows Server 2012 o Windows Server 2008 R2.

Office Professional Plus 2013, Office 365 Pro Plus, Excel 2013 Standalone u Office Professional Plus 2010.

Instalación de Microsoft Power Query para Excel

Se puede usar Power Query para importar datos desde diferentes orígenes en Microsoft Excel, donde se pueden habilitar herramientas de inteligencia empresarial (BI) como PowerPivot y Power View. En

concreto, Power Query puede importar datos ofrecidos o generados por un trabajo de Hadoop ejecutado en un cluster de HDInsight.

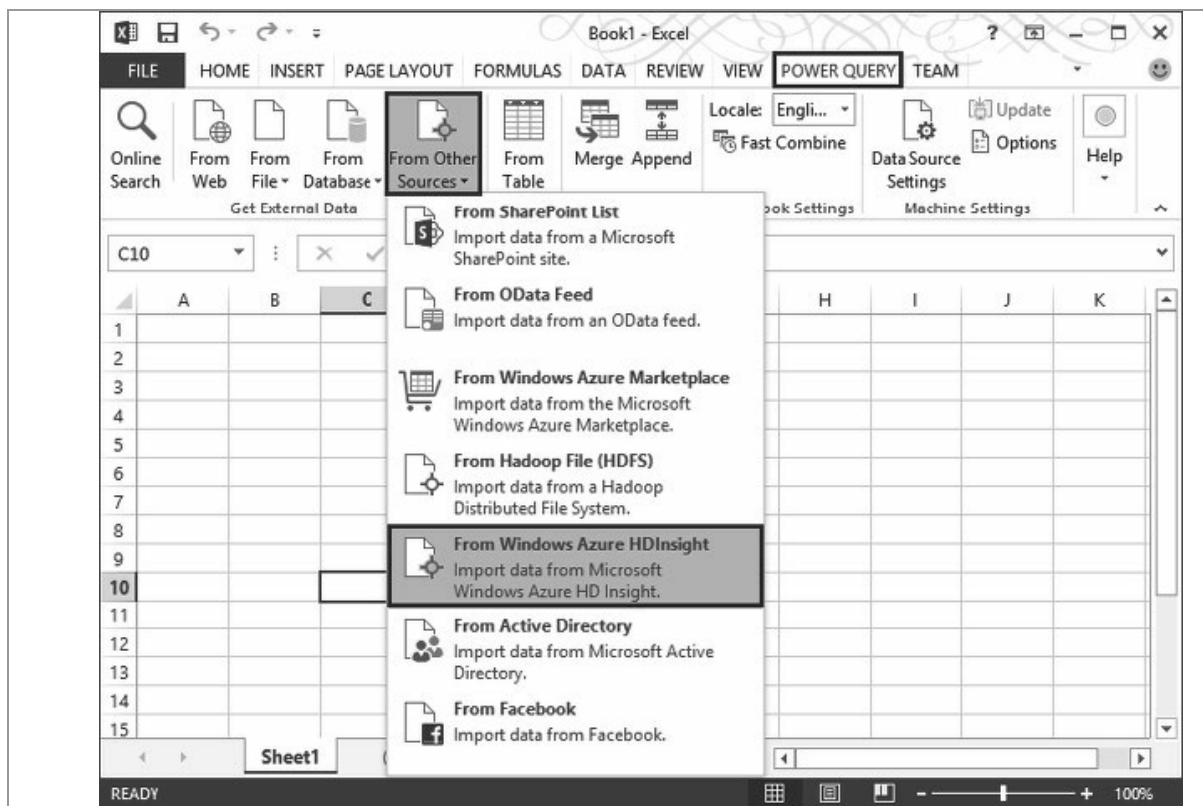
Descargue Microsoft Power Query para Excel en el Centro de descarga de Microsoft e instálelo.

Importación de datos de HDInsight a Excel

El complemento de Power Query para Excel facilita la importación de datos desde su cluster de HDInsight a Excel, donde se pueden usar herramientas de inteligencia empresarial como PowerPivot y Power Map para la inspección, el análisis y la presentación de los datos.

Importación de datos desde un cluster de HDInsight

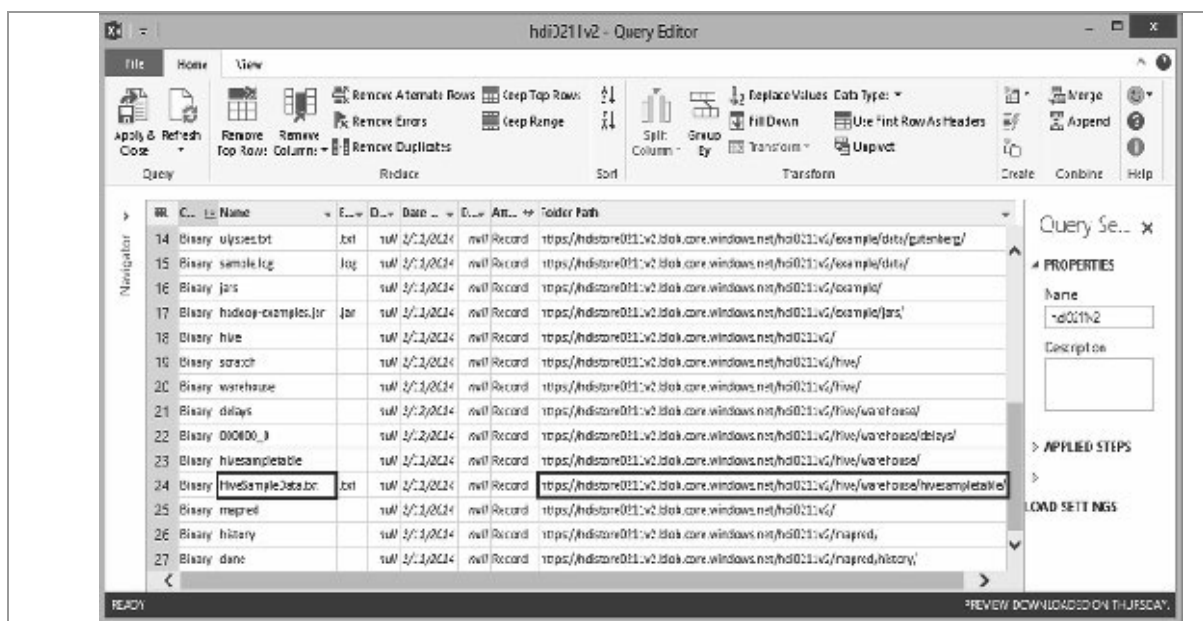
- Abra Excel.
- Cree un libro vacío.
- Haga clic en el menú **Power Query**, en **From Other Sources** y, a continuación, en **From Azure HDInsight**.



Nota: si no ve el menú **Power Query**, vaya a **File > Options > Add-Ins** y seleccione **COM Addins** en el cuadro desplegable **Manager** situado

al final de la página. Seleccione el botón **Go...** y compruebe que la casilla del complemento de Microsoft Office Power Query para Excel esté activada.

- En **Account Name**, escriba el nombre de la cuenta de almacenamiento de blobs de Azure asociada a su cluster y, a continuación, haga clic en **OK**.
- En **Account Key**, escriba la clave de cuenta para la cuenta de almacenamiento de blobs y, a continuación, haga clic en **Save**. (Solo tendrá que hacer esto la primera vez que obtenga acceso a este almacén).
- En el panel **Navigator** situado a la Izquierda de **Query Editor**, haga doble clic en el nombre del contenedor de almacenamiento de blobs. De forma predeterminada, el nombre del contenedor es el mismo que el del cluster.
- Busque **HiveSampleData.txt** en la columna **Name** (la ruta de la carpeta es **../hive/warehouse/hivesampletable/**) y, a continuación, haga clic en **Binary** a la izquierda de HiveSampleData.txt.



- Si quiere, puede cambiar el nombre de las columnas. Cuando haya terminado, haga clic en **Apply & Close**.

Book2 - Excel

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW POWER QUERY TEAM QUERY DESIGN Jonathan...

Edit Refresh Duplicate Reference Delete Merge Append Share

Data Manage Combine Share

A1 Column1

	A	B	C	D	E	F	G	H
1	Column1	Column2	Column3	Column4	Column5	Column6	Column7	Column8
2	8	6:54:20 PM	en-US	Android	Samsung	SCH-i500	California	United States
3	23	7:19:44 PM	en-US	Android	HTC	Incredible	Pennsylvania	United States
4	23	7:19:46 PM	en-US	Android	HTC	Incredible	Pennsylvania	United States
5	23	7:19:47 PM	en-US	Android	HTC	Incredible	Pennsylvania	United States
6	28	1:37:50 AM	en-US	Android	Motorola	Droid X	Colorado	United States
7	28	12:53:31 AM	en-US	Android	Motorola	Droid X	Colorado	United States
8	28	12:53:50 AM	en-US	Android	Motorola	Droid X	Colorado	United States
9	28	4:44:21 PM	en-US	Android	Motorola	Droid X	Utah	United States
10	28	4:43:41 PM	en-US	Android	Motorola	Droid X	Utah	United States
11	28	1:37:19 AM	en-US	Android	Motorola	Droid X	Colorado	United States
12	30	5:19:36 PM	en-US	RIM OS	RIM	9650	Massachusetts	United States
13	30	5:17:18 PM	en-US	RIM OS	RIM	9650	Massachusetts	United States
14	30	5:16:40 PM	en-US	RIM OS	RIM	9650	Massachusetts	United States
15	30	5:16:40 PM	en-US	RIM OS	RIM	9650	Massachusetts	United States

Sheet1 Sheet2

READY 100%

CONEXIÓN DE EXCEL A HDINSIGHT CON MICROSOFT HIVE ODBC DRIVER

Una característica clave de la solución de datos de gran tamaño de Microsoft es la integración de componentes de inteligencia empresarial de Microsoft con clusters Apache Hadoop implementados por HDInsight de Azure. Un ejemplo de esta integración es la capacidad de conectar Excel al almacén de datos de Hive de un cluster Hadoop de HDInsight mediante la utilización de Microsoft Hive Open Database Connectivity (ODBC) Driver.

También es posible conectar desde Excel los datos asociados con un cluster de HDInsight y otros orígenes de datos, incluidos otros clusters Hadoop (que no sean de HDInsight), con la utilización del complemento Microsoft Power Query para Excel.

Requisitos previos

Antes de empezar este artículo, debe tener lo siguiente:

Un cluster de HDInsight. Para configurarlo, consulte [Introducción a HDInsight de Azure](#).

Un equipo que ejecute Windows 8, Windows 7, Windows Server 2012 o Windows Server 2008 R2.

Office Professional Plus 2013, Office 365 Pro Plus, Excel 2013 Standalone u Office Professional Plus 2010.

Instalación de Microsoft Hive ODBC Driver

Descargue e instale Microsoft Hive ODBC Driver desde el Centro de descarga.

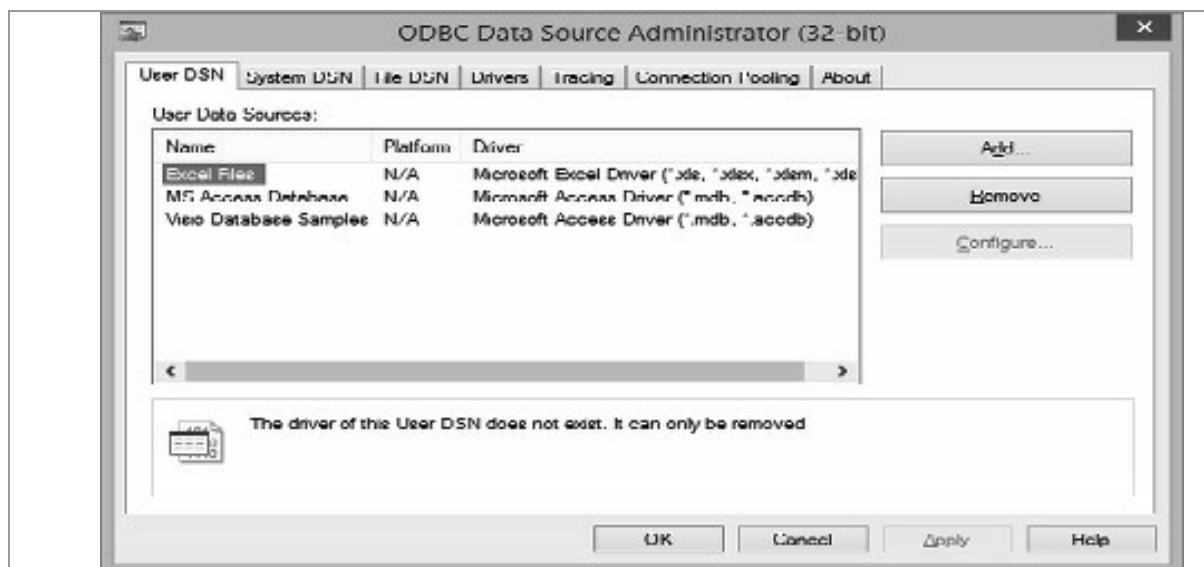
Este controlador puede instalarse en versiones de 32 o 64 bits de Windows 7, Windows 8, Windows Server 2008 R2 y Windows Server 2012 y permitirá la conexión con HDInsight de Azure (versión 1.6 y

posterior) y emulador de HDInsight de Azure (v.1.0.0.0 y posterior). Debe instalar la versión que coincida con la versión de la aplicación en que va a usar el controlador ODBC. Para este tutorial, el controlador se usará desde Office Excel.

Creación de un origen de datos de Hive ODBC

En los siguientes pasos se explica cómo crear un origen de datos de Hive ODBC.

- En Windows 8, presione la tecla Windows para abrir la pantalla Inicio y, a continuación, escriba **orígenes de datos**.
- Haga clic en **Configurar orígenes de datos ODBC (32 bits)** o **Configurar orígenes de datos ODBC (64 bits)**, en función de la versión de Office de que se trate. Si usa Windows 7, seleccione **Orígenes de datos ODBC (32 bits)** u **Orígenes de datos ODBC (64 bits)** en **Herramientas administrativas**. A continuación, se abrirá el cuadro de diálogo **Administrador de orígenes de datos ODBC**.



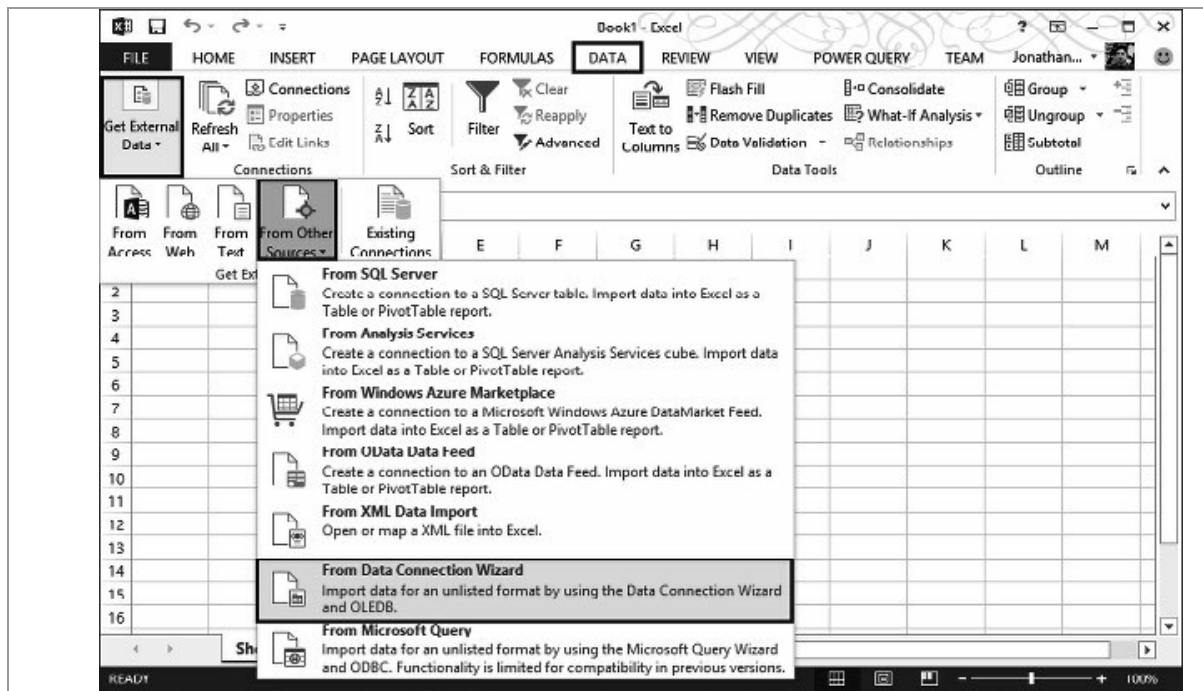
- En la pestaña DNS de usuario, haga clic en **Agregar** para abrir el asistente **Crear nuevo origen de datos**.
- Seleccione **Microsoft Hive ODBC Driver** y, a continuación, haga clic en Finalizar. Se abrirá el cuadro de diálogo **Microsoft Hive ODBC Driver DNS Setup**.
- Escriba o seleccione los siguientes valores:

- Hay algunos parámetros importantes que se deben tener en cuenta al hacer clic en ****Opciones avanzadas****:
- Haga clic en **Probar** para probar el origen de datos. Cuando el origen de datos esté configurado correctamente, aparecerá el mensaje *PRUEBAS COMPLETADAS CORRECTAMENTE*.
- Haga clic en **Aceptar** para cerrar el cuadro de diálogo de prueba. Ahora el nuevo origen de datos debería aparecer en el **Administrador de orígenes de datos ODBC**.
- Haga clic en **Aceptar** para salir del asistente.

Importación de datos a Excel desde un cluster de HDInsight

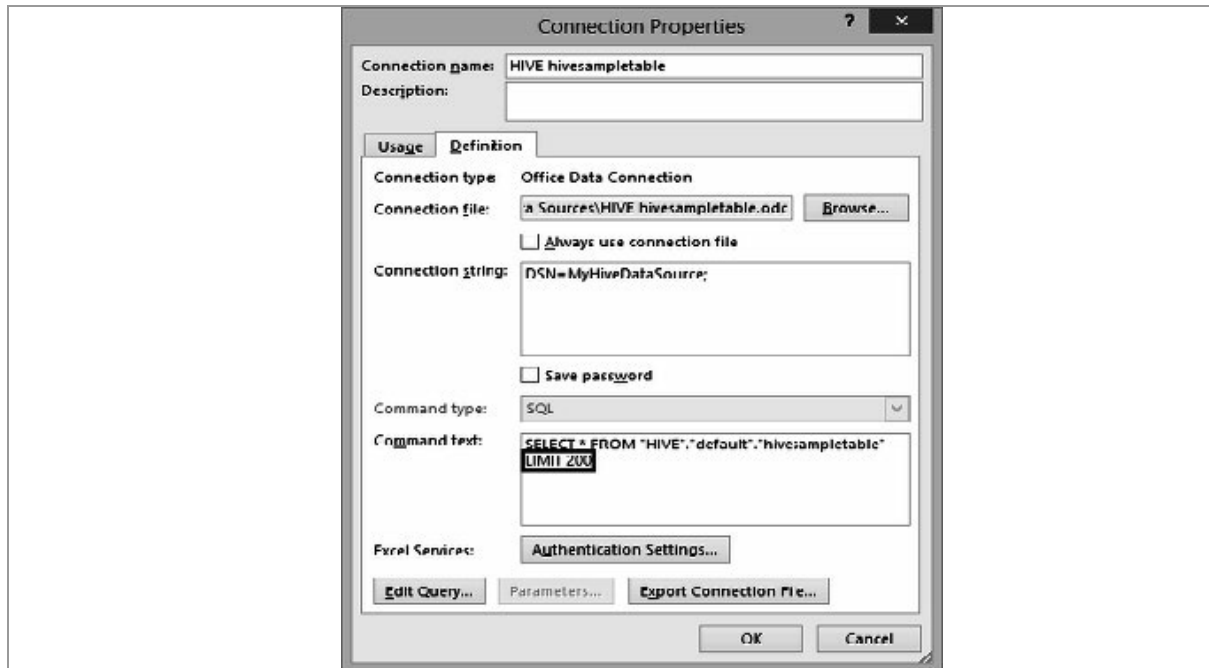
En los pasos siguientes se describe cómo importar datos desde una tabla de Hive a un libro de Excel mediante el origen de datos ODBC creado en los pasos anteriores.

- Abra un libro de Excel nuevo o existente.
- En la pestaña **Datos**, haga clic en **Obtener datos externos**, haga clic en **From Other Data Sources** y, a continuación, haga clic en **Desde el Asistente para la conexión de datos** para abrir el **Asistente para la conexión de datos**.



- Seleccione **DSN ODBC** como el origen de datos y, a continuación, haga clic en **Siguiente**.
- En los orígenes de datos ODBC, seleccione el nombre del origen de datos creado en el paso anterior y, a continuación, haga clic en **Siguiente**.
- Vuelva a escribir la contraseña para el cluster en el asistente y, a continuación, haga clic en **Probar** para comprobar la configuración.
- Haga clic en **Aceptar** para cerrar el cuadro de diálogo de prueba.
- Haga clic en **OK**. Espere a que se abra el cuadro de diálogo **Seleccionar base de datos y tabla**. Esta operación puede tardar unos segundos.
- Seleccione la tabla que desea importar y, a continuación, haga clic en **Siguiente**, *hivesampletable* es una tabla de Hive de muestra integrada en los clusters de HDInsight. Puede seleccionarla si no ha creado ninguna. Para obtener más información acerca de cómo ejecutar consultas de Hive y crear tablas de Hive, consulte *Uso de Hive con HDInsight*.
- Haga clic en **Finish**.
- En el cuadro de diálogo **Importar datos**, puede cambiar o especificar la consulta. Para ello, haga clic en **Propiedades**. Esta operación puede tardar unos segundos.

- Haga clic en la pestaña **Definición** y, a continuación, anexe **LIMIT 200** a la instrucción select de Hive en el cuadro de texto **Texto de comando**. La modificación limitará el conjunto de registros devueltos a 200.



- Haga clic en **Aceptar** para cerrar el cuadro de diálogo **Propiedades de conexión**.
- Haga clic en **Aceptar** para cerrar el cuadro de diálogo **Importar datos**.
- Vuelva a escribir la contraseña y, a continuación, haga clic en **Aceptar**. La importación de los datos a Excel tarda algunos segundos.

CAPÍTULO 7

BUSINESS INTELLIGENCE Y BIG DATA CON MICROSOFT SQL SERVER

SQL SERVER 2014 Y EL BIG DATA

SQL Server ofrece un rendimiento fiable gracias a la integración de tecnologías en memoria (IN-MEMORY OLTP), una rápida obtención de información útil a partir de cualquier tipo y volumen de datos de datos (BIG DATA), con herramientas adecuadas y una plataforma para compilar, implementar y administrar soluciones tanto locales como en nube.

SQL Server 2014 integra funcionalidad que, además de aportar fiabilidad, permite revelar información útil mediante el uso de herramientas de análisis conocidas y soluciones Big Data preparadas para las empresas. Su arquitectura y herramientas comunes para entornos locales y en nube hacen posibles las infraestructuras de TI híbridas. SQL Server 2014 incluye nueva funcionalidad en memoria en la base de datos principal y también proporciona nuevas funciones preparadas para la nube y el Big Data que simplifican su adopción.

SQL Server 2014 dispone de una solución de Business Intelligence completa que agiliza las operaciones de búsqueda, acceso y formato de datos internos y externos, así como la combinación de datos estructurados y no estructurados. De esta forma se habilita el trabajo con Big Data.

SQL Server 2014 proporciona nuevas soluciones de recuperación ante desastres y copia de seguridad con Microsoft Azure, así como nuevas herramientas para trasladar a la nube de forma sencilla las bases de datos SQL Server locales, lo que permite a los usuarios usar sus conocimientos actuales para aprovechar las ventajas de los centros de datos globales de

Microsoft.

El almacenamiento de datos escalable es otra de las características esenciales de SQL Server 2014. Se trata de escalabilidad para todos los datos mediante el almacenamiento de datos relacional. También permite la integración con orígenes de datos no relacionales como Hadoop, lo que capacita el uso de Big Data favorecido por el procesamiento paralelo masivo y las tecnologías en memoria.

Big Data y la nube plantean nuevos retos y oportunidades a empresas de todo el mundo. Con SQL Server 2014 se pueden aprovechar todos los datos, grandes y pequeños, para controlar su negocio en tiempo real.

Revele nuevos conocimientos y permita una toma de decisiones más informada con Azure HDInsight, una solución Big Data que utiliza la tecnología Apache Hadoop y que se integra en SQL Server 2014.

Cuando se necesite, es posible compilar rápidamente, en minutos, un cluster Hadoop y elimínelo una vez que haya terminado su trabajo. También es posible elegir el tamaño de cluster adecuado para optimizar el costo o el tiempo de acceso a los datos. Integre HDInsight de manera fluida en sus flujos de trabajo de análisis con Azure PowerShell y la interfaz de línea de comandos de Azure.

Basado completamente en Apache Hadoop, HDInsight permite ejecutar soluciones Big Data en Azure. Además, con Hortonworks Data Platform (HDP), se pueden ejecutar aplicaciones Big Data en Windows Server o Linux. Esta flexibilidad es única en el sector y permite aprovechar las ventajas de los proyectos existentes de Apache Hadoop, como Apache Pig, Hive, Sqoop y muchos más.

Para los desarrolladores, HDInsight ofrece la posibilidad de programación en su lenguaje preferido, como Java, .NET, etc. Los desarrolladores de .NET pueden

aprovechar la eficacia de la consulta integrada en el lenguaje con LINQto Hive. Los desarrolladores de bases de datos pueden aprovechar sus conocimientos de SQL para consultar y transformar a través de Hive.

Con las soluciones Big Data de Microsoft, puede implementar un cluster de Hadoop y estar preparado en unos minutos para consultar y combinar datos relacionales y no relacionales, con los mismos conocimientos que emplea con SQL Server. Además, cualquiera puede ser

un experto de BI y usar Excel y Power BI para crear eficaces y hermosas visualizaciones o contar historias con datos.

CARACTERÍSTICAS DE BIG DATA Y BUSINESS INTELLIGENCE EN MICROSOFT BI SQL SERVER

Las características de SQL Server que son parte de la plataforma Microsoft BI incluyen Analysis Services, Integration Services, Reporting Services y varias aplicaciones cliente que se usan para crear datos analíticos o para trabajar con ellos. Analysis Services y Reporting Services se pueden integrar con una granja de Microsoft SharePoint para habilitar las características de Business Intelligence (BI) en SharePoint. Entre las características se incluyen PowerPivot para SharePoint, Power View y Reporting Services. PowerPivot para SharePoint se usa para el acceso a datos PowerPivot en una granja de SharePoint. PowerPivot para SharePoint es el motor de datos para los libros creados en PowerPivot para Excel y accesibles desde una biblioteca de SharePoint. Una vez que guarde un libro de PowerPivot en SharePoint, puede utilizarlo como origen de datos para los informes de Power View.

Si desea usar, tanto PowerPivot para SharePoint como Reporting Services, ejecute dos veces el Asistente para la instalación de SQL Server. Reporting Services y PowerPivot son opciones diferentes de la página Rol de instalación del Asistente para la instalación de SQL Server. PowerPivot para SharePoint admite SharePoint 2010 y SharePoint 2013; sin embargo, se utilizan una arquitectura y un proceso de instalación distintos en función de la versión de SharePoint.

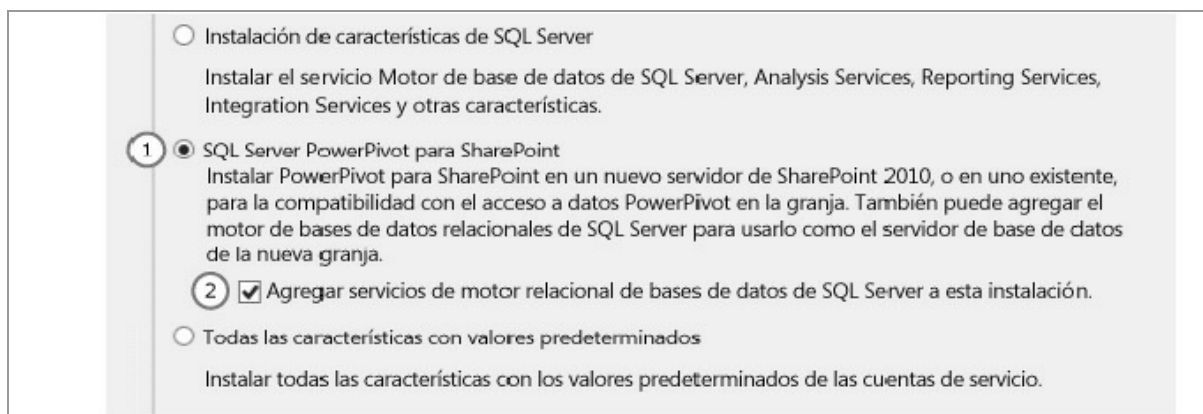
A continuación se muestra un resumen de los pasos de instalación para implementar características de BI de SQL Server 2014 en un único servidor.

PowerPivot para SharePoint 2013

En SharePoint 2013, la instalación de PowerPivot para SharePoint se puede ejecutar en un servidor que no tiene productos SharePoint instalados. La arquitectura de PowerPivot que se usa para SharePoint 2013 se ejecuta fuera de la granja de servidores de SharePoint y se puede

instalar en un servidor que también contiene una instalación de SharePoint o bien se puede instalar en un servidor que no contenga una instalación de SharePoint. Para instalar SQL Server PowerPivot:

1. Instale SharePoint Server 2013 y habilite Excel Services.
2. Instale Analysis Services en modo de SharePoint y conceda a las cuentas de servicios y de granja de SharePoint derechos de administrador del servidor en Analysis Services. En ambas versiones de SharePoint, el proceso de instalación de PowerPivot se inicia con la selección de la instalación del rol **SQL Server PowerPivot para SharePoint** en el Asistente para la instalación de SQL Server o utilizando una instalación desde el símbolo del sistema de SQL Server.



3. Para SharePoint 2013, puede ampliar las características y la experiencia de PowerPivot. Descargue y ejecute **spPowerPivot.msi** para agregar compatibilidad con procesamiento, colaboración y administración de actualización de datos del servidor para los libros PowerPivot. Ejecute el paquete de instalación de PowerPivot para SharePoint 2013 **spPowerPivot.msi** en cada servidor de la granja de SharePoint para asegurarse de que se instala la versión correcta de los proveedores de datos.

4. Para configurar PowerPivot para SharePoint 2013, use la herramienta **Configuración de PowerPivot para SharePoint 2013**. El asistente para la instalación de SQL Server instala dos herramientas de configuración de PowerPivot. Una de las herramientas de configuración admite SharePoint 2013 y la otra admite SharePoint 2010.

5. Configure Excel Services en SharePoint Server 2013 para usar la instancia de Analysis Services.

PowerPivot para SharePoint es un servidor de Analysis Services en modo de SharePoint que proporciona hospedaje de servidor de datos en una granja de SharePoint. Los datos de PowerPivot son un modelo de datos analíticos que se genera mediante el complemento PowerPivot para Excel 2010 o Excel 2013.

SQL Server 2014 PowerPivot para SharePoint admite el uso por parte de Excel Services de Microsoft SharePoint 2013 de libros de Excel que contienen modelos de datos e informes de Power View de Reporting Services.

Excel Services en SharePoint 2013 incluye funcionalidad de modelo de datos para habilitar la interacción con un libro PowerPivot en el explorador. No es necesario implementar el complemento PowerPivot para SharePoint 2013 en la granja de servidores. Solo se necesita instalar un servidor Analysis Services en modo de SharePoint y registrarlo dentro de la configuración de Modelo de datos de Excel Services.

La implementación del complemento PowerPivot para SharePoint 2013 habilita funcionalidad y características adicionales en la granja de servidores de SharePoint.

La Figura 7-1 muestra la implementación de servidores PowerPivot.

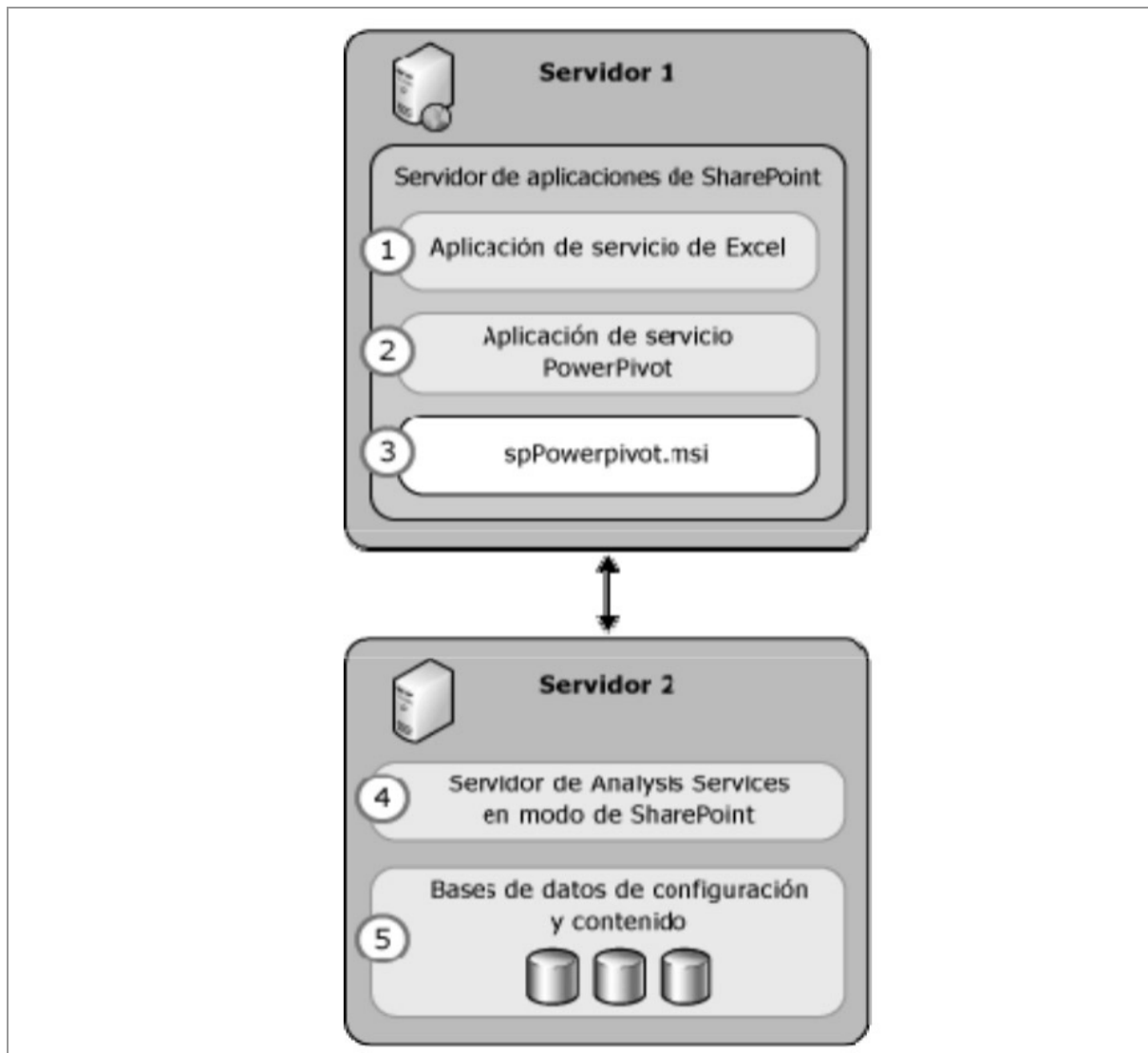


Figura 7-1

PowerPivot para SharePoint 2010

En SharePoint 2010, es necesario que la instalación de PowerPivot para SharePoint se ejecute en un servidor en el que se instalará SharePoint 2010 o que ya lo tenga instalado. La arquitectura de PowerPivot para SharePoint 2010 se ejecuta dentro de la granja y requiere SharePoint en el servidor en el que se ha instalado PowerPivot para SharePoint.

1. Instale Analysis Services en modo de SharePoint y conceda a las cuentas de servicios y de granja de SharePoint derechos de administrador del servidor en Analysis Services. Las implementaciones de SharePoint 2010 no admiten spPowerPivot.msi, y el archivo .msi no es necesario con SharePoint 2010. En ambas versiones de SharePoint,

el proceso de instalación de PowerPivot se inicia con la selección de la instalación del rol SQL Server PowerPivot para SharePoint en el Asistente para la instalación de SQL Server o utilizando una instalación desde el símbolo del sistema de SQL Server.

2. El asistente para la instalación de SQL Server instala dos herramientas de configuración de PowerPivot. Una de las herramientas de configuración admite SharePoint 2013 y la otra admite SharePoint 2010. Para configurar PowerPivot para SharePoint 2010, use Herramienta de configuración de PowerPivot.

PowerPivot para SharePoint 2010 proporciona hospedaje de servidor de datos PowerPivot en una granja de SharePoint 2010. Los datos de PowerPivot son un modelo de datos analíticos creado en Excel mediante el complemento PowerPivot para Excel.

El hospedaje de servidor de esos datos requiere SharePoint 2010, Excel Services y una instalación de PowerPivot para SharePoint. Los datos se cargan en las instancias de PowerPivot para SharePoint en la granja, donde pueden actualizarse a intervalos programados mediante la función de actualización de datos PowerPivot que el servidor proporciona.

PowerPivot para SharePoint se implementa como un servicio compartido, lo que significa que las características integradas y la infraestructura se pueden utilizar para administrar, proteger y usar una aplicación de servicio PowerPivot. La detección, redirección y administración de conexiones del servidor y la base de datos se administran en el nivel de granja.

Administración central proporciona la interfaz administrativa a los servicios utilizados para administrar la identidad del servidor, el estado del servidor y las propiedades de configuración.

Una implementación completa de cliente de PowerPivot para SharePoint incluye componentes de servidor que se integran con Excel y Excel Services en una granja de servidores de SharePoint. Los datos PowerPivot de un libro de Excel son una base de datos de Analysis Services que requiere un motor analítico en memoria xVelocity de Analysis Services (VertiPaq) para cargar y consultar los datos. En una estación de trabajo cliente, el motor xVelocity se ejecuta en proceso dentro de Excel.

En una granja de servidores de SharePoint, Analysis Services se ejecuta en un servidor de aplicaciones donde está emparejado con servicios relacionados que administran solicitudes de datos PowerPivot.

El siguiente diagrama (Figura 7-2) ilustra los componentes de servidor y cliente de PowerPivot:

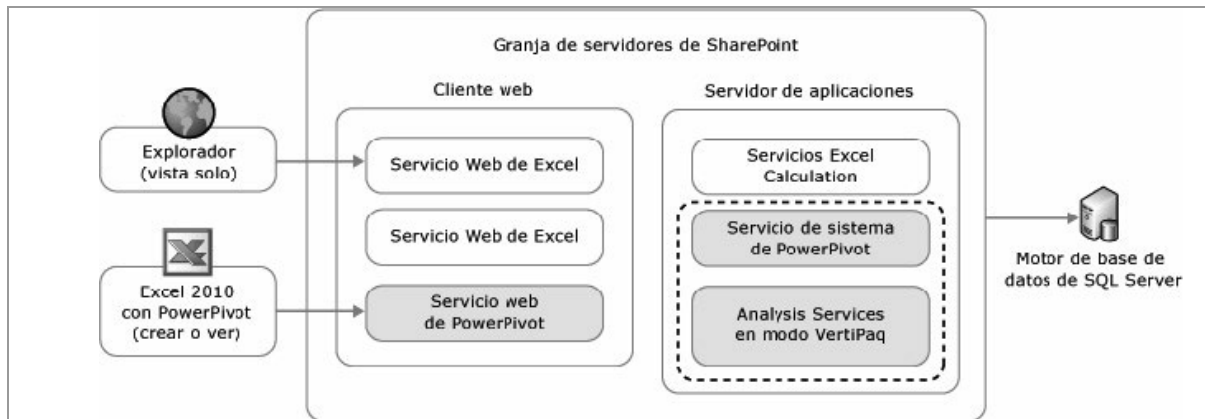


Figura 7-2

El servicio web de PowerPivot se ejecuta en un servidor de aplicaciones web. Redirige las solicitudes desde la aplicación web a una instancia de Servicio de sistema de PowerPivot en la granja.

El Servicio de sistema de PowerPivot emite solicitudes de carga al servidor de Analysis Services y administra las conexiones salientes a los datos que ya están cargados en la memoria, almacenando en memoria caché o descargando los datos, si ya no se utilizan o si se produce alguna contención con los recursos del sistema. También realiza un seguimiento de la actividad de los usuarios. Los datos de estado de servidor y otros datos de uso se recopilan y se presentan en informes para indicar el grado de idoneidad del funcionamiento del sistema.

Una instancia de servidor de Analysis Services en modo integrado de SharePoint completa la implementación. Carga, consulta y descarga los datos. También procesa los datos si el libro se configura para la actualización de datos PowerPivot. Cada instancia está unida estrechamente al Servicio de sistema de PowerPivot local que forma parte de la misma instalación.

POWER VIEW PARA SHAREPOINT SERVER: CREAR, GUARDAR E IMPRIMIR INFORMES

Power View para SharePoint es una aplicación de Silverlight basada en explorador, una característica del complemento SQL Server Service Reporting Services para Microsoft SharePoint Server 2010 y 2013. Crear, abrir y guardar informes de Power View (archivos RDLX) son acciones que se realizan en SharePoint Server 2010 y 2013.



Crear un informe en Power View para SharePoint Server

Para crear un informe de Power View, inicie Power View desde un archivo de modelo de datos en SharePoint Server 2010 y 2013. Los modelos, o las conexiones a los modelos, pueden estar en una biblioteca de documentos de SharePoint Server o en una galería de PowerPivot, que es una biblioteca de documentos especial de SharePoint Server que permite obtener una completa vista previa y administrar los documentos de libros de Microsoft Excel publicados que contienen modelos de datos.

Para crear un informe de Power View desde una galería de PowerPivot:

Haga clic en el icono *Crear informe* de Power View en la esquina superior derecha del archivo de Excel (XLSX).

Para crear un informe de Power View desde una biblioteca de documentos de SharePoint Server:

Haga clic en un origen de datos compartido (RSDS)  ¿o archivo de conexión de BISM (BISM)  para iniciar Power View.

El entorno de diseño de Power View se abre en la vista donde se crean los informes, con los campos del origen de datos compartido (RSDS) o el archivo de conexión de BISM (BISM) en la lista de campos.

Abrir un informe existente de Power View

Al abrir informes en una galería de PowerPivot, se puede optar por abrir el informe en una vista específica.

Para abrir un informe en una galería de PowerPivot:

- En Internet Explorer, vaya al sitio de galería de PowerPivot que hospeda los informes de Power View.
- Haga clic en cualquiera de las imágenes del informe.
- El informe se abre en la vista en modo de lectura.
- Para modificar el informe, haga clic en *Editar informe* en la esquina superior izquierda.

Para abrir un informe en una biblioteca de documentos de SharePoint Server.

- En Internet Explorer, vaya a la página principal de la biblioteca de documentos de SharePoint que hospeda los informes de Power View.
- Para abrir un informe en modo de lectura, haga clic en el título del informe.
- Para modificar el informe, haga clic en *Editar informe* en la esquina superior izquierda.

Guardar un informe

Los informes de Power View (archivo RDLX) se guardan en una biblioteca de documentos compartidos o una galería de Power Pivot, en el mismo servidor de SharePoint que el modelo desde el que se ha iniciado Power View.

- Para guardar el informe, en el menú *Archivo* de Power View ssCrescent, haga clic en *Guardar* o en *Guardar como*. La primera vez que guarde el informe, la ubicación predeterminada será la carpeta donde se encuentra el modelo.
- Para guardarlo en una ubicación distinta, desplácese a esa ubicación y haga clic en *Guardar*. En el cuadro de diálogo *Guardar como*, en el

campo *Nombre de archivo*, escriba el nombre del archivo.

- De forma predeterminada, la casilla *Guardar imágenes de vista previa con el informe* está seleccionada. Por razones de privacidad, quizá desee desactivarla y no guardar las imágenes de vista previa.
- Haga clic en *Guardar*.

El informe se guardará. Para salir de Power View y volver al sitio de SharePoint, haga clic en el botón *Atrás* del explorador.

Permisos para Power View

Power View usa permisos de SharePoint para controlar el acceso a los informes de Power View. Si tiene los permisos *Abrir elementos* para una carpeta de SharePoint, puede abrir cualquier informe de Power View de dicha carpeta en modo de edición o lectura. Por lo tanto, puede modificar el informe en modo de edición tanto como desee. Sin embargo, solo podrá guardar los cambios si dispone de los permisos *Agregar elementos en la biblioteca o carpeta de destino* o los permisos *Editar elementos* para sobrescribir el documento existente.

También podrá exportar un informe a PowerPoint si dispone de los permisos *Abrir elementos*. Sin embargo, no podrá exportar informes a PowerPoint con cambios no guardados. Por lo tanto, si solo tiene permisos del tipo *Abrir elementos*, podrá exportar informes tal cual, pero no podrá modificarlos ni exportarlos. Para ello, primero deberá guardar los cambios, lo que implica que necesita los permisos *Agregar elementos* o *Editar elementos*.

Formato de archivo RDLX: Power View crea los archivos con el formato de archivo RDLX. Estos no son compatibles con los archivos RDL creados en el Generador de informes o SQL Server Reporting Services (SSRS). Los archivos RDL no se pueden abrir ni modificar en Power View, y viceversa.

Exportar a PowerPoint desde Power View en SharePoint

Puede exportar una versión interactiva del informe de Power View de

SharePoint a PowerPoint. Cada vista de Power View se convierte en una diapositiva independiente de PowerPoint.

Hay dos versiones de Power View: Power View para Excel 2013 y Power View para SharePoint Server 2010 y 2013. Solo se puede exportar a PowerPoint desde Power View para SharePoint Server.

La interacción con los informes de Power View exportados a PowerPoint es similar a la Interacción con las vistas de Power View en los modos de lectura y de pantalla completa de Power View. En los modos de presentación de diapositivas y de lectura de Microsoft PowerPoint, puede interactuar con las visualizaciones y los filtros que haya agregado el creador del Informe a cada vista, pero no puede crear visualizaciones o filtros.

Actualizar los datos del informe

Puede actualizar los datos en un Informe de Power View sin actualizar también la página. Haga clic en el botón *Actualizar* en la barra de herramientas de acceso rápido de Power View.

Si hace clic en el botón *Actualizar* en el explorador y, a continuación, hace clic en *Salir* de esta página, perderá los cambios realizados en el Informe desde que lo guardó por última vez.

Imprimir vistas en un informe de Power View para SharePoint Server

Puede imprimir un informe de Power View en los modos de diseño o lectura, pero no en modo de pantalla completa. Power View Imprime una vista a la vez: la vista actual.

1. Para Imprimir una vista, en el menú Archivo de Power View haga clic en Imprimir. Se abre el cuadro de diálogo Imprimir del explorador.
2. Haga clic en Imprimir. La vista siempre se Imprime en orientación horizontal, Independientemente de la configuración del cuadro de diálogo Imprimir. Imprime exactamente lo que aparece en la vista. Por ejemplo, Power View imprime:

- La parte de una visualización que está visible al Imprimir, si la visualización tiene una barra de desplazamiento.
- El mosaico seleccionado en un contenedor en mosaico.
- El área de filtros, si se expande.
- El marco actual de un gráfico de dispersión o burbujas con un eje de reproducción.

Informes de Power View basados en modelos de datos

En SharePoint, Power View se inicia siempre desde un modelo de datos. El modelo puede ser:

1. Un libro de Excel (XLSX) con un modelo de datos en una galería de PowerPivot en SharePoint Server. Usar una galería de PowerPivot.
2. Un origen de datos compartidos (RSDS) ¿2.en SharePoint Server con un tipo de origen de datos de Microsoft Business Intelligence Semantic Model, basado en:
 - Un libro de Excel.
 - Un modelo tabular en un servidor de Analysis Services. Crear un origen de datos compartido para un modelo de datos.
 - Un modelo multidimensional en un servidor SSAS. Descripción de los objetos de modelo multidimensional de Power View.
 - Un archivo de conexión de BISM (BISM) Abasado en un modelo tabular en un servidor de Analysis Services. Los archivos de conexión de BISM pueden estar en una biblioteca de documentos estándar de SharePoint Server o en una galería de PowerPivot. Usar una conexión de modelo semántico de BI.

Los modelos de datos son la siguiente generación de modelos diseñada específicamente para las cargas de trabajo de informes y análisis. Los modelos de datos pueden importar datos desde una gran variedad de orígenes de datos, entre los que se incluyen:

- SQL Server
- DB2

- O Data
- Oracle
- Teradata
- Y otros

Se puede crear un modelo de datos en Excel, mejorar el modelo de datos con PowerPivot para Excel y después guardar el modelo en una biblioteca de documentos de SharePoint Server 2013 o en una galería de PowerPivot. Los programadores de modelos de una organización de TI crean modelos en SQL Server Data Tools (SSDT) y, después, los implementan en un servidor de SQL Server Analysis Services (SSAS).

Descripción de los objetos de modelo multidimensional de Power View

Se puede usar Power View iniciado desde SharePoint Server para explorar datos de manera interactiva y crear visualizaciones dinámicas de modelos multidimensionales de Analysis Services en Microsoft SQL Server.

Cuando se usa Power View para visualizar modelos multidimensionales, es importante que tenga en cuenta que está trabajando con una representación de tipo de modelo tabular de un modelo multidimensional. Los modelos tabulares tienen objetos como tablas y columnas, y al igual que con los modelos multidimensionales, medidas y KPI.

Cuando usted o un administrador crea una conexión de origen de datos compartida en SharePoint, se especifica un nombre o perspectiva de cubo nombre en la cadena de conexión. Solo puede especificarse un cubo o perspectiva. El cubo o la perspectiva especificada en la conexión a un origen de datos compartida se exponen como modelo en la Lista de campos de Power View. Los objetos del modelo se exponen como campos que se pueden usar como visualizaciones de tabla en una vista. Hay, sin embargo, algunas diferencias en cómo aparecen ciertos objetos multidimensionales en Power View. Al igual que con los modelos tabulares, la Lista de campos muestra todos los objetos que se pueden usar en una vista.

Los modelos multidimensionales tienen dimensiones. En este ejemplo, la Lista de campos contiene objetos de la dimensión Producto (Figura 7-3). El producto aparece como una tabla en el modelo Adventure Works (cubo). Una tabla o dimensión tiene también otros objetos.

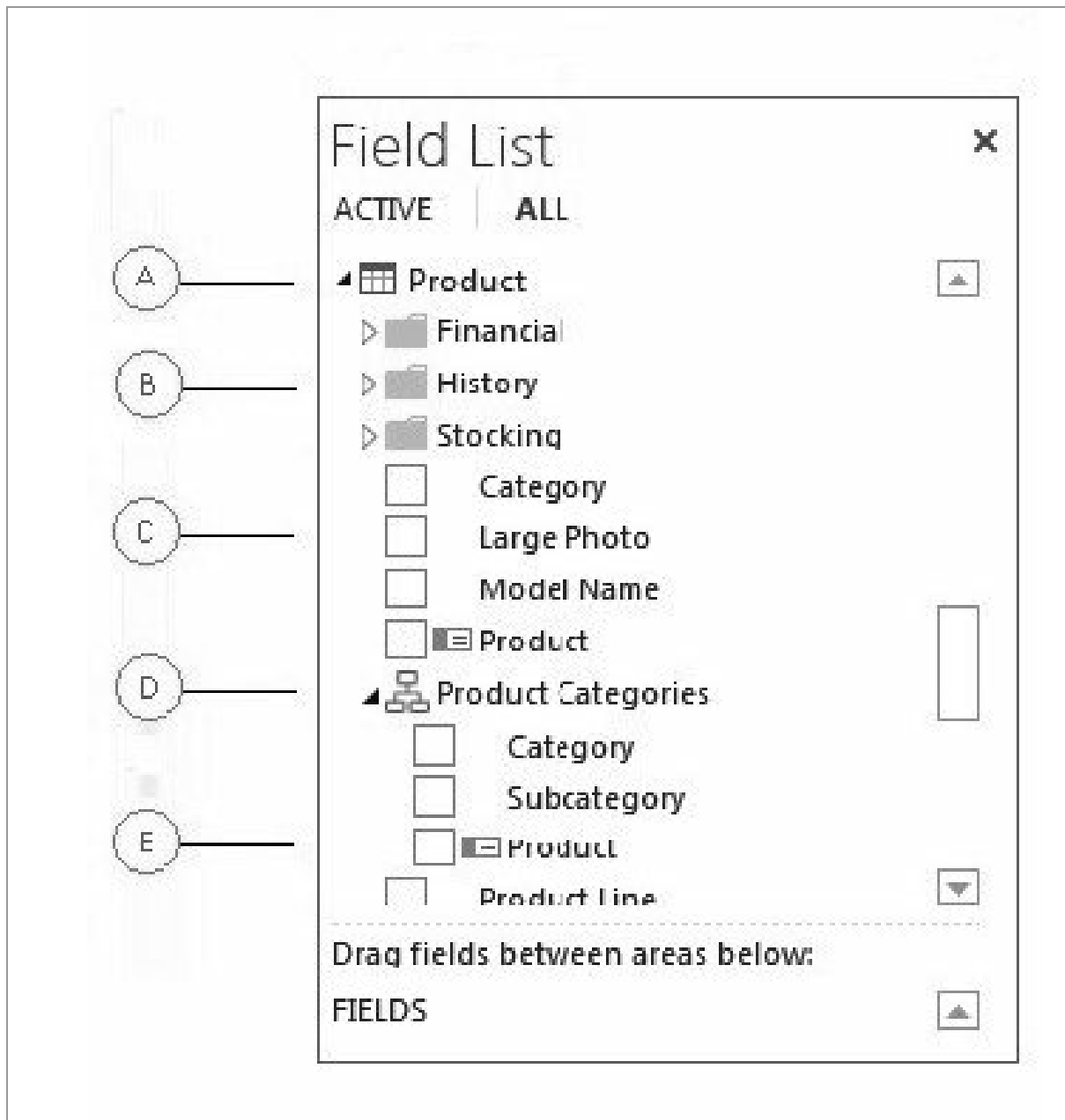


Figura 7-3

- Las dimensiones aparecen como tablas y pueden expandirse para mostrar otros objetos en la tabla (dimensión). El cubo de Adventure Works tiene muchas tablas, cuenta a través de moneda de origen.
- Mostrar carpetas aún más, dividir y clasificar cómo aparecen los objetos asociados en las herramientas de cliente. Mostrar carpetas

puede aparecer en la Lista de campos debajo de las tablas de dimensiones y de las tablas de grupos de medida.

c) Los atributos de dimensión aparecen como columnas en una tabla. Una única columna (atributo) puede aparecer en una tabla y de nuevo en una jerarquía, a menos que se oculte explícitamente.

d) Las jerarquías de usuario y elementos primarios y secundarios se Incluyen en las tablas (dimensiones). Las jerarquías se pueden expandir para mostrar columnas (niveles) en ellas. Cuando se selecciona un nivel, todos los niveles anteriores también se seleccionan automáticamente. Puede deseleccionar los niveles más altos para quitarlos de la visualización. Esto puede resultar útil cuando determinados campos se exponen solo en una jerarquía.

e) Los objetos con este icono indican que objeto es una clave

Los modelos multidimensionales también tienen grupos de medidas (Figura 7-4), también conocidos como dimensiones de medidas, que contienen las medidas que se pueden utilizar para agregar datos en el análisis.

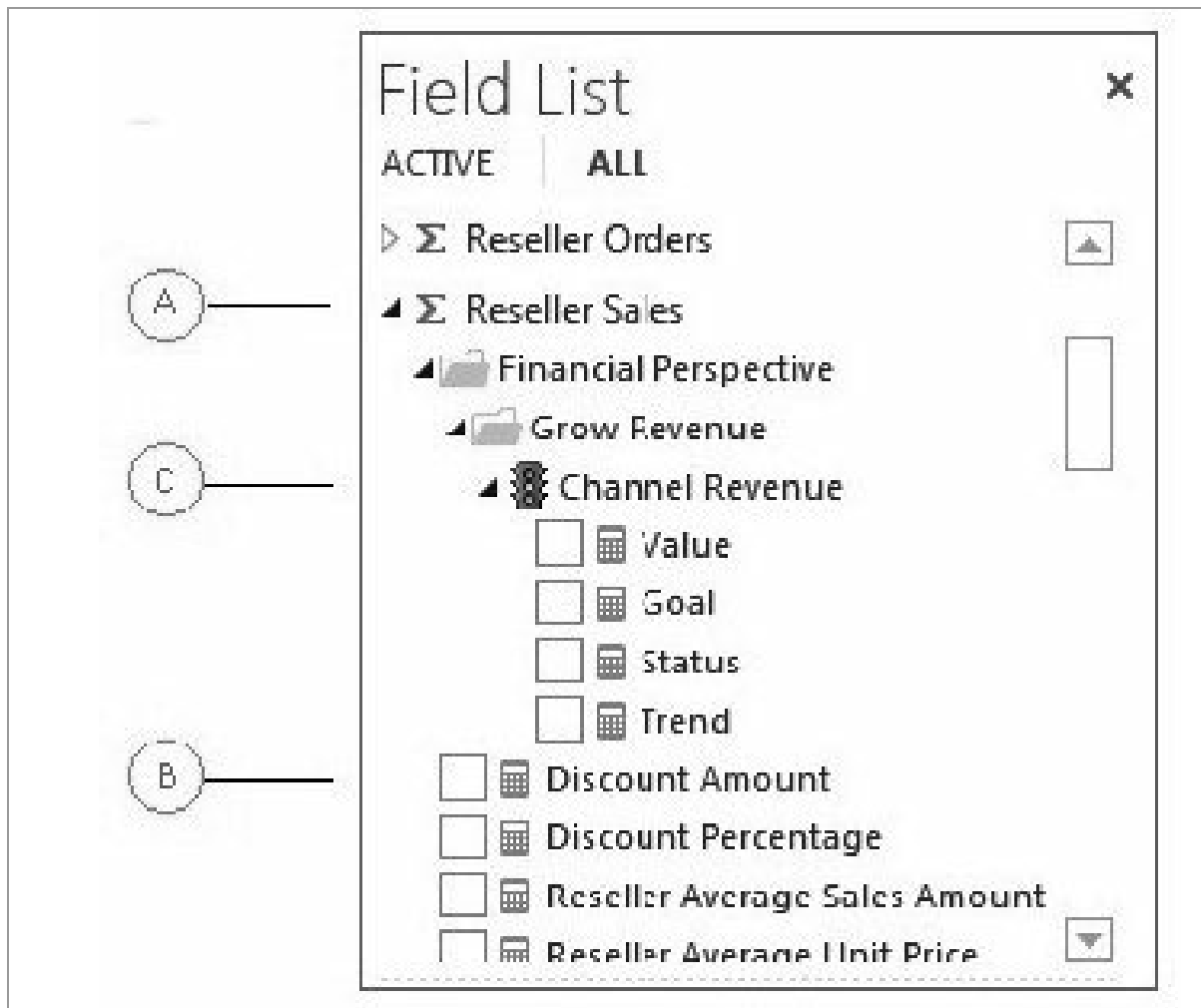


Figura 7-4

a) Los grupos de medidas aparecen como tablas; sin embargo, a diferencia de las tablas de una dimensión, una tabla de un grupo de medidas se identifica con un icono sigma.

b) Las medidas aparecen en tablas de grupos de medidas y se identifican con un icono de calculadora. Si el cubo tiene solo una medida, se incluirá en un grupo de medidas asociado si hay uno, o en una sola tabla denominada Medidas.

c) Los KPI se incluyen en tablas de grupos de medidas asociadas y se identifican con un icono de luz. Por ejemplo, si en el modelo de Adventure Works expande Venta del distribuidor > Perspectivas financieras > Aumentar los ingresos, verá el KPI de Ingresos de canal y sus cuatro medidas: Valor, Objetivos, Estado y Tendencia.

Gráficos y otras visualizaciones en Power View

En Power View en SharePoint 2013 y Excel 2013, puede crear rápidamente una variedad de visualizaciones de datos, desde tablas y matrices hasta gráficos de barras, columnas y burbujas, así como conjuntos de gráficos de múltiples. Sea cual sea la visualización que desee crear, siempre comenzará en una hoja de Power View creando una tabla, que más adelante se convertirá fácilmente en otras visualizaciones para determinar cuál es la que mejor ilustra los datos.

Para crear una visualización:

1. Cree una tabla en la hoja de Power View activando una tabla o campo en la lista de campos o arrastrando un campo de la lista de campos a la hoja. Power View dibuja la tabla en la hoja, muestra los datos reales y agrega automáticamente encabezados de columna.
2. Convierta la tabla en una visualización eligiendo un tipo de visualización en la pestaña Diseño. Dependiendo de los datos de la tabla, diferentes tipos de visualizaciones estarán disponibles para darle la mejor visualización para esos datos.

A continuación se presenta una visualización típica en Power View (Figura 7-5):

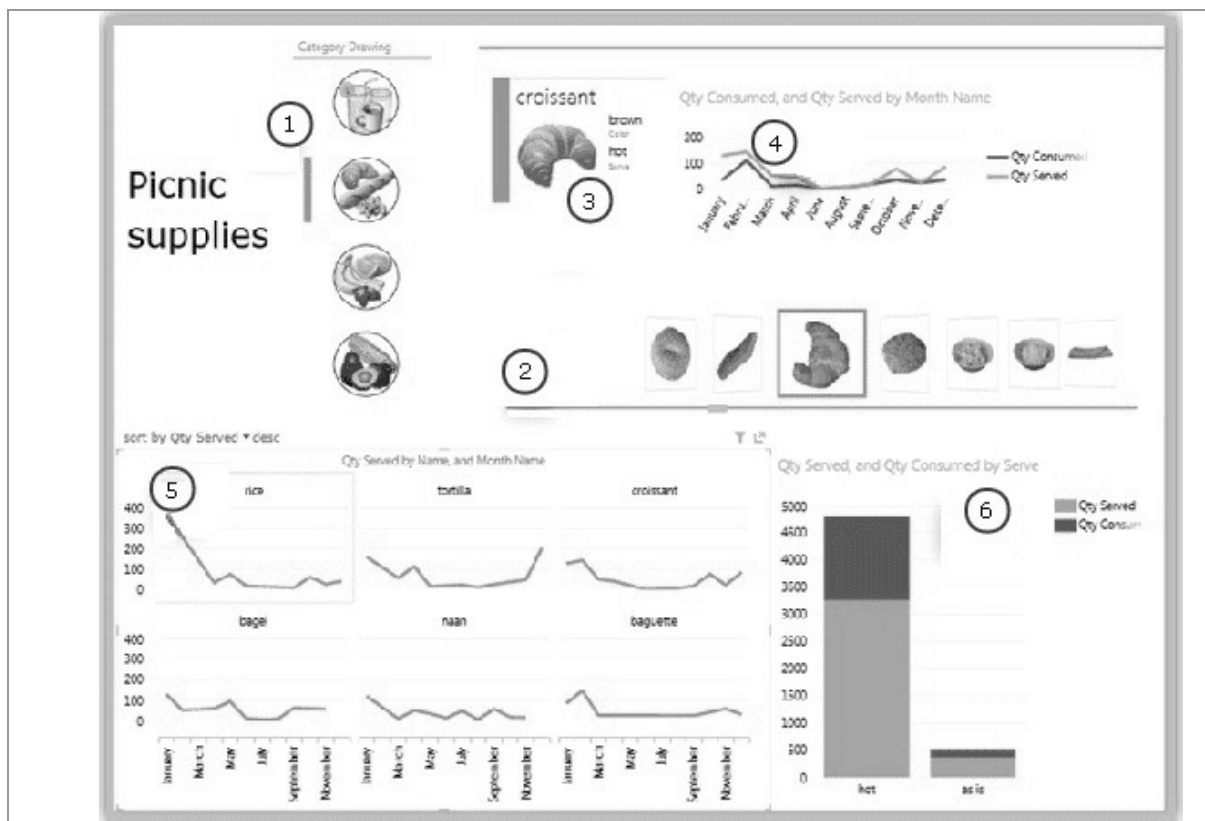


Figura 7-5

Sus elementos son los siguientes:

1. Segmentación que filtra el informe por panes.
2. Navegación de flujo de mosaicos para los mosaicos, actualmente en croissant.
3. Tarjeta en un contenedor en mosaico, filtrado por el mosaico actual (croissant).
4. Gráfico de líneas en el contenedor en mosaico que muestra las cantidades consumidas y servidas, filtrado por croissants de enero a diciembre.
5. Múltiplos, filtrados por panes y ordenados en orden descendente por cantidad servida.
6. Gráfico de columnas filtrado por panes, que muestra las cantidades servidas y consumidas.

Todas las visualizaciones comienzan con una tabla (Figura 7-6).

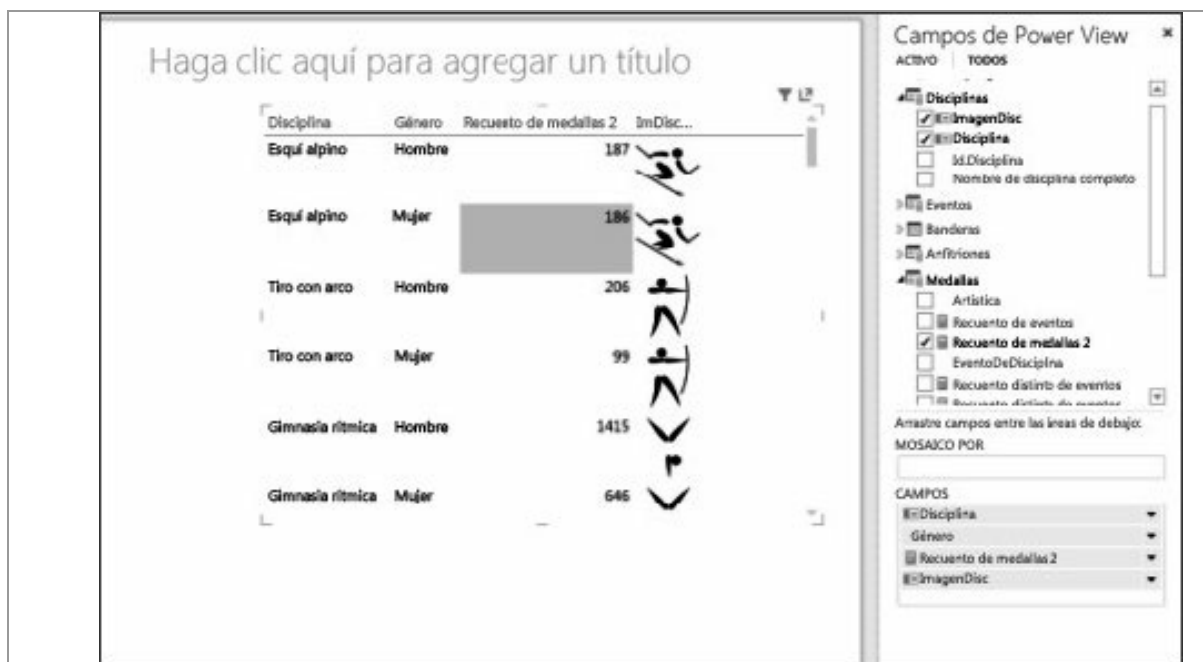


Figura 7-6

Power View ofrece una serie de opciones de gráficos: circulares, columnas, barras, líneas, dispersión y burbujas. Los gráficos pueden tener varios campos numéricos y varias series. En cuanto a las opciones de diseño en un gráfico puede mostrar u ocultar etiquetas, leyendas y títulos.

Los gráficos son Interactivos. Al hacer clic en un valor de un gráfico,

se:

- Resalta ese valor en el gráfico.
- Filtra por ese valor en todas las tablas, matrices y mosaicos del informe.
- Destaca ese valor en todos los demás gráficos del informe.

Los gráficos también son interactivos en las presentaciones. Por ejemplo, en los modos de lectura y de pantalla completa de Power View en SharePoint o de una hoja de Power View en un libro de Excel guardado en Excel Services o visto en Office 365.

Gráficos circulares

Los gráficos circulares de Power View pueden ser simples o sofisticados. Puede crear un gráfico circular que desglose la información al hacer doble clic en un sector o un gráfico circular que muestre subsectores dentro de los sectores de color más grandes (Figura 7-7).

Puede aplicar filtros cruzados de un gráfico circular a otro gráfico. Imagine que hace clic en una barra de un gráfico de barras: la parte del gráfico circular que se aplica a esa barra se resalta y se atenúa el color del resto del gráfico.

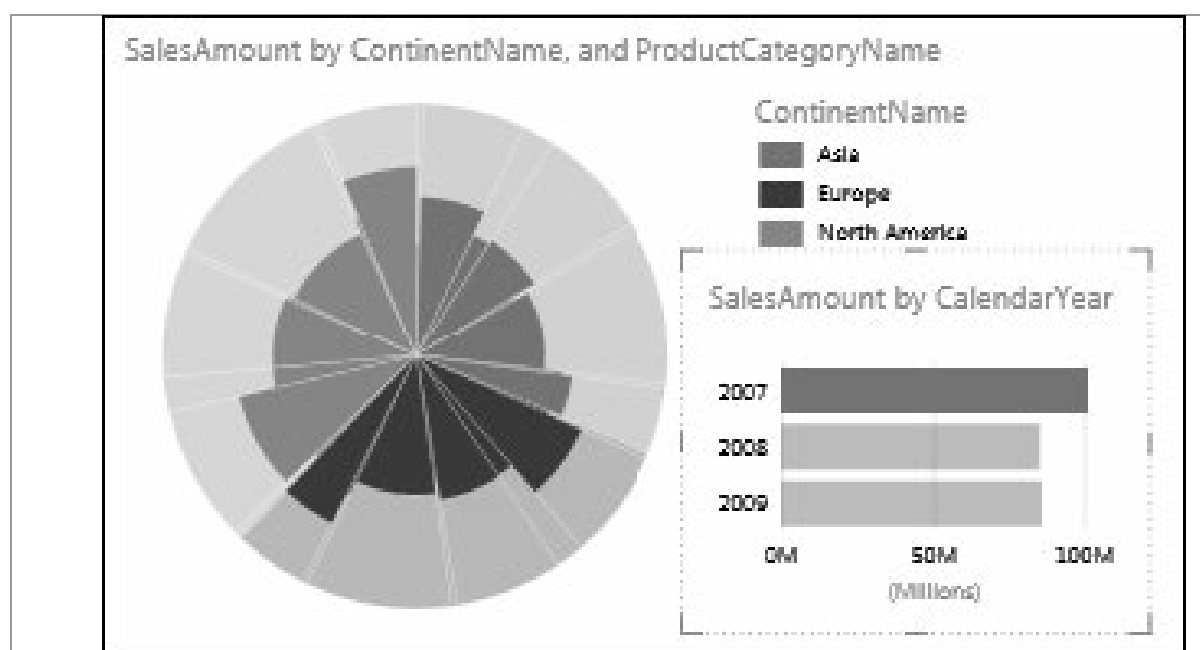


Figura 7-7

Gráficos de dispersión y de burbujas

Los gráficos de dispersión y de burbujas son una magnífica forma de mostrar grandes cantidades de datos relacionados en un mismo gráfico. En los gráficos de dispersión (Figura 7-8), el eje X muestra un campo numérico y el eje Y muestra otro, de modo que resulta fácil ver la relación entre los dos valores para todos los elementos del gráfico.

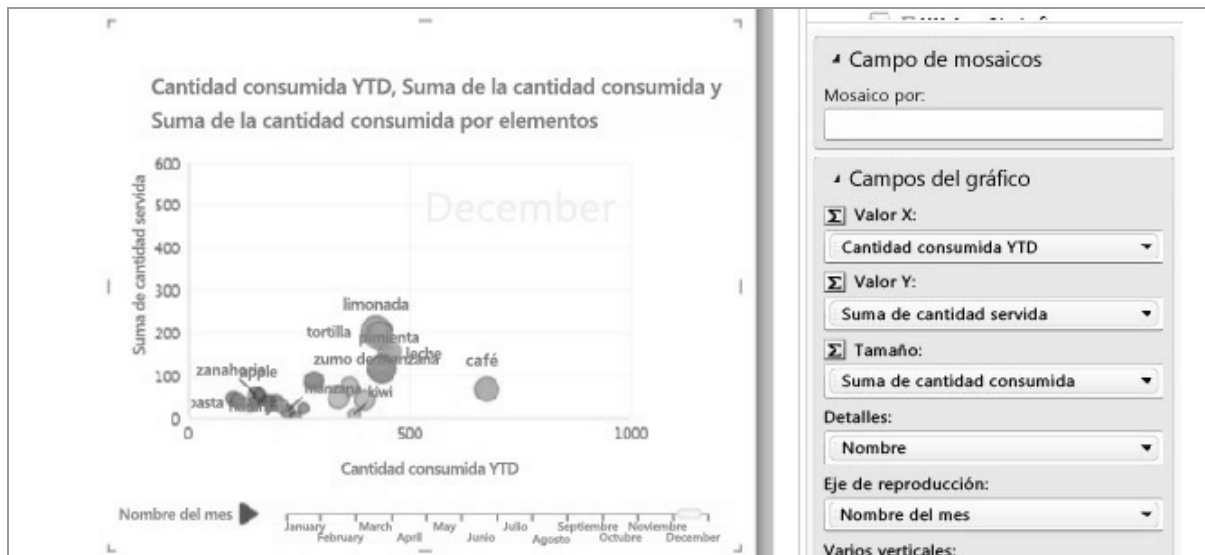


Figura 7-8

En un gráfico de burbujas (Figura 7-9), un tercer campo numérico controla el tamaño de los puntos de datos. Puede agregar un eje de “reproducción” a un gráfico de burbujas o dispersión para ver los datos conforme cambian con el tiempo.

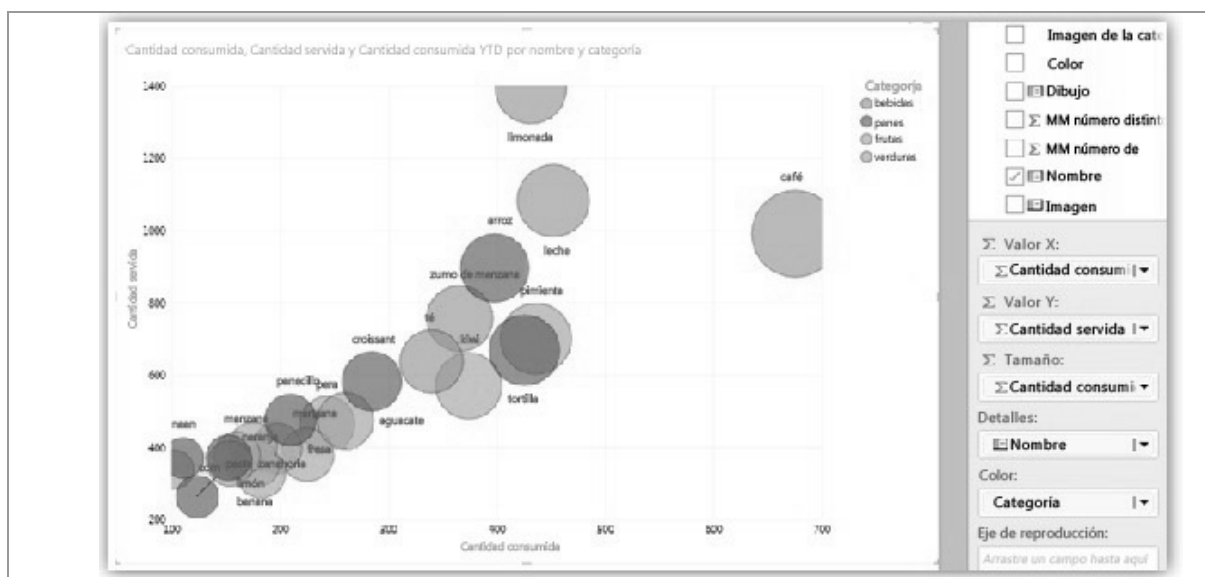


Figura 7-9

Gráficos de líneas, barras y columnas

Los gráficos de líneas, barras y columnas son útiles para comparar puntos de datos en una o varias series de datos. En los gráficos de líneas, barras y columnas, el eje X muestra un campo y el eje Y otro. De este modo es fácil ver la relación entre los dos valores de todos los elementos del gráfico.

Gráficos de barras

En los gráficos de barras, las categorías se organizan en torno al eje vertical y los valores en torno al eje horizontal. Considere la posibilidad de usar un gráfico de barras en cualquiera de estos casos:

- Tiene una o más series de datos que desea representar.
- Sus datos incluyen valores positivos, negativos y cero (0).
- Desea comparar los datos de varias categorías.
- Las etiquetas de eje son largas.

En Power View, dispone de tres subtipos de gráficos de barras para elegir: apilados, 100% apilados y agrupados.

Gráficos de columnas

Los datos organizados en columnas o filas en una hoja de cálculo se pueden representar en un gráfico de columnas. Los gráficos de columnas resultan útiles para mostrar cambios en los datos a lo largo de un período de tiempo o para ilustrar comparaciones entre elementos. En los gráficos de columnas, las categorías se organizan en torno al eje horizontal y los valores se organizan en torno al eje vertical.

En Power View, dispone de tres subtipos de gráficos de columnas para elegir: apilados, 100% apilados y agrupados.

Gráficos de líneas

Los gráficos de líneas (Figura 7-10) distribuyen los datos de categorías

de manera uniforme en un eje horizontal (categoría) y distribuyen todos los datos de valores numéricos a lo largo de un eje vertical (valor).

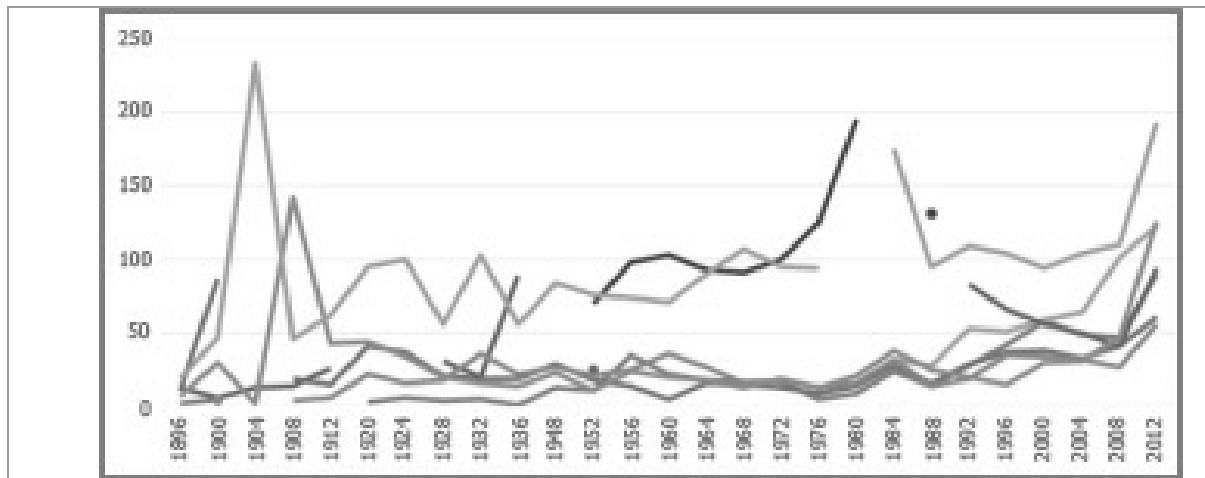


Figura 7-10

Considere la posibilidad de usar un gráfico de líneas con escala de tiempo en el eje horizontal. Los gráficos de líneas muestran las fechas en orden cronológico a intervalos concretos o unidades básicas, como el número de días, meses o años, incluso si las fechas de la hoja de cálculo no están en orden o en las mismas unidades básicas.

Mapas

Los mapas de Power View (Figura 7-11) utilizan los mosaicos de los mapas de Bing, por lo que puede hacer zoom y desplazarse como haría en cualquier otro mapa de Bing. Al agregar lugares y campos, se colocan puntos en el mapa. Cuanto mayor sea el valor, mayor será el punto. Cuando se agrega una serie de valores múltiples, se colocan gráficos circulares en el mapa y el tamaño del gráfico circular muestra el tamaño del total. Encontrará más información sobre los mapas en Power View.

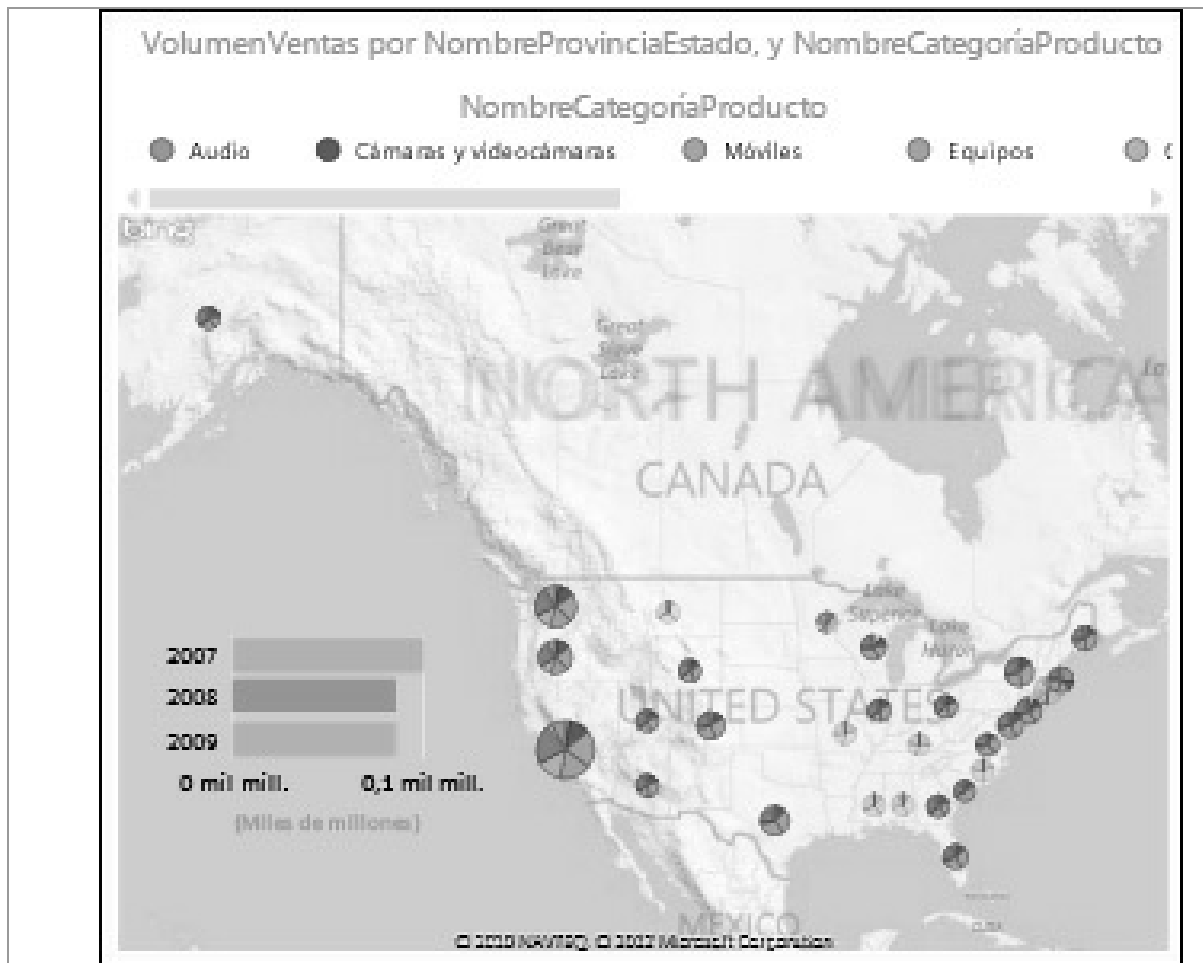


Figura 7-11

Múltiplos: un conjunto de gráficos con los mismos ejes

Con los múltiplos, puede crear una serie de gráficos con ejes X e Y idénticos para, a continuación, organizados unos junto a otros, lo que facilita la comparación de distintos valores al mismo tiempo. Los múltiplos también suelen llamarse “gráficos enrejados” (Figura 7-12).

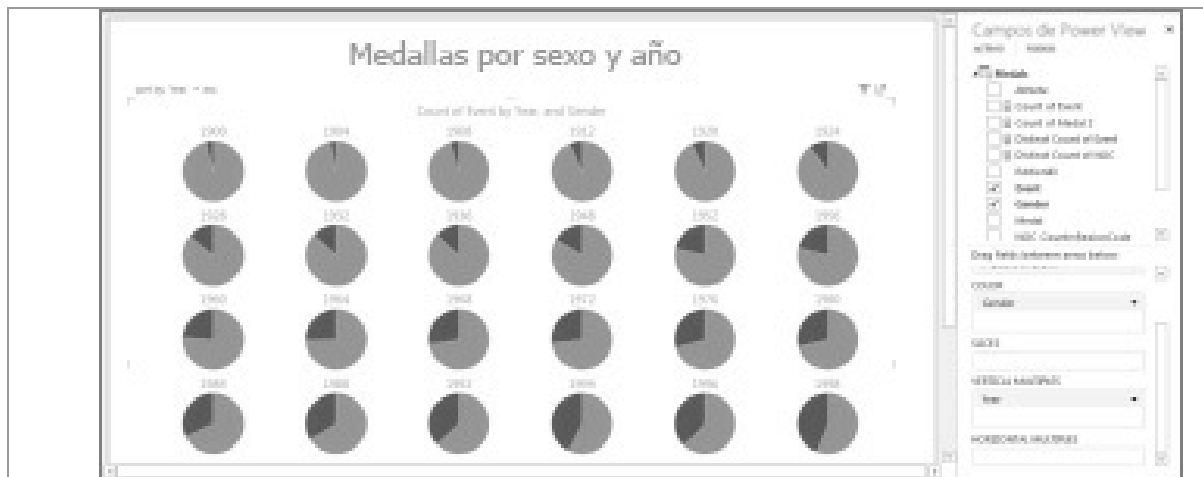


Figura 7-12

Matrices

Una matriz es similar a una tabla por el hecho de que está compuesta por filas y columnas. Sin embargo, las matrices tienen unas capacidades que no tienen las tablas:

- Muestran los datos sin repetir los valores.
- Muestran los totales y los subtotales por fila y columna.
- Con una jerarquía, es posible agregar detalles y resúmenes.
- Es posible expandir y contraer la visualización.

Tarjetas

Puede convertir una tabla en una serie de tarjetas que muestren los datos de cada fila de la tabla diseñada en formato de tarjeta, como una tarjeta de índice (Figura 7-13).


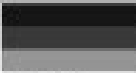

	Argentina 239
FlagURL	CountryRegion Medal Count
	Armenia 9
FlagURL	CountryRegion Medal Count
	Australia 1079
FlagURL	CountryRegion Medal Count

Figura 7-13

Mosaicos

Puede convertir una tabla o matriz en mosaicos para presentar los datos tabulares interactivamente. Los mosaicos son contenedores con una franja de navegación dinámica. Los mosaicos sirven de filtros: filtran el contenido que hay en el mosaico en función del valor seleccionado en la franja de pestañas. Puede agregar más de una visualización al mosaico y filtrar todas las visualizaciones por el mismo valor. Puede usar texto o imágenes como pestañas. Más información acerca de los mosaicos en Power View.

Esta imagen (Figura 7-14) muestra el número de medallas que han ganado los distintos países en una competición de patinaje de velocidad.

Medal Count by CountryRegion

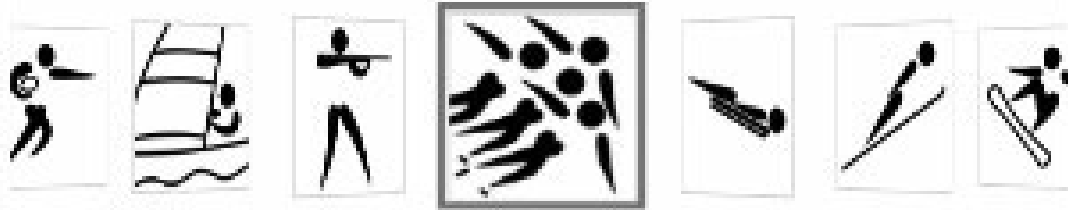
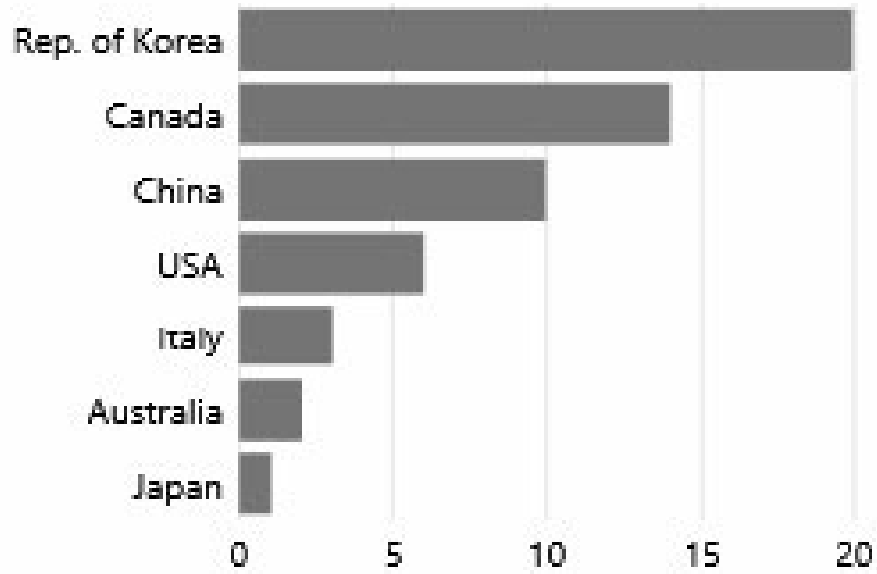
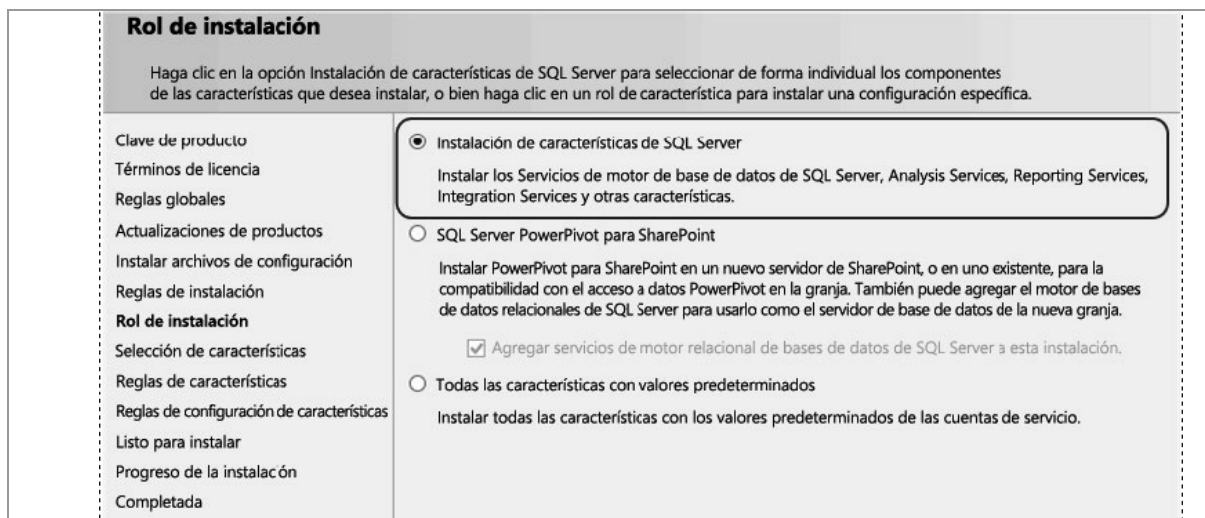


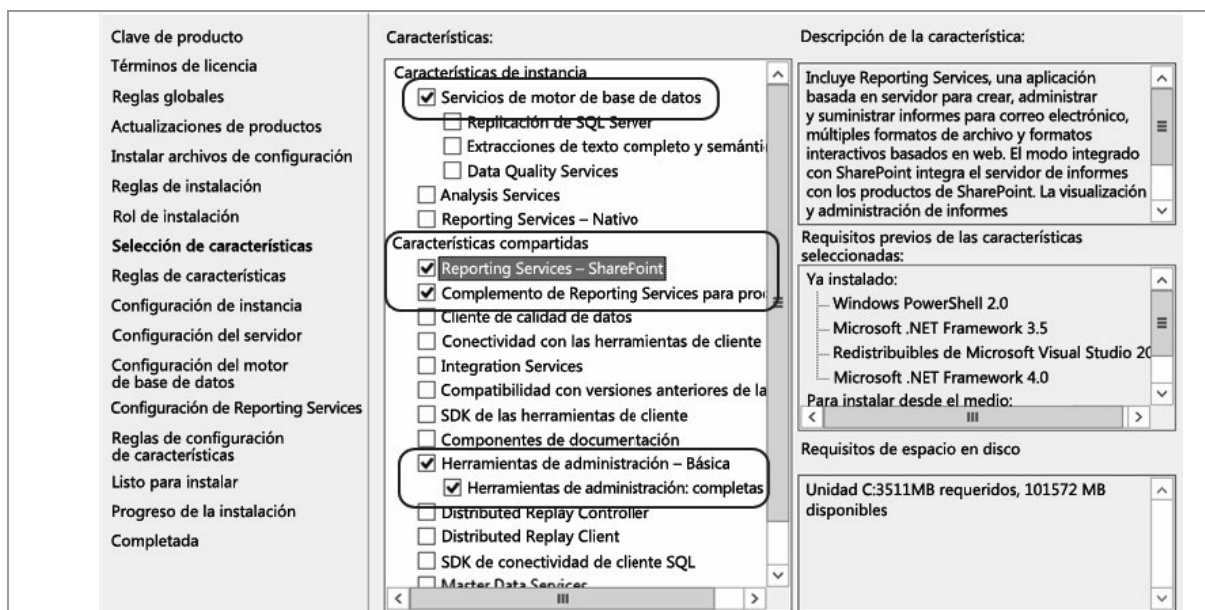
Figura 7-14

ANALYSIS SERVICES, INTEGRATION SERVICES Y REPORTING SERVICES

Para instalar Analysis Services, Integration Services y Reporting Services, basta elegir estas características BI de SQL Server, en la página *Rol de instalación del Asistente* para la instalación de SQL Server:



Para instalar Complemento de Reporting Services para productos de SharePoint, basta elegirlo en la página Selección de características del Asistente para la instalación de SQL Server:



Informes de Reporting Services (SSRS)

Los informes de SQL Server Reporting Services son definiciones de informe basadas en XML que incluyen los datos y elementos de diseño de los informes (Figura 7-15). En un sistema de archivos de cliente, las definiciones de informe tienen la extensión de archivo .rdl. Una vez que se publica un informe, se convierte en un elemento de informe que se almacena en el servidor de informes o en el sitio de SharePoint. Los informes constituyen un único componente de la plataforma de generación de informes basada en servidor que proporciona Reporting Services.

Puede usar las soluciones de informes de Reporting Services para:

- Usar un solo conjunto de orígenes de datos que proporcione una única versión de los hechos. Puede basar los informes en esos orígenes de datos para proporcionar una vista de datos unificada que facilite la toma de decisiones comerciales.
- Visualizar los datos de formas diversas e interconectadas a través de las regiones de datos. Puede mostrar los datos organizados en tablas, matrices o tablas de referencias cruzadas; también puede expandir o contraer grupos, gráficos, medidores, indicadores o KPI, y mapas, e incluso tiene la posibilidad de incluir gráficos en las tablas.
- Ver los informes para su propio uso o publicar informes en un servidor de informes o un sitio de SharePoint para compartirlos con el equipo o la organización.
- Definir un informe una sola vez y presentarlo de diversas maneras. Puede exportar el informe en varios formatos de archivo o entregar el informe a los suscriptores en forma de correo electrónico o de un archivo compartido. Puede crear informes vinculados que apliquen distintos conjuntos de parámetros a la misma definición de informe.
- Usar elementos de informe, orígenes de datos compartidos, consultas compartidas y subinformes para definir las visualizaciones de datos para su reutilización.
- Administrar los orígenes de datos del informe con independencia de la definición de informe. Por ejemplo, puede cambiar de un origen de datos de prueba a un origen de datos de producción sin cambiar el informe.

- Crear informes con un diseño libre. El diseño del informe no está restringido a bandas de información. Puede organizar la visualización de los datos de la página de forma que facilite su comprensión, mejore su entendimiento y promueva la entrada en acción.
- Habilitar acciones para la obtención de detalles, alternadores para expandir y contraer, botones de ordenación, información sobre herramientas y parámetros de informe que permitan al lector interactuar con el informe. Puede combinar los parámetros de informe con sus propias expresiones para que los lectores del informe puedan controlar el modo en que se filtran, agrupan y ordenan los datos.
- Definir expresiones que le proporcionan la capacidad de personalizar el modo en que se filtran, agrupan y ordenan los datos.

Al crear un informe, tiene que definir un archivo de definición de informe (.rdl) en formato XML. Este archivo contiene toda la información necesaria para combinar los datos y el diseño del informe mediante el procesador de informes. Cuando consulte un informe, este avanzará a través de los pasos siguientes:

- **Compilación.** Se evalúan las expresiones de la definición de informe y se almacena el formato intermedio de compilación internamente en el servidor de informes.
- **Proceso.** Se ejecutan las consultas de conjuntos de datos y se combina el formato intermedio con los datos y el diseño.
- **Representación.** El informe procesado se envía a una extensión de representación para determinar cuánta información cabe en cada página y crear el informe paginado.
- **Exportación (opcional).** El informe se exporta a un formato de archivo diferente.

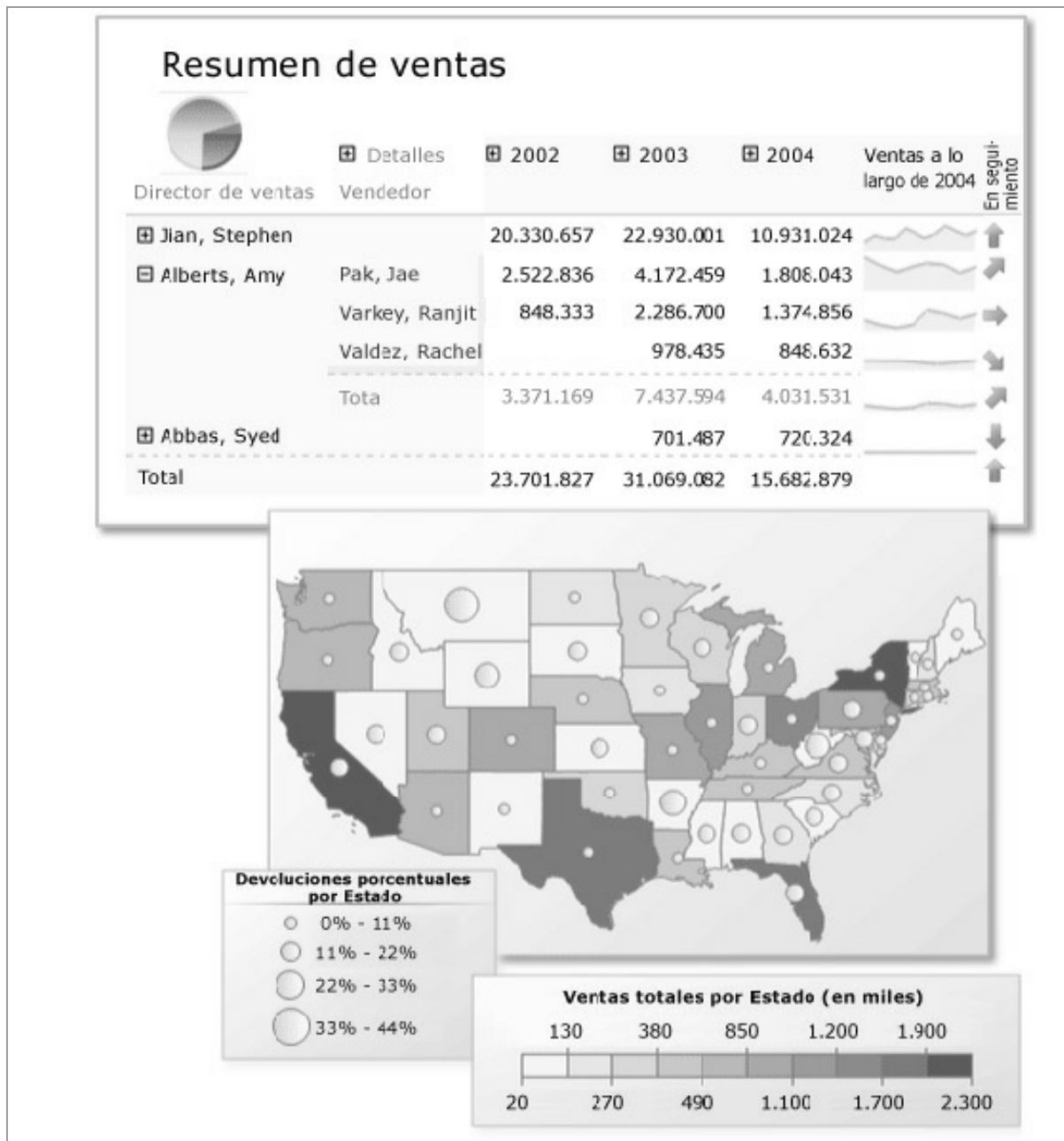


Figura 7-15

SQL Server Integration Services

Microsoft Integration Services es una plataforma para la creación de soluciones empresariales de transformaciones de datos e Integración de datos. Integration Services sirve para resolver complejos problemas empresariales mediante la copia o descarga de archivos, el envío de mensajes de correo electrónico como respuesta a eventos, la actualización de almacenamientos de datos, la limpieza y minería de datos, y la administración de objetos y datos de SQL Server. Los paquetes pueden

funcionar en solitario o junto con otros paquetes para hacer frente a las complejas necesidades de la empresa. Integration Services puede extraer y transformar datos de diversos orígenes como archivos de datos XML, archivos planos y orígenes de datos relacionales y, después, cargar los datos en uno o varios destinos.

Integration Services contiene un variado conjunto de tareas y transformaciones integradas, herramientas para la creación de paquetes y el servicio Integration Services para ejecutar y administrar los paquetes. Las herramientas gráficas de Integration Services se pueden usar para crear soluciones sin escribir una sola línea de código. También se puede programar el amplio modelo de objetos de Integration Services para crear paquetes mediante programación y codificar tareas personalizadas y otros objetos de paquete.

Un paquete es una colección organizada de conexiones, elementos de flujo de control, elementos de flujo de datos, controladores de eventos, variables, parámetros y configuraciones que se pueden ensamblar con la ayuda de las herramientas gráficas de diseño proporcionadas por SQL Server Integration Services o mediante programación. A continuación guarda el paquete completado en SQL Server, el Almacén de paquetes SSIS o el sistema de archivos, o puede implementar el proyecto de *SSnover* en el servidor SSIS. El paquete es la unidad de trabajo que se recupera, ejecuta y guarda.

Al crear por primera vez un paquete, es un objeto vacío que no hace nada. Para agregar funcionalidad a un paquete, debe agregarle un flujo de control y, opcionalmente, uno o más flujos de datos.

El siguiente diagrama (Figura 7-16) muestra un paquete individual que contiene un flujo de control con una tarea Flujo de datos que, a su vez, contiene un flujo de datos.

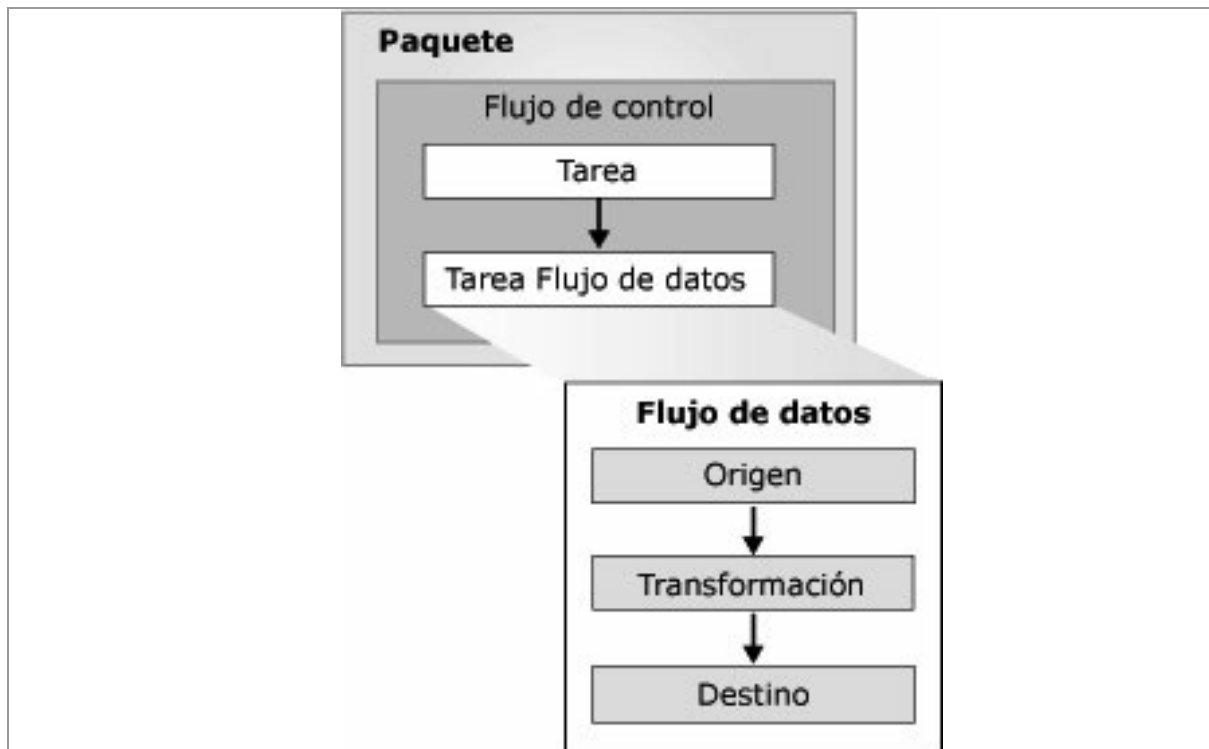


Figura 7-16

Una vez creado el paquete básico, puede agregarle características avanzadas como registro y variables para extender su funcionalidad.

Una vez completado el paquete, puede configurarse estableciendo propiedades de nivel de paquete que permitan implementar la seguridad, reiniciar paquetes desde puntos de comprobación o incorporar transacciones en el flujo de trabajo de paquetes.

Un flujo de control consta de una o más tareas y contenedores que se ejecutan cuando se ejecuta el paquete. Para controlar el orden o definir las condiciones para ejecutar la siguiente tarea o contenedor del flujo de control del paquete, puede usar restricciones de precedencia para conectar las tareas y los contenedores de un paquete. También se puede agrupar y ejecutar repetidamente un subconjunto de tareas y contenedores como una unidad en el flujo de control del paquete.

Un flujo de datos consta de los orígenes y destinos que extraen y cargan datos, las transformaciones que modifican y extienden datos, y las rutas que vinculan orígenes, transformaciones y destinos. Para poder agregar un flujo de datos a un paquete, el flujo de control de paquetes debe incluir una tarea Flujo de datos. La tarea Flujo de datos es el ejecutable del paquete SSIS que crea, organiza y ejecuta el flujo de datos. Se abre una Instancia Independiente del motor de flujo de datos para cada tarea Flujo

de datos de un paquete.

Un paquete suele incluir al menos un administrador de conexiones. Un administrador de conexiones es un vínculo entre un paquete y un origen de datos que define la cadena de conexión para acceder a los datos que las tareas, transformaciones y controladores de eventos del paquete usan. Integratlon Services Incluye tipos de conexiones para orígenes de datos tales como archivos de texto y XML, bases de datos relacionales y proyectos y bases de datos de Analysis Services.

Los paquetes se utilizan con frecuencia como plantillas para generar paquetes que compartan una funcionalidad básica. Puede generar el paquete básico y luego copiarlo, o puede designar que el paquete sea una plantilla. Por ejemplo, un paquete que descarga y copia archivos, y luego extrae los datos, puede incluir las tareas FTP y Sistema de archivos en un bucle Foreach que enumera archivos de una carpeta. También podría incluir administradores de conexión de archivos planos para el acceso a los datos, y orígenes de archivos planos para la extracción de los datos. El destino de los datos varía y se agrega a cada nuevo paquete una vez copiado del paquete básico. También puede crear paquetes y luego usarlos como plantillas para los nuevos paquetes que agregue a un proyecto de integration Services.

Analysis Services

Analysis Services es un motor de datos analíticos en línea que se usa en soluciones de ayuda a la toma de decisiones y Business Intelligence (BI), y proporciona los datos analíticos para informes empresariales y aplicaciones cliente como Excel, informes de Reporting Services y otras herramientas de BI de terceros. Un flujo de trabajo típico para Analysis Services incluye la creación de un modelo de datos OLAP o tabular, la implementación del modelo como base de datos en una instancia de Analysis Services, el procesamiento de la base de datos para cargarla con datos y, a continuación, la asignación de permisos para permitir el acceso a datos. Cuando esté listo, se puede obtener acceso a este modelo de datos con varios fines desde cualquier aplicación cliente que admita Analysis Services como origen de datos.

Para crear un modelo, use SQL Server Data Tools y elija una plantilla

de proyecto Tabular o Multidimensional y Minería de datos. Los modelos se rellenan con datos procedentes de sistemas de datos externos, normalmente almacenamientos de datos hospedados en un motor de base de datos relacional de SQL Server o de Oracle (los modelos tabulares admiten tipos de orígenes de datos adicionales). Los modelos especifican objetos de consulta, como cubos y dimensiones, cálculos y KPI, además de interacciones como los comportamientos de navegación y de obtención de detalles.

Para usar un modelo, se implementa en una instancia de Analysis Services que ejecuta bases de datos en un modo de servidor determinado, haciendo que los datos estén disponibles para los usuarios autorizados que se conectan a través de Excel u otras aplicaciones.

Puede instalar una instancia de Analysis Services en uno de estos tres modos de servidor:

- Como instancia tabular, ejecutando modelos tabulares.
- Como una instancia multidimensional y de minería de datos, ejecutando cubos OLAP y modelos de minería de datos (es el valor predeterminado).
- Como PowerPivot para SharePoint, ejecutando modelos de datos PowerPivot y de Excel en SharePoint (PowerPivot para SharePoint es un motor de datos de nivel intermedio que carga, consulta y actualiza modelos de datos hospedados en SharePoint).

El mismo motor de datos tiene tres formas de usarlo. Tenga en cuenta que los modos de servidor se establecen durante la instalación y no se pueden cambiar posteriormente. Debe instalar una nueva instancia si necesita otro modo diferente.

Analysis Services proporciona tres métodos para crear un modelo semántico de Business Intelligence: tabular, multidimensional y PowerPivot. Las soluciones tabulares usan construcciones de modelado relacional como tablas y relaciones para modelar los datos y el motor de análisis en memoria para almacenarlos y calcularlos. Las soluciones multidimensionales y de minería de datos usan construcciones de modelado OLAP (cubos y dimensiones) y almacenamiento MOLAP, ROLAP u HOLAP. PowerPivot es una solución BI de autoservicio que permite a los analistas de negocios generar un modelo de datos analíticos

en un libro de Excel mediante el complemento PowerPivot para Excel. PowerPivot usa también el motor de análisis en memoria en Excel y en SharePoint. Dado que usa las soluciones PowerPivot usan Excel tanto para el modelado de datos como para la representación, para implementar un libro en un servidor para el acceso a datos centralizados y controlados, se requiere SharePoint y Excel Services.

Las soluciones tabulares y multidimensionales se generan con SQL Server Data Tools y se han diseñado para proyectos BI corporativos que se ejecutan en una instancia independiente de Analysis Services. Ambas soluciones producen bases de datos analíticas de alto rendimiento que se integran con facilidad con informes de Reporting Services, Excel y otras aplicaciones BI desde aplicaciones de Microsoft y de otros fabricantes. Con todo, cada solución difiere en cómo se crea, se usa y se implementa. En este tema se exploran las diferencias, lo que le permite comparar e identificar la solución que mejor cumpla los requisitos del proyecto.

Dado que la tabular es la solución más reciente, puede que piense que migrar una solución MDX existente a un formato tabular lo más correcto, pero este no suele ser el caso. La solución tabular no reemplaza a la multidimensional y los dos formatos no son intercambiables. A menos que tenga una razón concreta para ello, no recompila una solución MDX existente si cumple ya las necesidades de su organización. Para los proyectos nuevos, considere el método tabular. Se agiliza el diseño, la prueba y la implementación; y funcionará mejor con las aplicaciones BI de autoservicio más recientes de Microsoft.

Los **modelos tabulares son bases de datos “en memoria”** de Analysis Services. Gracias a los algoritmos de compresión avanzados y al procesador de consultas multiproceso, el motor analítico en memoria xVelocity (VertiPaq) ofrece un acceso rápido a los objetos y los datos de los modelos tabulares para aplicaciones cliente de Informes como Microsoft Excel y Microsoft Power View.

Los modelos tabulares admiten el acceso a los datos mediante dos modos: modo de almacenamiento en caché y modo DirectQuery. En el modo de almacenamiento en caché, puede integrar datos de varios orígenes como bases de datos relacionales, fuentes de distribución de datos y archivos de texto planos. En el modo DirectQuery, puede omitir el modelo en memoria, lo que permite a las aplicaciones cliente consultar los datos directamente en el origen relacional (SQL Server).

Los modelos tabulares se crean en SQL Server Data Tools (SSDT) mediante las nuevas plantillas de proyectos de modelos tabulares. Puede importar datos de varios orígenes y, a continuación, enriquecer el modelo agregando relaciones, columnas calculadas, medidas, KPI y jerarquías. A continuación, los modelos se pueden implementar en una Instancia de Analysis Services que permite a las aplicaciones cliente de Informes conectarse con ellos. Los modelos implementados se pueden administrar en SQL Server Management Studio del mismo modo que los modelos multidimensionales. También se pueden crear particiones de los mismos para optimizar el procesamiento y protegerlos en el nivel de fila usando la seguridad basada en roles.

Una **solución multidimensional** de Analysis Services usa estructuras de cubos para analizar datos de negocio en varias dimensiones. El modo multidimensional es el modo de servidor predeterminado de Analysis Services. Incluye un motor de cálculo y consulta de datos OLAP, con los modos de almacenamiento MOLAP, ROLAP y HOLAP para equilibrar el rendimiento con los requisitos de datos escalables. El motor OLAP de Analysis Services es un servidor OLAP líder en el sector, que funciona con una amplia variedad de herramientas de BI. La mayoría de las implementaciones de Analysis Services se instalan como servidores OLAP clásicos.

La razón principal para generar un modelo multidimensional de Analysis Services es lograr el rendimiento rápido de consultas ad hoc en los datos empresariales. Un modelo multidimensional se compone de cubos y dimensiones que se pueden anotar y ampliar para admitir construcciones de consultas complejas. Los desarrolladores de BI crean cubos para admitir tiempos de respuesta rápida y para proporcionar un único origen de datos para Informes empresariales. Debido a la mayor importancia de business intelligence en todos los niveles de una organización, el hecho de tener un solo origen de datos analíticos garantiza que las discrepancias se mantienen al mínimo, si no se eliminan por completo.

Otra ventaja importante del uso de las bases de datos multidimensionales de Analysis Services es la integración con las herramientas de informes BI utilizadas habitualmente, como Excel, Reporting Services y PerformancePoint, así como las aplicaciones personalizadas y las soluciones de terceros.

Analysis Services también proporciona una plataforma integrada para las soluciones que incorporan la **minería de datos**. Puede usar datos relacionales o de cubo para crear soluciones de Business Intelligence con análisis predictivos.

La minería de datos usa principios estadísticos contrastados para detectar patrones en los datos, ayudándole a tomar decisiones inteligentes sobre problemas complejos. La aplicación de los algoritmos de minería de datos de Analysis Services a los datos le permitirá predecir tendencias, identificar patrones, crear reglas y recomendaciones, analizar la secuencia de eventos en conjuntos de datos complejos y obtener nuevos puntos de vista.

En SQL Server 2014, la minería de datos es eficaz y accesible, y está integrada con las herramientas preferidas de los usuarios para el análisis y la creación de informes. Vea los vínculos de esta sección para obtener toda la información sobre la minería de datos que necesita para empezar.

SQL Server proporciona las siguientes características para las soluciones integradas de minería de datos:

- **Varios orígenes de datos:** no es necesario crear un almacenamiento de datos o un cubo OLAP para realizar la minería de datos. Puede usar datos tabulares de proveedores externos, hojas de cálculo e incluso archivos de texto. También puede realizar fácilmente la minería de los cubos OLAP creados en Analysis Services. Sin embargo, no puede usar datos de una base de datos en memoria.
- **Limpieza de los datos integrados, administración de datos y ETL:** Data Quality Services proporciona herramientas avanzadas para la generación de perfiles y la limpieza de datos. Se puede usar Integration Services para generar procesos ETL de limpieza de datos, y también para tareas de creación, procesamiento, entrenamiento y actualización de modelos.
- **Varios algoritmos personalizables:** además de proporcionar algoritmos como la agrupación en clusters, las redes neuronales y los árboles de decisión, la plataforma le permite desarrollar sus propios complementos con algoritmos personalizados.
- **Infraestructura de prueba del modelo:** pruebe los modelos y los conjuntos de datos usando herramientas estadísticas tan importantes

como la validación cruzada, las matrices de clasificación, los gráficos de mejora respecto al modelo predictivo y los gráficos de dispersión. Cree y administre fácilmente conjuntos de prueba y entrenamiento.

- **Consultas y obtención de detalles:** cree consultas de predicción, recupere patrones y estadísticas de modelos, y obtenga información detallada de los datos de los casos.

- **Herramientas de cliente:** además de los estudios de desarrollo y diseño proporcionados por SQL Server, puede usar los Complementos de minería de datos para Excel para crear, consultar y examinar los modelos. O bien crear clientes personalizados, incluidos servicios web.

- **Compatibilidad con el lenguaje de scripting y API administrada:** todos los objetos de minería de datos son completamente programables. El scripting es posible mediante MDX, XMLA o las extensiones de PowerShell para Analysis Services. Use el lenguaje DMX (Extensiones de minería de datos) para crear rápidamente consultas y Scripts.

- **Seguridad e implementación:** proporciona seguridad basada en roles a través de Analysis Services, incluyendo permisos distintos para la obtención de detalles del modelo y los datos de la estructura. Fácil implementación de modelos en otros servidores, de forma que los usuarios puedan tener acceso a los patrones o realizar predicciones.

CAPÍTULO 8

HERRAMIENTAS DE BIG DATA EN SAS

HADOOP Y BIG DATA EN SAS

SAS soporta implementaciones de Big Data, incluyendo Hadoop. Independientemente de cómo usar la tecnología, cada proyecto debe pasar por un ciclo iterativo y mejora continua. Y eso incluye preparación de datos y gestión, visualización de datos y exploración, desarrollo de modelos, despliegue y supervisión.

SAS se centra en el análisis, no en el almacenamiento. Por ello ofrece un enfoque flexible para elegir proveedores de hardware y bases de datos. SAS trabaja para implementar la combinación adecuada de tecnologías, incluyendo la habilidad para desplegar Hadoop con otras tecnologías de almacén de datos. Hadoop es una plataforma de datos integrada en SAS como un componente esencial de proceso analítico y de plataforma de próxima generación. La Figura 8-1 muestra la integración entre componentes SAS y Hadoop.

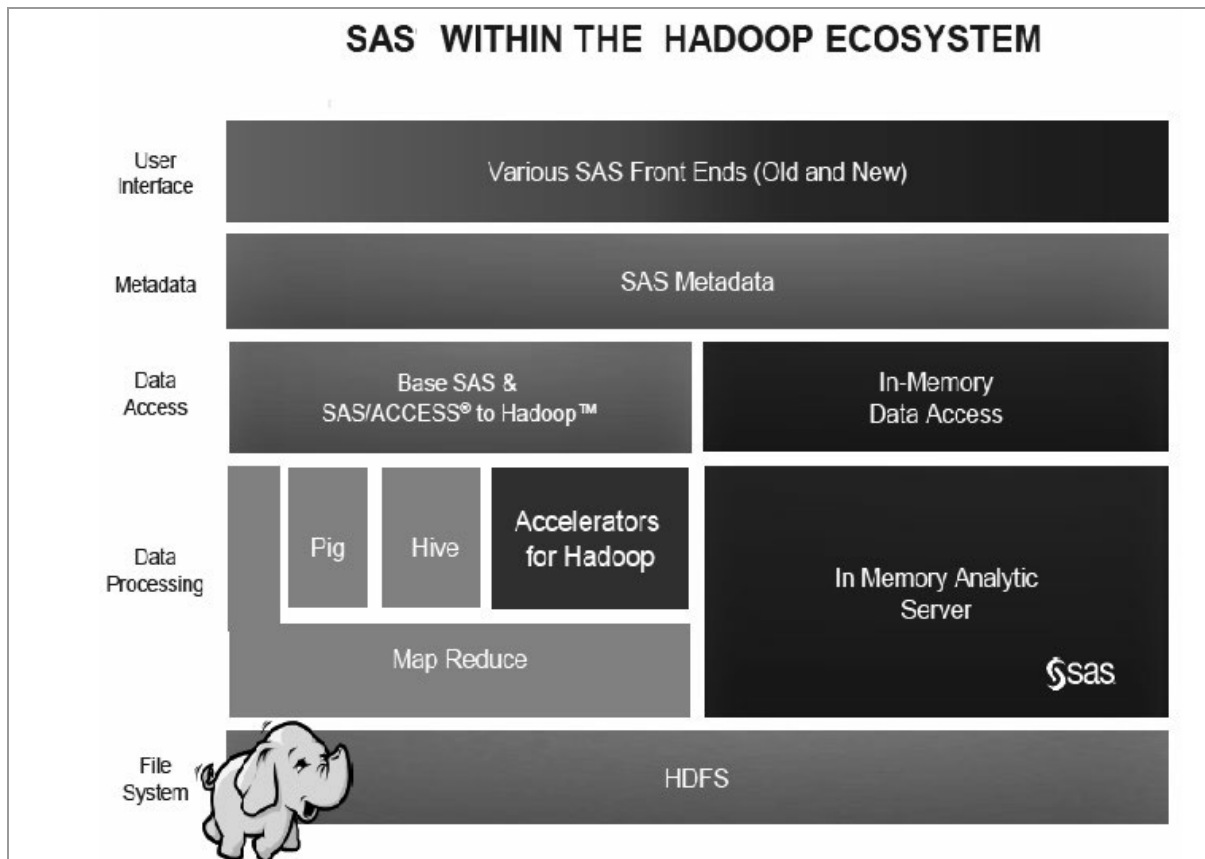


Figura 8-1

Observamos que SAS utiliza varias interfaces de usuario para tener acceso a la información y la metainformación. El acceso a datos de Hadoop se realiza mediante SAS Base y SAS/ACCESS to Hadoop aportando capacidades de almacenamiento en memoria In Memory OLTP típicas del Big Data. Posteriormente realiza el procesamiento de los datos utilizando Map Reduce, Pig, Hive y aceleradores para Hadoop, perfeccionándolo con los análisis derivados de las herramientas analíticas In Memory. También utiliza el sistema de ficheros HDFS. La Figura 8-2 se amplía la integración entre los componentes SAS y Hadoop.

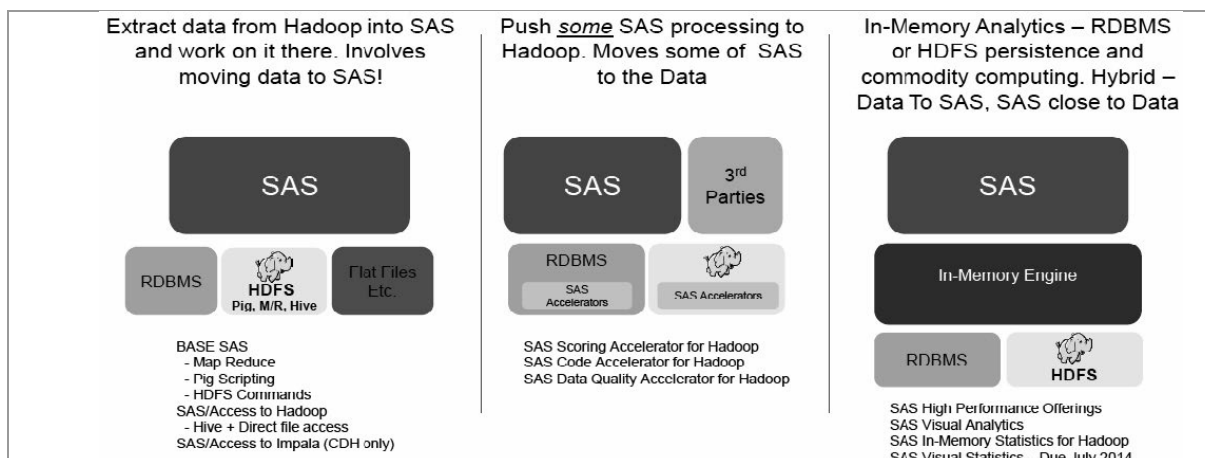


Figura 8-2

SAS, HADOOP Y EL PROCESO ANALÍTICO

A través de SAS y Hadoop es posible trabajar en todas las etapas del proceso analítico: identificar/formular, preparación de datos, exploración de datos, transformar y seleccionar, construir el modelo, validar el modelo, implementar el modelo y evaluar el modelo (Figura 8-3).

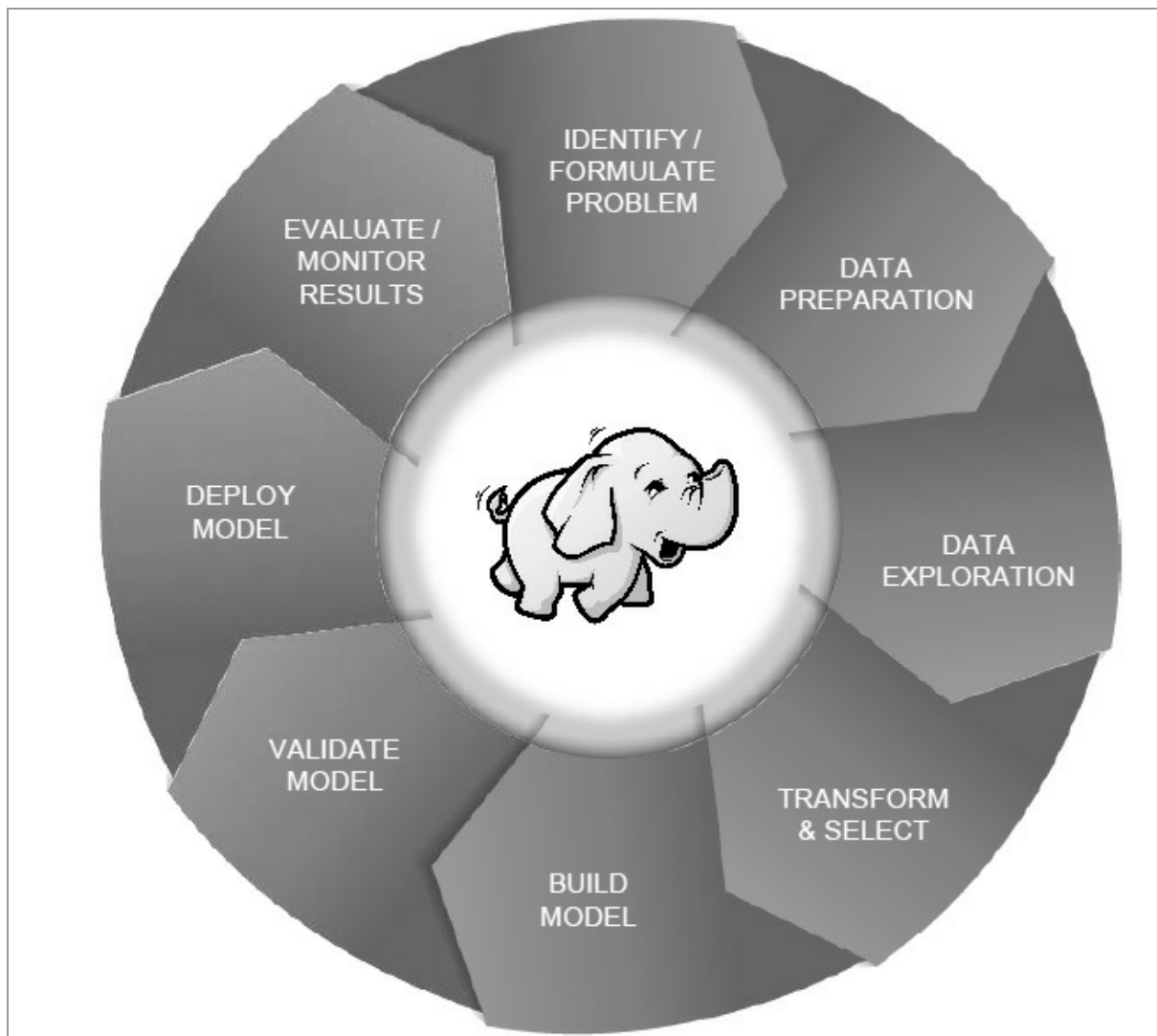


Figura 8-3

Algunas compañías están utilizando Hadoop como un componente esencial de un proceso analítico y de plataforma de última generación. A veces por encima de una plataforma de datos y a veces en solo un amplio despliegue de Hadoop.

Los productos de SAS implementan el ciclo de vida completo

alrededor de Hadoop para que funcione en todas las etapas del proceso analítico (Figura 8-4).

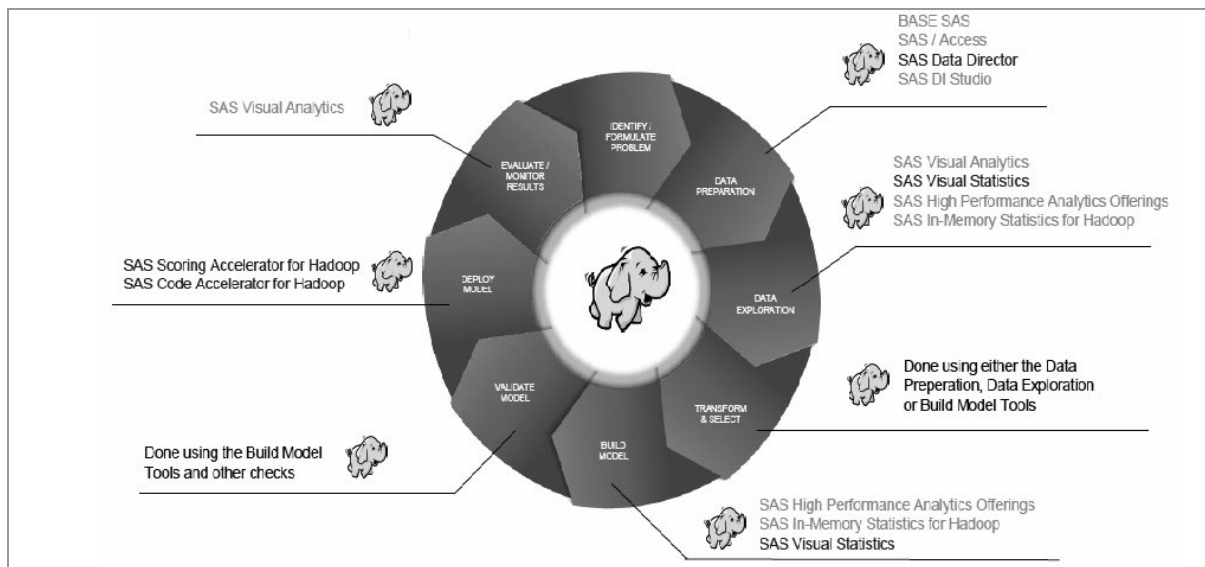


Figura 8-4

BIG DATA Y SOLUCIONES HADOOP DE SAS

Acceso y administración de datos de Hadoop

- **SAS/ACCESS Interface to Hadoop.** Permite obtener conectividad entre SAS y Hadoop, vía Hive.
- **SAS Data Management.** Permite asegurar la integración fiable de datos de cualquier fuente.
- **SAS Federation Server.** Permite centralizar y optimizar vistas de los datos negocio.
- **Software SAS Base.** Utilizar un lenguaje de programación flexible para acceso a datos de gran alcance, transformación y presentación de informes.

Explorar, visualizar y tratar datos científicos

- **SAS Visual Analytics.** Realiza exploración visual de los datos para descubrir nuevos patrones y publicar informes en la web y dispositivos móviles.
- **SAS In-Memory Statistics for Hadoop.** Permite enriquecer a SAS Visual Analytics con análisis estadísticos profundos para encontrar ideas con los datos utilizando Hadoop en un entorno que se mueve rápidamente a través de cada fase del ciclo de vida analítico.

La Figura 8-5 muestra herramientas SAS para trabajo en memoria en Hadoop.

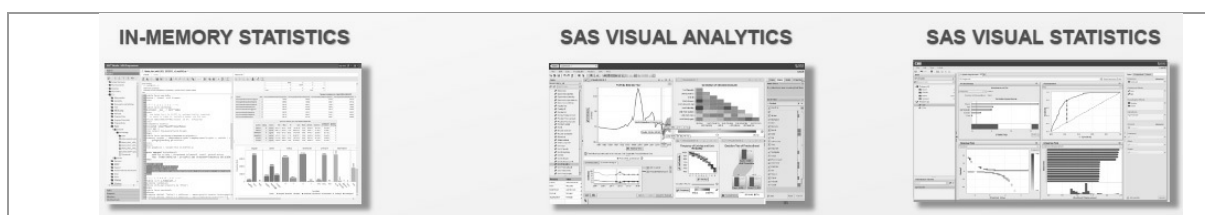


Figura 8-5

Analizar y modelizar

- **SAS Visual Statistics.** Permite crear y modificar modelos predictivos más rápido que nunca usando una interfaz visual y procesamiento en la memoria.
- **SAS High-Performance Data Mining.** Permite analizar rápidamente grandes datos (Big Data) para obtener información más precisa y tomar decisiones empresariales oportunas.
- **SAS High-Performance Text Mining.** Rápidamente descubre categorías y temas en grandes volúmenes de datos no estructurados.
- **SAS High-Performance Statistics.** Permite solucionar problemas de datos grandes (Big Data) con software estadístico potente.
- **SAS High-Performance Optimization.** Permite modelar y resolver problemas de optimización muy grandes o cuyas otras características hacen difíciles las soluciones.

La Figura 8-6 muestra herramientas de SAS para trabajar en minería de datos y minería de textos en Hadoop.

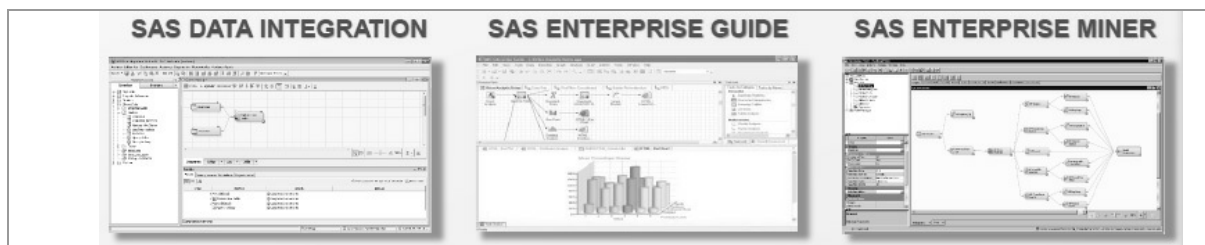


Figura 8-6

Implementar e integrar

- **SAS Scoring Accelerator.** Permite automatizar los procesos de los modelos de scoring dentro de la base de datos para mejorar el rendimiento de los modelos y la obtención de puntuación para obtener resultados más rápidos.
- **SAS Event Stream Processing Engine.** Posibilita la elaboración de analíticas de datos en tiempo real.

SAS/ACCESS INTERFACE TO HADOOP

Una de las ventajas de Hadoop es que presenta una arquitectura de procesamiento distribuido que aporta una escalabilidad excepcional para resolver una amplia gama de problemas de conectividad entre SAS y Hadoop, vía Hive (Figura 8-7). Precisamente esta conectividad óptima es la finalidad de SAS/ACCESS Interface to Hadoop.

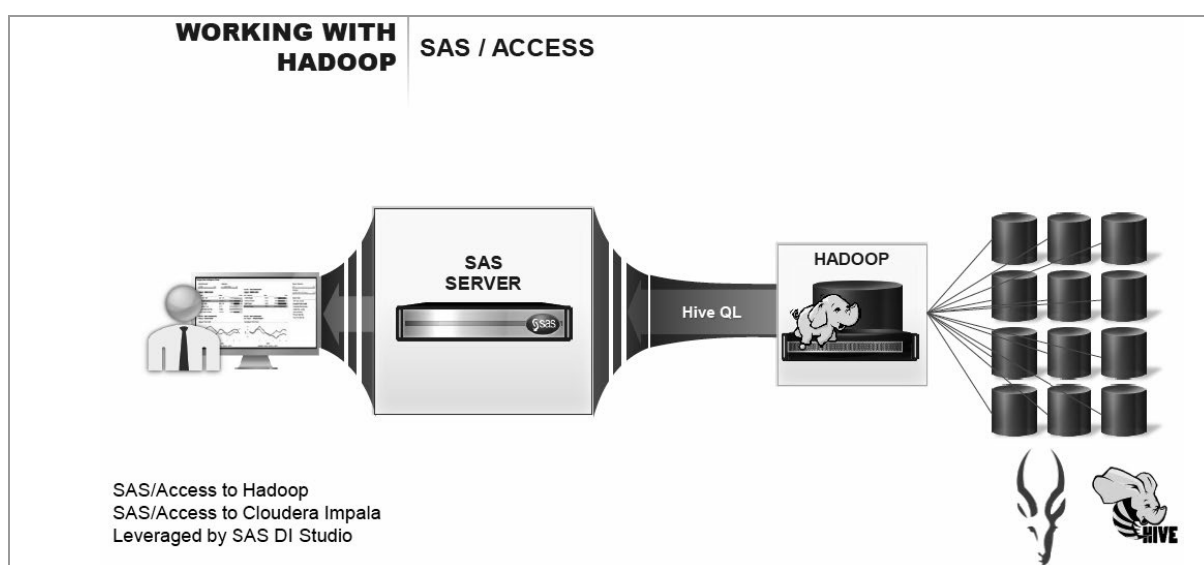


Figura 8-7

A través de la interfaz SAS/Access to Hadoop es posible obtener acceso de lectura/escritura transparente de más de 60 fuentes de datos, incluyendo bases de datos relacionales y no relacionales, ficheros informáticos y herramientas de almacén de datos. SAS Data Surveyor proporciona motores especializados específicamente para acceder a las aplicaciones empresariales. Puede integrar los datos almacenados en Hadoop con otros datos y utilizar fácilmente todos los datos de su empresa con SAS Data Analytics y utilizar fácilmente todos los datos de su empresa con SAS Analytics.

SAS crea los flujos de datos que combinan el procesamiento de Hadoop con el procesamiento con SAS, obteniendo un desempeño óptimo utilizando el mejor entorno de ejecución. SAS/ACCESS también soporta carga de alto rendimiento de Hadoop, proporcionando la capacidad para

cargar rápidamente datos en Hadoop de una gran variedad de fuentes, incluyendo archivos planos y otras fuentes de bases de datos relacionales.

En SAS/Acces interfaz No Hadoop o HiveQL se requiere experiencia. Las consultas apropiadas de HiveQL son generadas a partir de la interfaz y pasadas a la fuente de datos para su ejecución. SAS permite un enfoque transparente de datos fuentes que aparecen como si fuesen nativos de SAS, lo que facilita la interacción con los procedimientos de SAS, el paso data de SAS y las soluciones SAS. También se puede interactuar directamente con el origen de datos utilizando las capacidades de fuente de datos nativos.

Esta solución requiere una formación mínima y transferencia de conocimientos básica. Los usuarios suelen utilizar informes de SAS y capacidades analíticas que perfectamente pueden acceder a datos de terceros. Los usuarios técnicos pueden acceder a los datos de forma transparente, o interactuar directamente con los datos almacenados en Hadoop usando HiveQL.

Características

Acceso de datos transparente y sin problemas

- Amplia cobertura de plataformas y fuentes de datos que permite integrar datos procedentes de toda la organización.
- Incluye la opción de acceso a datos sin conocimiento detallado de Hadoop o HiveQL. Las consultas apropiadas de HiveQL son generadas a partir de la interfaz y enviadas a la fuente de datos para su ejecución.
- Incluye la opción de usar HiveQL nativo si así se desea.
- Tiene la capacidad de crear consultas federadas que abarcan múltiples fuentes de datos. Las consultas aparecen como un objeto de base de datos que puede utilizarse con todas las soluciones de SAS.

Soporte de idiomas de consulta flexible

- Una opción de acceso transparente a datos con un mínimo conocimiento de los datos o el HiveQL necesario para el acceso.
- Una opción personalizada de SQL que permite a los usuarios crear

sus propias declaraciones HiveQL o modificar HiveQL generado automáticamente.

- Funciones específicas de base de datos que permiten que todas las declaraciones HiveQL apropiadas sean procesados directamente dentro de Hadoop, proporcionando el mejor rendimiento posible.

Características de rendimiento

- Lectura y escritura de datos altamente optimizados con almacenamiento en búfer, etc.
- Soporta el interfaz de tablas externas de Hive, incluyendo archivos HDFS externos.
- Tiene la capacidad de controlar las capacidades de procesamiento de JOINS.
- Carga datos en HDFS muy rápidas con grandes velocidades de transmisión.

Integración de metadatos

- Los metadatos de Hadoop pueden mantenerse con precisión dentro del repositorio de metadatos de SAS.
- Los trabajos con los datos se registran mediante el servidor de metadatos SAS, de modo que pueden ser utilizados en las soluciones SAS.

Optimización

- Proporciona soporte para las opciones de almacenamiento nativo, incluidas las tablas externas.
- Traduce los datos nativos Hadoop al tipo de datos SAS apropiado para su procesamiento con SAS. Proporciona compatibilidad con idioma UTF-8.

Requisitos del sistema

- HP/UXen Itanium: Ili3
- IBM AIX on POWER architectures: 6.1 y 7.1

- Linux x 64 (64 bits): Novell SuSE 10 y 11; RHEL 5 y 6; Oracle Linux 5.5 y 6
- Microsoft Windows (32 bits): Windows XP Professional, Windows Vista, Windows 7, familia de Windows Server 2003, y familia de Windows Server 2008
- Microsoft Windows x 64 (64 bits): Windows XP Professional x 64, Windows Vista para x 64, Windows 7 para x 64, la familia de Windows Server 2003 x 64, la familia de Windows Server 2008 x 64
- Solaris en SPARC: actualización de la versión 10 8
- Solaris x 64 (x 64-86): actualización de la versión 10 8
- Ediciones de Windows Vista compatibles son: Enterprise, Ultimate y Business.
- Ediciones de Windows 7 compatibles son: Enterprise, Ultimate y profesional.

Software de SAS requerido

- SAS Base es necesario para la instalación del software SAS/ACCES Interface to Hadoop.
- HiveServer2 y Kerberos se admiten en SAS 9.4.

SAS DATA MANAGEMENT

SAS Data Management es una solución líder en la industria construida sobre una plataforma de calidad de datos que le ayudará a mejorar, integrar y gobernar sus datos.

No importa cuál sea el almacenamiento de los datos. SAS ayuda a acceder a los datos que se necesitan desde sistemas heredados de Hadoop. SAS Data Management permite crear reglas de gestión de datos y reutilizarlas, proporcionando un método estándar y repetible para mejorar la integración de datos sin costo adicional.

La tecnología de SAS Data Management está realmente integrada, lo que significa que no está obligado a trabajar con una solución que se improvisó. Todas las componentes son parte de la misma arquitectura.

SAS Data Management permite a los usuarios de negocio en la organización actualizar datos, ajustar los procesos y analizar resultados, promoviendo la colaboración y liberando recursos para otros proyectos.

SAS Data Management posibilita un acceso rápido y fácil a datos Hadoop, permitiendo añadir datos grandes (Big Data) a los procesos de IT existentes. Y con las transformaciones de MapReduce, Pig y Hive, se puede administrar datos grandes (Big Data).

SAS Data Management presenta una consola centralizada (Figura 8-8), puestos de control y procesos (Figura 8-9), enlace trabajos (Figura 8-10), propagación de la información (Figura 8-11), depuración integrada (Figura 8-12) y estadísticas de tiempo de ejecución (Figura 8-13).

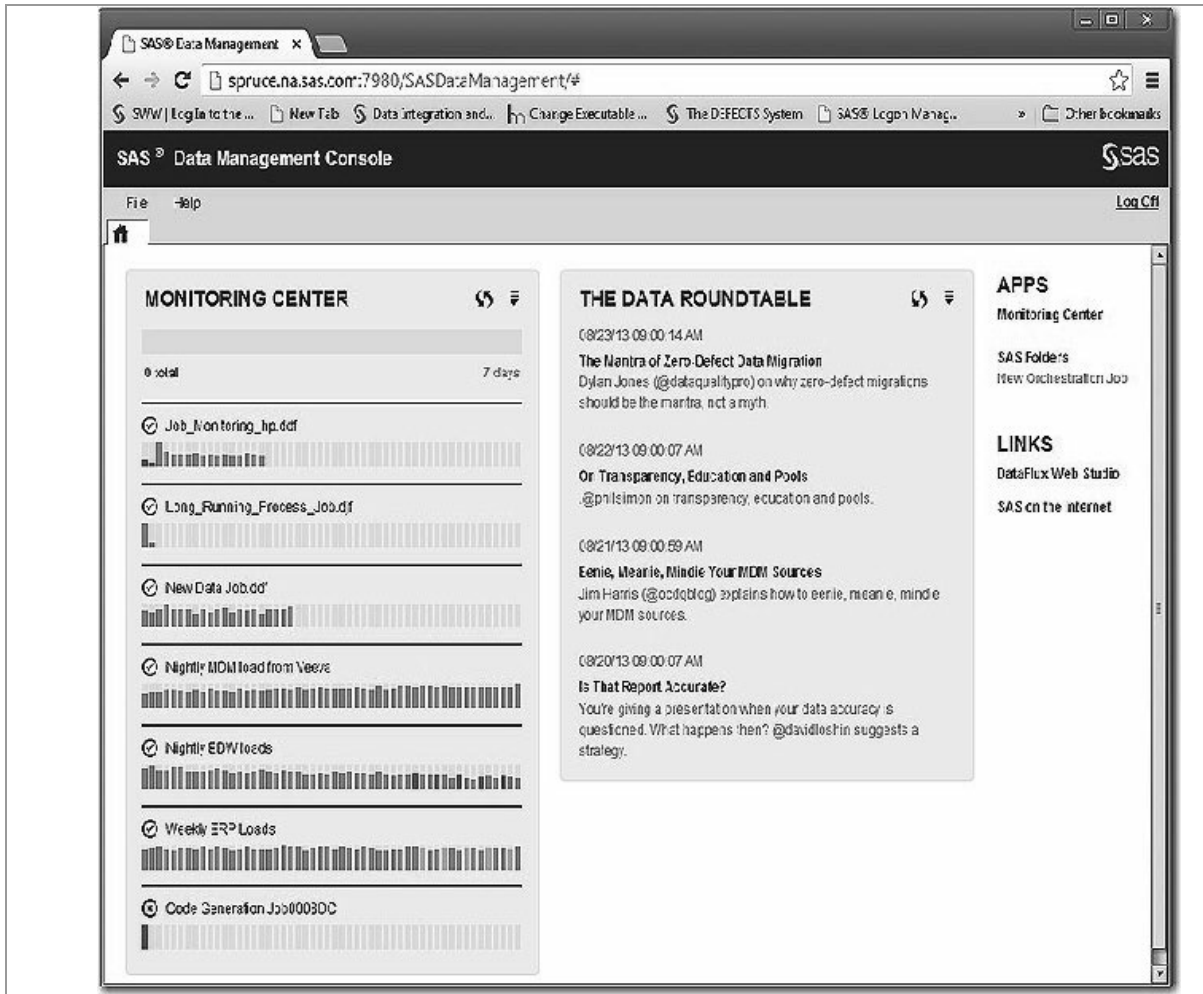


Figura 8-8

The screenshot shows the SAS Environment Manager Monitoring Center. It features a table with columns for Job Name, Job Type, Platform, Status, % Complete, Start Time, End Time, Run Time, 20 Run Mean, Trend, and User ID. The table lists several 'Copy of ExpressionJob.dcf' jobs, all of which are 'Completed' with a 100% completion rate. The jobs were executed on the '476664 na SAS.com' platform between 8/27/13 1:55:33 PM and 8/27/13 10:24:19 AM.

Job Name	Job Type	Platform	Status	% Complete	Start Time	End Time	Run Time	20 Run Mean	Trend	User ID
Copy of ExpressionJob.dcf	DM Process Job	476664 na SAS.com	Completed	100	8/27/13 1:55:33 PM	8/27/13 1:55:33 PM	2:049s	2:036s	→	PUBLIC
Copy of ExpressionJob.dcf	DM Process Job	476664 na SAS.com	Completed	100	8/27/13 11:38:55 AM	8/27/13 11:38:57 AM	2:04s	2:036s		PUBLIC
Copy of ExpressionJob.dcf	DM Process Job	476664 na SAS.com	Completed	100	8/27/13 11:24:24 AM	8/27/13 11:24:26 AM	2:056s			PUBLIC
ExpressionJob.dcf	DM Process Job	476664 na SAS.com	Completed	100	8/27/13 11:21:04 AM	8/27/13 11:21:06 AM	2:043s	2:003s	→	PUBLIC
ExpressionJob.dcf	DM Process Job	476664 na SAS.com	Completed	100	8/27/13 11:07:14 AM	8/27/13 11:07:16 AM	2:034s	2:071s	→	PUBLIC
ExpressionJob.dcf	DM Process Job	476664 na SAS.com	Completed	100	8/27/13 10:44:09 AM	8/27/13 10:44:11 AM	2:034s	2:003s	→	PUBLIC
ExpressionJob.dcf	DM Process Job	476664 na SAS.com	Completed	100	8/27/13 10:28:05 AM	8/27/13 10:28:07 AM	2:029s	2:11s	→	PUBLIC
ExpressionJob.dcf	DM Process Job	476664 na SAS.com	Completed	100	8/27/13 10:27:50 AM	8/27/13 10:27:52 AM	2:044s	2:176s		PUBLIC
ExpressionJob.dcf	DM Process Job	476664 na SAS.com	Completed	100	8/27/13 10:24:17 AM	8/27/13 10:24:19 AM	2:176s			PUBLIC

Figura 8-9

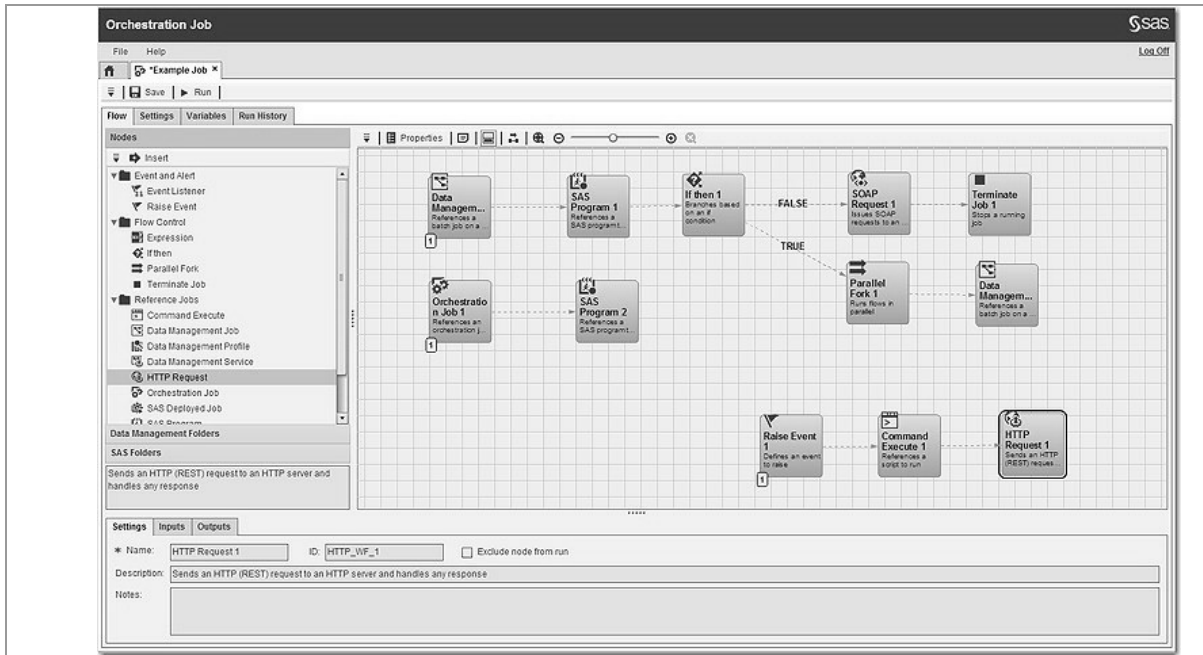


Figura 8-10

The screenshot shows the 'Mappings' window in SAS. It has tabs for 'Status', 'Warnings and Errors', 'Statistics', and 'Control Flow'. The 'Control Flow' tab is active. Below the tabs is a toolbar with various icons. The main area is divided into two tables: 'Source table' and 'Target table'. The 'Source table' has columns for '#', 'Column', and 'Type'. The 'Target table' has columns for '#', 'Column', 'Expression', and 'Type'. Arrows indicate the mapping between columns in the source and target tables.

#	Column	Type	#	Column	Expression	Type
18	customer_geo_id	Numeric	18	customer_geo_id	TRIM(LEFT(PUT(customer_g...	Character
17	customer_rk	Numeric	17	customer_rk		Numeric
22	date_id	Numeric	22	date_id		Numeric
6	discount_type_cd	Character	6	discount_type_cd	INPUT(discount_type_cd,BE...	Numeric
20	effective_from_dtm	Numeric	20	effective_from_...		Numeric
21	effective_to_dtm	Numeric	21	effective_to_dtm		Numeric
19	employee_rk	Numeric	19	employee_rk		Numeric

At the bottom of the window, there is a status bar that says '1 Error'.

Figura 8-11

Retail Data * (Read-Only)

Up Run Stop [Icons]

BI_ORDER_F... (BI_ORDER_...)
bi order fact

1 Rank variables
This transform will rank variable...

2 Filter results
This transform will filter...

3 Analyze

BI_ORGANIZ... (BI_ORGANI...)

This job will append...

Diagram Code Log Output

Details

Status Warnings and Errors Statistics Control Flow

Clear All

Node	Name	Status	Details
0	Precode	Completed successfully	
1	Rank variables	Completed successfully	
2	Filter results	Error	
3	Analyze	In progress	

Figura 8-12

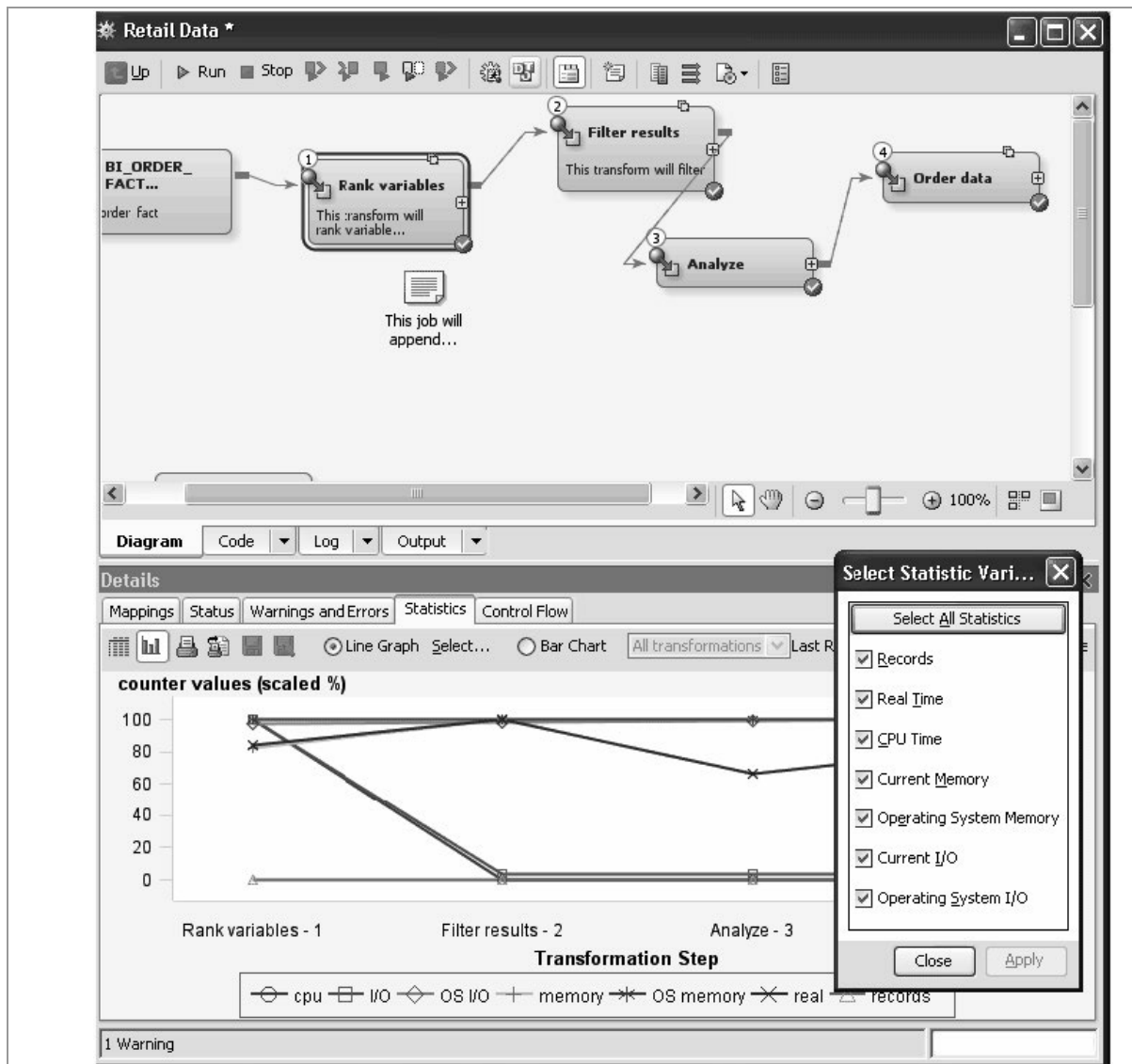


Figura 8-13

Características

Entorno de desarrollo de integración de datos

- Un entorno fácil de usar, basado en GUI con un sistema intuitivo de ventanas configurables para administrar los procesos. La funcionalidad de arrastrar y soltar elimina la necesidad de programación.
- Asistentes de acceso a sistemas de origen, creando estructuras objetivo, funciones de importación y exportación metadatos y ejecución y construcción de flujos de procesos ETL.

- Presenta vistas de árbol de metadatos personalizados permitiendo mostrar, visualizar y comprender metadatos.
- Un GUI dedicado para perfiles de datos facilita reparar problemas en el sistema fuente conservando las reglas de negocio para su uso en otros procesos de gestión de datos.
- Proceso interactivo de depuración y prueba de puestos de trabajo durante el desarrollo y compatible con pleno acceso a los registros.
- Los procesos de auditoría y check-in/check-out permite a los diseñadores ver qué puestos de trabajo registraron cambios, cuándo y por quién.
- Capacidad para distribuir las tareas de integración de datos a través de cualquier plataforma y conectar prácticamente cualquier almacén de datos de origen o destino.
- Integración con terceros mejorando las funciones de control de fuentes.
- Capacidades mejoradas de importación de código SAS que proporcionan a los usuarios una forma fácil de importar sus trabajos de SAS y código.
- Opciones de implementación de trabajo en línea de comandos para implementar trabajos individuales y múltiples.

Diseñador de procesos integrados

- Crear y editar los procesos de gestión de datos con un diseñador visual, end-to- end de eventos.
- Control de la ejecución de integración de datos, trabajos de calidad y procedimientos almacenados de SAS.
- Ejecución condicional de trabajos utilizado la cláusula lógica IF THEN.
- Ejecución de trabajos y procesos en paralelo.
- Publicar trabajos de entradas y salidas parametrizadas.
- Escucha de eventos internos y externos así como condicional para provocar eventos.
- Ejecutar comandos de nivel OS externos tales como llamada scripts

de shell.

- Llamar a servicios Web SOAP y otros.
- Listar y abrir las versiones antiguas de puestos de trabajo (en modo de solo lectura) y las versiones históricas actuales con versiones incorporadas.
- Proporcionar soporte completo para promoción y migración de puestos de trabajo.
- Utilizar lenguajes comunes para implementar los trabajos por lotes de manera automatizada con el despliegue de trabajo automatizado.

Acceso de datos y conectividad avanzada

- Proporciona conectividad en proceso por lotes o en tiempo real con gran cantidad de fuentes de datos y plataformas.
- Los motores de acceso de datos están disponibles para las aplicaciones empresariales, bases de datos no relacionales, RDBMS, herramientas de almacén de datos, formatos de archivo de PC, etc.
- Carga optimizada de Teradata, Oracle y DB2.
- Lectura y escritura de archivo disponible para Hadoop File System (HDFS) y soporte para de Hadoop MapReduce, Pig y Hive así como para Hortonworks.
- La compartición de metadatos proporciona la definición de datos consistentes a través de todas las fuentes de datos.
- Métodos de acceso nativo ofrecen un rendimiento superior, reducen el movimiento de datos y reducen la necesidad de codificación personalizada.
- Soporte para middleware orientado a mensajes, incluyendo WebSphere MQ de IBM, MSMQ de Microsoft, Java Message Service (JMS) y TIBCO Rendezvous. Soporte para datos no estructurados y semiestructurados y para analizar y procesar archivos.
- Acceso a datos estáticos y de transmisión para enviar y recibir vía Web Services.
- Soporte para bases de datos de MPP ampliado: datos Aster nCluster, Greenplum Pivotal y Sybase IQ, más soporte para utilidades

de carga de ELT.

- Soporte nativo para el procesamiento basado en SQL.
- Conectividad mejorada con datos Aster, Greenplum Pivotal, Hadoop y Sybase IQ con capacidad de procesamiento para las bases de datos.

Administración de metadatos consistente

- Metadatos documentados a lo largo de las transformaciones y los procesos de integración y disponibles para su reutilización inmediata.
- Mapas de metadatos sofisticados propagan las definiciones de las columnas de fuentes a los objetivos y crean tablas automatizadas e inteligentes.
- La búsqueda de metadatos permite la localización rápida de los componentes deseados.
- Análisis del impacto para evaluar el alcance e impacto de hacer cambios a los objetos existentes tales como columnas, tablas y trabajos de proceso antes de que ocurran.
- Capacidad para determinar la ruta, los procesos y transformaciones para producir la información resultante.
- Análisis de impacto inverso, fundamental para validar las dependencias y ayudar a crear confianza en los datos del usuario.
- Análisis de cambio para metadatos, comparación, análisis y propagación selectiva.
- Soporte de múltiples usuarios con colaboración objeto check-in y check out.
- Promoción y replicación de metadatos a través de entornos de desarrollo, prueba y producción.
- Trabajo con metadatos guiado por el asistente de importación y exportación así como la normalización de las columnas.
- Flexibilidad de implementación basados en metadatos para que trabajos de proceso puedan ser desplegados para la ejecución por lotes, como reutilizables, almacenan los procesos o como servicios web.

Fundación de calidad de datos

- La calidad de los datos está incrustada en el proceso por lotes, casi en tiempo real de los procesos.
- Limpieza de datos proporcionada en lenguas nativas con conciencia del lenguaje específico y localizaciones para más de 38 regiones del mundo.
- Las fundones de calidad de datos están disponibles en entornos operacionales y de reporting (transacción y por lotes).
- Una interfaz interactiva le permite analizar datos operacionales del perfil para identificar los datos incompletos, inexactos o ambiguos.
- Reglas de negocio de calidad de datos personalizables y reutilizables son accesibles directamente dentro de los flujos de trabajo de proceso.
- Reglas de normalización ajustadas a datos estándares corporativos. Se pueden crear también reglas personalizadas para situaciones especiales.
- Los metadatos construidos y compartidos a través de todo el proceso proporcionan un rastro preciso de acciones aplicado a los datos limpios.
- Se puede obtener valor añadido a los datos existentes mediante la generación y anexo de direcciones postales, geocodificación, datos demográficos o datos de otras fuentes de información.
- Los administradores de datos pueden perfilar los datos operativos y vigilar las actividades de datos en curso con una interfaz gráfica interactiva diseñada específicamente para sus necesidades.
- Proceso simple para institucionalizar reglas de negocio de calidad de datos. Se pueden aplicar reglas básicas o complejas para validar los datos según los requerimientos específicos del negocio de un determinado proceso, proyecto u organización. Estas reglas podrán aplicarse en modo por lotes o como una transacción en tiempo real en procesos de limpieza.
- El monitoreo de la calidad de datos permite examinar continuamente los datos en tiempo real y con el tiempo para descubrir

cuándo la calidad cae por debajo de los límites aceptables.

- Se pueden emitir alertas cuando hay una necesidad de medidas correctivas.

Extracción, transformación y carga (ETL) y extracción y transformación (ELT).

- Se dispone de una interfaz de usuario de transformación de gran alcance, fácil de usar que soporta colaboración, reutilización de metadatos comunes y procesos.
- Posibilidad de ofrecer capacidades de ELT, incluidas las tablas de crear, unir, insertar filas, eliminar filas, actualizar filas, merge y SQL conjunto.
- La adquisición de datos individuales de múltiples fuentes, la transformación, limpieza y carga permiten la fácil creación de almacenes de datos, puestos de datos o almacenes de datos BI y analítica.
- Los metadatos son capturados y documentados a lo largo de la integración de datos y los procesos de transformación y están disponibles para su reutilización inmediata.
- Las transformaciones se pueden ejecutar en cualquier plataforma con cualquier fuente de datos.
- Más de 300 transformaciones a nivel de tabla y columna predefinidos.
- Se pueden utilizar transformaciones analíticas, incluyendo las correlaciones y frecuencias, análisis de distribuciones y las estadísticas de Resumen.
- Asistente de transformación o Java Plug-In con diseño de plantillas para posibilitar fácilmente generar transformaciones reutilizables y repetibles que son rastreadas y registradas en metadatos.
- Los procesos de transformación, accesibles a través de salidas personalizadas, colas de mensajes y servicios web son reutilizables en diferentes proyectos y entornos.
- Las transformaciones pueden ser ejecutadas de forma Interactiva y programadas para ejecutarse en lote o basándose en eventos que

desencadenan la ejecución.

- Entorno adecuado para publicar Información en archivos, un canal editorial, correo electrónico o diversos mddleware de Message Queue Server.
- Se facilitan los procesos de añadir y actualizar durante la carga.
- Se optimizan las técnicas de carga con opciones seleccionares por el usuario.
- Las transformaciones generan automáticamente código SAS de alto rendimiento que está diseñado para el procesamiento rápido y eficiente.
- La transformación de las tablas comparan fuentes de datos y detectan cambios en los datos.
- Está habilitada la capacidad de llamar a servicios Web SOAP u otros.

Federación de datos

- Acceso virtual a las estructuras de base de datos, aplicaciones empresariales, archivos heredados para mainframe, texto, XML, colas de mensajes y otras fuentes.
- Capacidad para unirse a datos a través de fuentes de datos para acceso en tiempo real y análisis.
- Acceso instantáneo a una vista en tiempo real de los datos utilizando el visor de datos integrado.
- Se proporciona optimización de consultas tanto automáticamente como parte de las solicitudes DBMS y manualmente dentro del editor SQL avanzado. Pueden utilizarse fuentes de datos homogéneas y heterogéneas.

Gestión de datos maestros

- Funciones de búsqueda mejorada permiten realizar la búsqueda por tipo, nombre, fecha u otras palabras clave, por carpetas u otras opciones y guardar búsquedas para uso futuro.
- Soporte para descripciones semánticas de orígenes de datos de entrada y salida para identificar de forma exclusiva cada instancia de

un elemento de negocio (cliente, producto, cuenta, etc.).

- Herramientas de transformación de gran alcance y procesos de calidad de datos integrados mejoran la calidad de los datos maestros.
- Sofisticada tecnología fuzzy implementa técnicas de matching y agrupamiento que permiten validar y consolidar los registros maestros en grupos de datos identificables.
- Monitoreo de datos en tiempo real para ver y controlar la integridad de los datos en el tiempo.
- Las fuentes de datos pueden llegar en una sola transacción o en cientos de transacciones al mismo tiempo.
- Los conjuntos de datos se pueden procesar en una sola pasada de los datos de origen.

Manejo de datos

- Manejo de datos mejorado basado en la Wweb.
- Un glosario de datos de negocios integrado permite ser organizados jerárquicamente y relacionarse con metadatos técnicos tales como tablas y procesos de gestión de datos.
- Capacidades de datos extensos basados en web y monitoreo para informar y realizar procesos de mantenimiento.

Migración y sincronización

- Capacidad de migrar o sincronizar datos entre las estructuras de base de datos, aplicaciones empresariales, archivos heredados para mainframe, texto, XML, colas de mensajes y anfitriones de otras fuentes.
- Acceso a fuentes y destinos basados en metadatos.
- Disponibilidad de una amplia biblioteca de transformaciones predefinidas que puede ser extendida y compartida con otros procesos de integración.
- Reglas de negocio de calidad de datos incrustados, reutilizables para limpiar datos, sincronizados o replicados.
- Programador opcional integrado permite que los cambios realizados

en uno o más sistemas puedan propagarse a otros sistemas de forma programada.

- Proporciona servicios de datos en tiempo real para proyectos de migración y sincronización.

Message Queue Server

- Integración de procesos empresariales asincronos mediante conectividad basada en mensajes.
- Interfaces para los productos principales de Message Queue Server, incluyendo Microsoft MSMQ, IBM WebSphere, Tibco Rendezvous y Java Message Service (JMS).
- Mensaje/transacción de entrega garantizada reduce el coste de las interrupciones.
- Acceso optimizado para cada administrador de cola de mensajes diseñado para el mínimo esfuerzo administrativo.
- Integración en una sola aplicación de aplicaciones basadas en eventos y actividades que automáticamente desencadenan acciones en otras aplicaciones.
- Secuencias de ejecución dinámicas, orientadas a eventos y alertas.
- Capacidad para enviar y recibir mensajes entre sistemas distribuidos y dispares.

Administración mejorada y monitoreo

- Informes de estado y de rendimiento de trabajo e información de tendencias ofrecen la posibilidad de rastrear indicadores tales como el uso de CPU, memoria, I/O, etc. y entregar información actualizada sobre cómo funciona recientemente el proceso comparado con situaciones anteriores.
- Permite a los usuarios administrar y monitorear su entorno con una integración completa, incluyendo los siguientes tipos de trabajos y actividades:
 - o Trabajos de calidad de datos.
 - o Consultas programadas para actualizar la caché de la Federación.

- o Flujos de procesos.
- o Archivos de registro de acceso desde un panel central basado en la web, más rápido y de mejor solución de problemas.
- o Procesos SAS almacenados.
- o Trabajos de integración de datos.

Requisitos del sistema

Nivel de plataformas/servidor host

- HP/UX en Itanium: l1v3 (11.31)
- IBM AIX R64 sobre arquitectura POWER 7.1
- IBM z/OS: V1R11 y superiores
- Linux x 64 (64 bits): Novell SuSE 11 SP1; Red Hat Enterprise Linux 6.1; Oracle Linux 6.1
- Microsoft Windows x 64 (64 bits):
- Escritorio: Windows 7 * x 64 SP1; Windows 8 ** x 64
- Servidores: La familia Windows Server 2008 x 64 SP2; Familia de Windows Server 2008 R2 SP1; Familia Windows Server 2012
- Solaris en SPARC: actualización de la versión 10 9
- Solaris x 64 (x 64-86): actualización de la versión 10 9; Versión 11

Nivel de cliente

- Microsoft Windows (32 bits): Windows 7 x 88-64; Windows 8 x 88-64
- Microsoft Windows (64 bits): Windows 7 x 64 SP1; Windows 8 x 64

Nivel intermedio

- HP/UX on Itanium
- IBM AlXon POWER
- Linux x64 (x88-64)

- Microsoft Windows x64 (x88-64)
- Solaris (SPARC and x64)

Navegadores compatibles

- Internet Explorer 9: Windows 7 (32-bit y 64-bit)
- Internet Explorer 10: Windows 7 y Windows 8 (32-bit y 64-bit)
- Internet Explorer 11: Windows 7 y Windows 8 (32-bit y 64-bit)
- Firefox 6 y arriba: Windows 7 y Windows 8 (32-bit y 64-bit); Linux x 64: RHEL 6 y SLES 11 (32-bit)
- Windows 7 y Windows 8 (32-bit y 64-bit navegadores); Linux x 64: RHEL 6.1 y SLES 11 SP 1 (navegadores web de 32-bit)
- También puede consultar la página de soporte de terceros para obtener más información acerca de requisitos de software de terceros para su uso con SAS 9.4.

SAS SERVIDOR DE FEDERACIÓN

Los datos no tienen el máximo valor si son de difícil acceso. Además, la accesibilidad no debe poner en riesgo la seguridad. Federación de datos de SAS proporciona la accesibilidad y la flexibilidad que se necesita como parte de su estrategia de virtualización de datos.

Las opciones avanzadas de seguridad permiten determinar no solo la autorización del acceso a la información, sino también el detalle de la información a la que se puede acceder. Todo ello contribuye a garantizar que los datos no caigan en las manos equivocadas.

Una consola centralizada, basada en la web facilita gráficamente administrar, monitorear y mantener las conexiones, privilegios de usuario y actualizar los horarios. De esta forma es más fácil tratar con varias permutaciones y combinaciones de acceso a la información del usuario.

No hay ninguna razón para limitar el número de veces que se consulta tu sistema con nuestros datos de capacidad de almacenamiento en caché. Se facilita el acceso a la información que se necesita, tantas veces como sea necesario, sin afectar al sistema o elevando los costos.

SAS Federación Server proporciona acceso a fuentes como Hadoop, Netezza y SAP HANA virtualización de Big Data. Se unifica y normaliza el acceso a través de sistemas de una sola conexión, eliminando la necesidad de crear estrategias de acceso separado para cada origen de datos.

El Servidor de Federación SAS presenta caché refrescada (Figura 8-14), modelo de seguridad robusta (Figura 8-15), consola SQL (Figura 8-16) y funcionalidad de almacenamiento de datos en memoria (Figura 8-17).

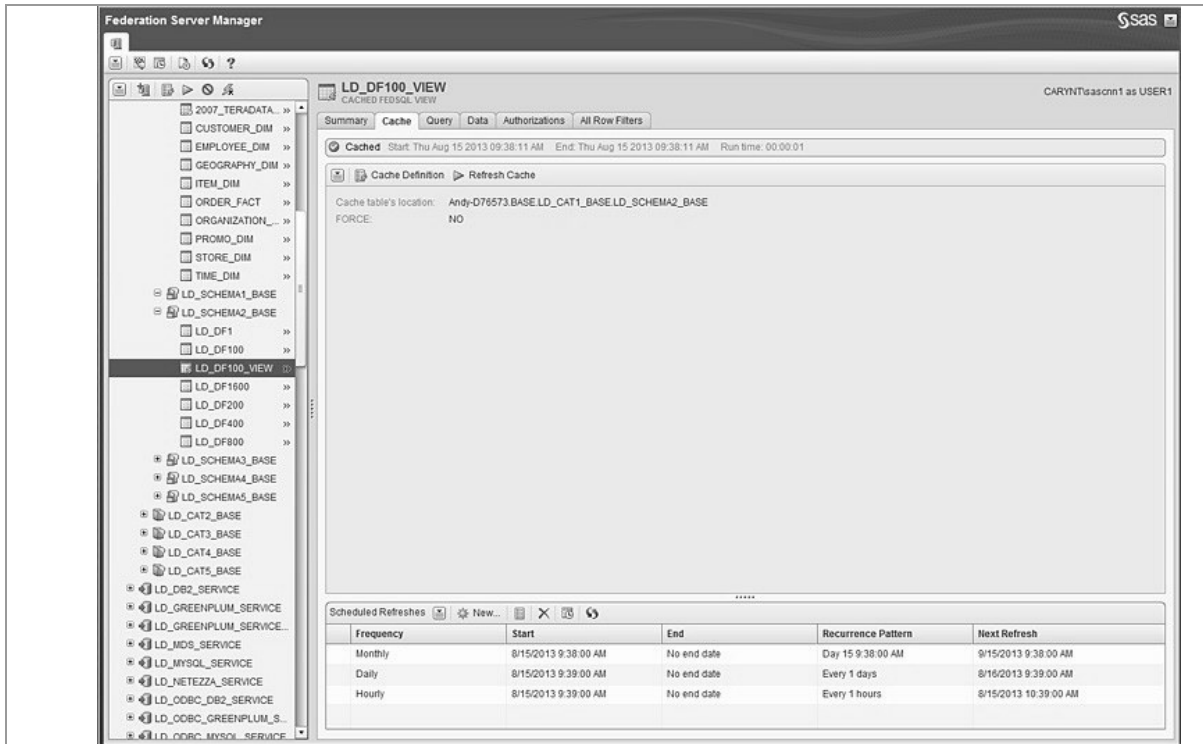


Figura 8-14

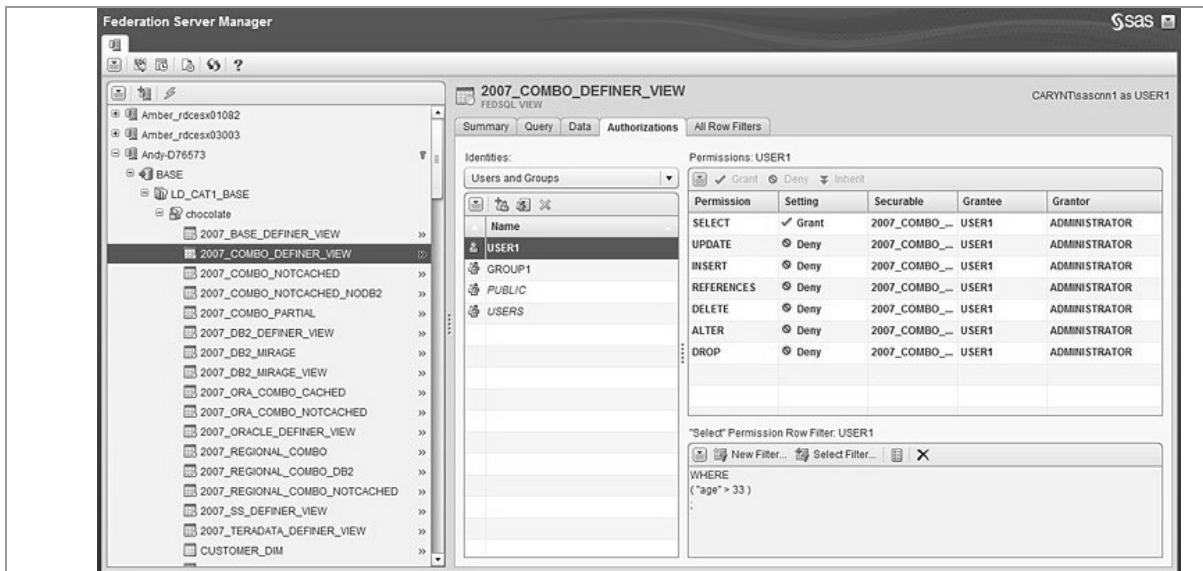


Figura 8-15

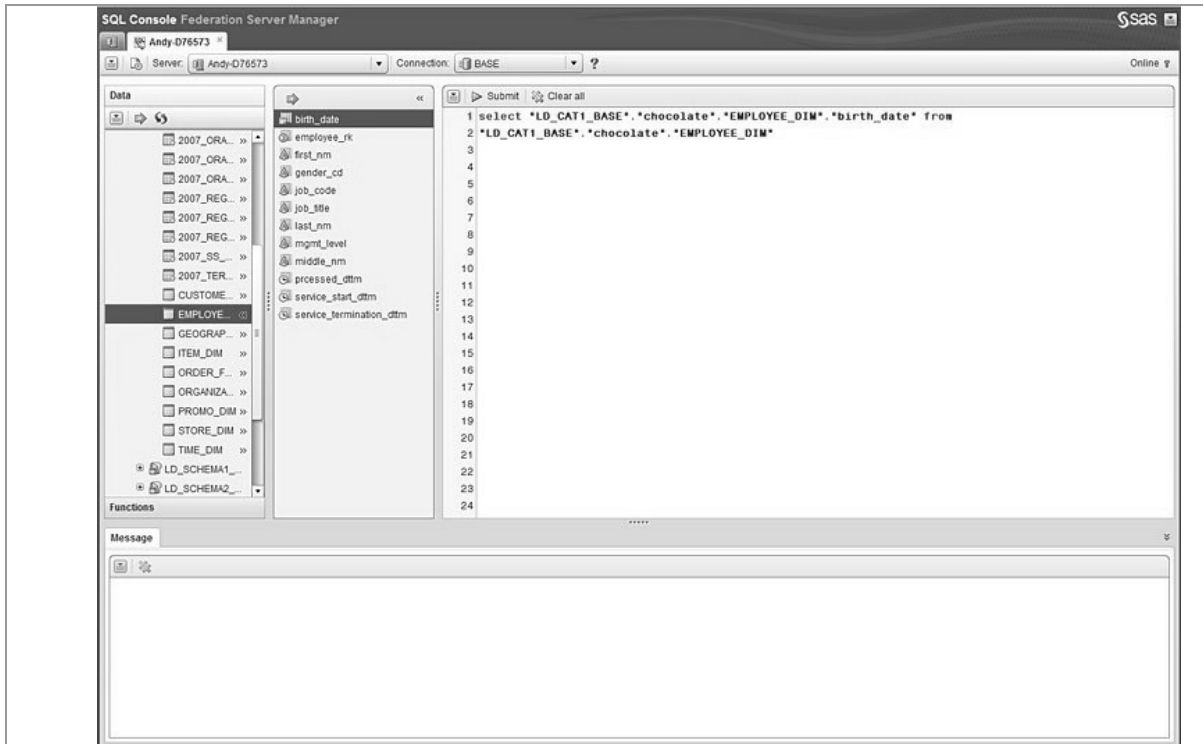


Figura 8-16

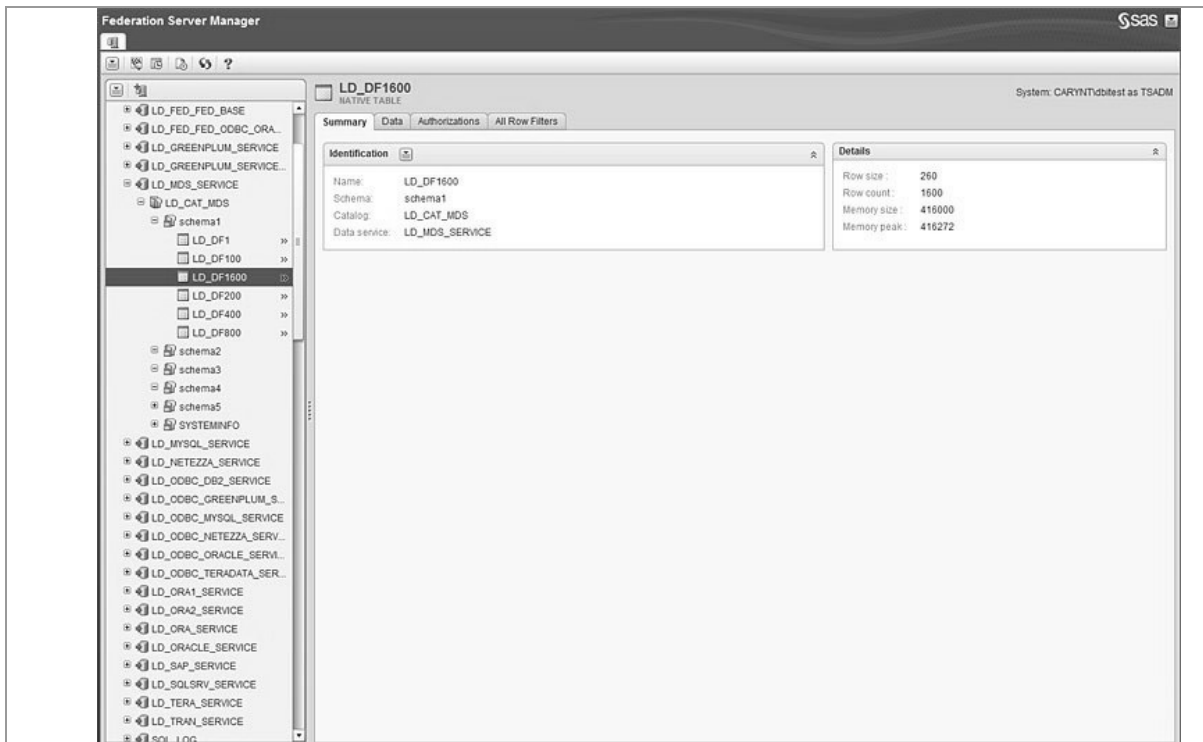


Figura 8-17

Características

Las fuentes de datos unificadas proporcionan una

visión de negocio solo a los almacenes de datos virtualizados

- Crear nombres de origen de datos federados (DSN), permitiendo a los usuarios acceder a múltiples fuentes de datos heterogéneas en la misma conexión.
- Acceder a aplicaciones de terceros en un entorno federado utilizando los controladores JDBC y ODBC. Acceso a datos Aster, Pivotal, Netezza, Sybase IQ, Oracle y Hadoop así como DB2, Teradata y SAS.

Administración centralizada, configuración y monitoreo

- Utilice la consola administrativa basada en Web para mantenimiento simplificado de acceso de los usuarios, privilegios y autorizaciones.
- Crear vistas materializadas rápidamente con un generador de consultas integradas.
- Activar o desactivar fácilmente la caché de datos.

Rendimiento mejorado al acceso a datos con cachés de datos y programación .

- Mantener consultas de aplicación actual y disponibles para los usuarios a través de vistas materializadas que son apoyados por el cacheo de datos.
- Conexiones y declaraciones entre los sistemas de datos de aplicación, servidor de Federación SAS y fuente de consulta.
- Información asegurada con ajustes administrativos.
- Definir permisos de acceso de un usuario o grupo en el catálogo, esquema, tabla, columna y nivel de fila.
- Registro del historial de acceso de los usuario para auditorías, mitigación de riesgos y presentación de informes de cumplimiento de políticas de seguridad.

Requisitos del sistema

Nivel de plataformas/servidor host

- HP/UX en Itanium: I11, I31
- IBM AIX R64 sobre arquitectura POWER 7.1
- IBM z/OS: V1R11 y superiores
- Linux x 64 (64 bits): Novell SuSE 11 SP1; Red Hat Enterprise Linux 6.1; Oracle Linux 6.1
- Microsoft Windows x 64 (64 bits): escritorio: Windows 7 x 64 SP1; Windows 8
x 64
Servidores: La familia Windows Server 2008 x 64 SP2; Familia de Windows Server 2008 R2 SP1; Familia Windows Server 2012
- Solaris en SPARC: actualización de la versión 10 9
- Solaris x 64 (x 64-86): actualización de la versión 10 9; Versión 11

Nivel intermedio

- HP/UX on Itanium
- IBM AIX on POWER
- Linux x64 (x86-64)
- Microsoft Windows x64 (x86-64)
- Solaris (SPARC and x64)

Navegadores compatibles

- Internet Explorer 9: Windows 7 (32 bits y x 64 32 bits)
- Internet Explorer 10: Windows 7 y Windows 8 (32 bits y x 64 32 bits)
- Firefox 6 y arriba: Windows 7 y Windows 8 (32 bits y x 64 32 bits); Linux x 64: RHEL 6 y SLES 11 (navegadores de 32 bits)
- Windows 7 y Windows 8 (32 bits y x 64 32 bits); Linux x 64:

RHEL 6.1 y SLES 11 SP 1 (navegadores de 32 bits)

Nivel de cliente

- Microsoft Windows (64 bits): Windows 7 x 64 SP1; Windows 8 x 64

SOFTWARE SAS BASE

El software SAS Base es la base para todo el software SAS. Junto con un lenguaje de programación fácil de aprender y flexible, se obtiene una interfaz de programación basada en web para manipulación de datos, almacenamiento de información y recuperación, estadística descriptiva e informes. También se dispone de un repositorio centralizado de metadatos y una instalación de marco que reduce la programación de mantenimiento y el tiempo.

Basado en una arquitectura abierta, multiplataforma, SAS Base es hardware-ágil y se integra en cualquier infraestructura informática de medio ambiente, que permite unificar sus esfuerzos informáticos y obtener una vista única de los datos.

Un lenguaje de programación intuitivo y fácil de aprender y programas empaquetados llamados procedimientos reducen significativamente la cantidad de código necesario para entregar información. Los procedimientos de SAS encapsulan y ofrecen la funcionalidad con unos pocos comandos simples, aumentando la productividad de programadores. SAS Studio, una interfaz basada en desarrollo web, permite acceder a datos archivos, bibliotecas y programas existentes desde cualquier dispositivo que tenga un navegador, haciendo la codificación de SAS más fácil y más accesible que nunca.

Capacidades de multiprocesos, escalables y de alto rendimiento permiten la ventaja del procesamiento paralelo para obtener el máximo de los recursos informáticos y producir respuestas más rápidas. Las rutinas de manipulación de datos pueden ser ejecutadas en los datos en bases de datos o en Hadoop, reduciendo la necesidad de mover los datos. Esto mejora el rendimiento y la seguridad.

SAS/SECURE™ se entrega como parte de la SAS Base y proporciona acceso al cifrado de datos estándar de la industria, incluyendo el Advanced Encryption Standard (AES). De esta forma se pueden cifrar los datos en los discos SAS y aumentar la seguridad de las contraseñas almacenadas.

Los programadores pueden leer, formatear, analizar e informar sobre

datos rápidamente, independientemente del formato.

SAS Base proporciona máxima flexibilidad de presentación de informes. Es posible crear fácilmente informes en formatos estándar de oficina tales como RTF, PDF, Microsoft PowerPoint, HTML y un formato de libros electrónicos que puede ser leído con iBooks en el iPad y iPhone. Es fácil producir informes y visualizaciones de resultados analíticos automáticamente de métodos estadísticos y entregarlos al personal de plataformas y aplicaciones de uso más.

El procedimiento de Hadoop le permite enviar comandos HDFS, programas de MapReduce y código de Pig contra datos Hadoop desde dentro de SAS. Es fácil almacenar grandes conjuntos de datos SAS en Hadoop y agregar los beneficios de SAS Base relativos a funciones de seguridad como el cifrado y protección mediante contraseña para implementaciones de Hadoop.

El software SAS Base presenta un entorno integrado (Figura 8-18).

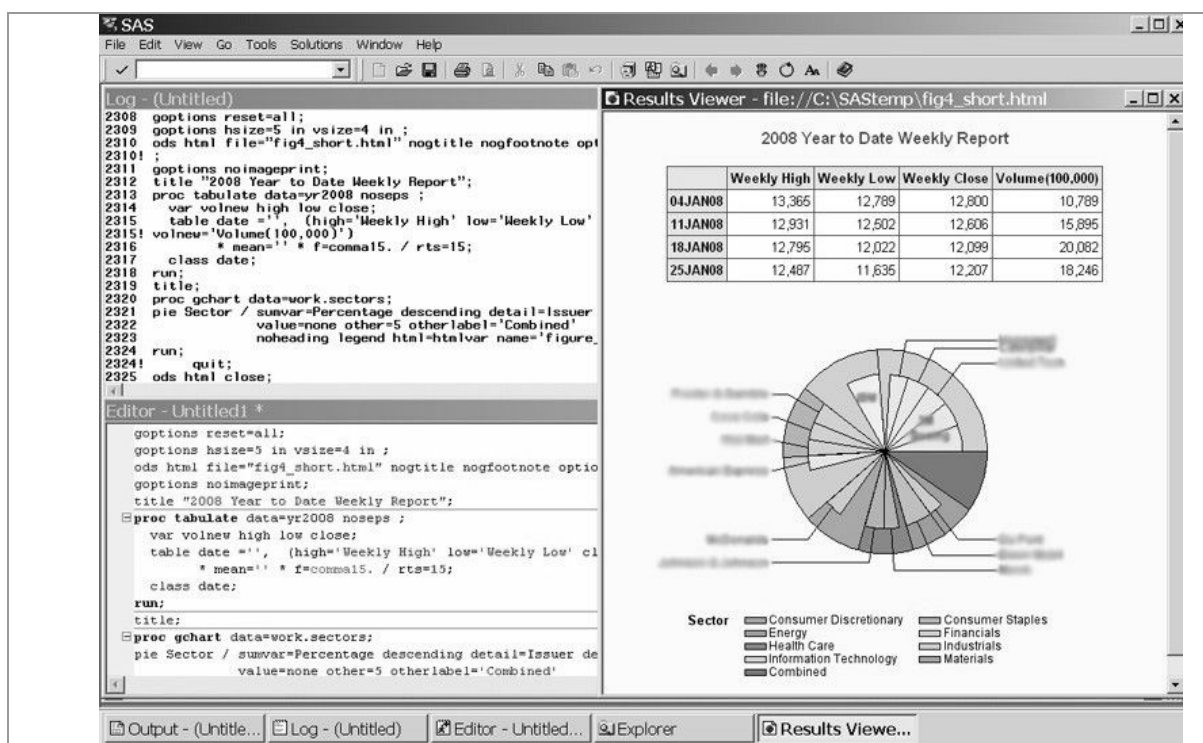


Figura 8-18

Características

Potentes capacidades de análisis de datos

- Capacidades de análisis que van desde simples estadísticas descriptivas a las correlaciones de datos avanzadas.
- Biblioteca de procedimientos de programación preconfigurados para administrar, analizar y presentar los datos.

4GL flexible

- 4GL Intuitivo con sintaxis fácil de aprender.
- SAS macro reduce la codificación para las tareas comunes y le permite modularizar trabajo fácil, reutilización y mantenimiento.
- Ejecuta interactivamente o en modo por lotes.
- Incluye SAS FedSQL, una aplicación propietaria SAS de ANSI SQL: 1999 con núcleo estándar, que proporciona de forma escalable un enfoque de alto rendimiento para acceder, manipular y administrar datos.
- Incluye DS2, un nuevo lenguaje de programación que permite la manipulación de datos avanzada a realizarse en los datos en donde residen en bases de datos.

Entorno de desarrollo basado en navegador fácilmente accesible (SAS Studio).

- Permite acceder a SAS desde cualquier lugar, utilizando cualquier dispositivo con un navegador web, sin ninguna instalación del cliente y con huella cero.
- Todos los programas SAS, archivos de datos y bibliotecas se acceden desde el escritorio, Mac y iPad a través del navegador web.
- Compatible con el sistema gestor de pantalla.
- La función Autocompletar muestra una lista de procedimientos SAS cuando se comienza a escribir un nombre de procedimiento.
- Genera automáticamente consultas SQL y permite acceder al código SQL generado detrás de los trabajos.
- Permite crear y añadir sus propios fragmentos de código a la biblioteca de fragmentos de código.
- Una interfaz de Point-and-click guía a través de analíticas o

procesos de manipulación de datos.

Soporte para amplia gama de formatos de datos

- Capacidad para leer los datos en cualquier formato, de cualquier tipo de archivo, incluyendo registros de longitud variable, archivos binarios, datos con formato libre e incluso archivos con datos faltantes o desordenados.
- Soporte para el lenguaje de consulta estructurado (SQL).
- Globalización con soporte completo para las codificaciones de caracteres más utilizados (por ejemplo, Latin1, Latin2 y conjuntos de caracteres de múltiples bytes para los idiomas hebreo, árabe y asiático).

Soporte para Hadoop

- Soporte para Pig, MapReduce y HDFS con comandos desde el entorno de ejecución de SAS.
- Es compatible con las referencias de archivo externo desde dentro de cualquier componente SAS. Es fácil trabajar con archivos de Hadoop.

Performance y escalabilidad

- Optimización de I/O en paralelo permite trabajar con grandes volúmenes de datos de manera oportuna. Los datos pueden ser repartidos a través de dispositivos para un acceso más rápido, pero se hace referencia a todos ellos como un conjunto único de datos.
- La creación de índice paralelo reduce el tiempo necesario para crear conjuntos de datos grandes con varios índices o anexar datos a los conjuntos de datos existentes.
- Los procedimientos clave SAS son multiproceso para una ejecución más rápida de tareas estándar como recapitulación de datos y clasificación.
- La escalabilidad de SAS Base permite abarcar múltiples máquinas y redes utilizando el software SAS/CONNECT[®], que está disponible por separado.
- Varios procedimientos SAS Base de optimización de SQL se enfocan a Hive en Hadoop, datos Aster, Pivotal Greenplum Database,

IBM DB2, Netezza, Oracle y Teradata.

- Catálogos de formato SAS pueden ser publicados y compilados dentro de bases de datos (Hive de Hadoop, datos de Aster, Pivotal Greenplum Database, IBM DB2, Netezza, Oracle y Teradata) haciendo que formatos pueden ser aplicados a los valores de datos reales durante la ejecución de la consulta.

Interoperabilidad e implementación de plataforma múltiple

- La arquitectura MultiVendor permite que los programas se escriban una vez y se ejecuten en cualquier lugar, independientemente del hardware o sistema operativo.
- La arquitectura abierta de metadatos SAS permite diferentes aplicaciones de intercambio de metadatos.
- Puede ejecutar código Groovy en la máquina Virtual de Java.

Capacidad de administración

- SAS Environment Manager proporciona una vista única al sistema y los recursos SAS para monitoreo proactivo, alertas y administrar entornos SAS Business Analytics.
- SAS Management Console proporciona una GUI extensible Java para administrar tareas SAS.
- Motor XML importa y exporta una gran variedad de documentos XML.
- Una interfaz de arrastrar y soltar crea mapas XML.
- La capacidad de control y reinicio permite a los usuarios presentar un programa fallido en el modo de reinicio de ejecución completa, reanudando con el paso que no se ha completado cuando se produjo el error.
- La interfaz de aplicaciones de medición de respuesta (ARM) controla la disponibilidad y el rendimiento de las transacciones dentro y a través de diversas aplicaciones.

Formatos de informes flexible de salida

- ODS proporciona un número casi ilimitado de opciones de presentación de informes y visualización de resultados analíticos.
- Gráficos de alta calidad se elaboran con Base SAS 9.4. Se incluye:
 - o ODS para gráficos estadísticos.
 - o La familia de SG de procedimientos.
 - o El lenguaje de plantillas de gráfico.
 - o El diseñador gráfico de ODS.
 - o El editor de gráficos de ODS.
- Crear informes en formatos estándar como RTF, Microsoft PowerPoint y PDF. Todos los formatos están disponibles en todas las plataformas.
- Crear informes como libros electrónicos que pueden ser leídos con iBooks en el iPad y iPhone.
- Crear visualmente atractivos gráficos de salida analítica por defecto (ninguna programación adicional).
- HTML 4, HTML 5 y XML se encuentran entre los lenguajes de marcado proporcionados. Se puede modificar cualquier lenguaje de marcado que proporciona SAS o crear su propio lenguaje de marcado para la salida. HTML es ahora el destino predeterminado para la salida.
- Personalizar o modificar la jerarquía de salida. Salida de reproducción a los diferentes destinos sin volver a ejecutar el programa.

Algoritmos de cifrado de datos estándar de la industria

- Estableciendo una conexión entre las tablas de datos físicos y los metadatos, SAS asegura que la seguridad se aplica constantemente, independientemente de cómo un usuario solicita acceso de SAS.
- SAS/SECURE ahora se entrega junto con SAS Base. Todavía es un producto independiente pero sin ningún costo adicional para utilizar SAS/SECURE en SAS 9.4.
- Incorporación de Advanced Encryption Standard (AES). Mediante

este algoritmo estándar de la industria, puede cifrar los datos en los discos SAS.

Requisitos del sistema

Nivel de plataformas/servidor host

- HP/UX en Itanium: I113 (11,31)
- IBM AIX R64 sobre arquitectura POWER 7.1
- IBM Z/OS:V1R11 y superiores
- Linux x 64 (64 bits): Novell SuSE 11 SP1; Red Hat Enterprise Linux 6.1; Oracle Linux 6.1
- Microsoft Windows x 64 (64 bits):
- Escritorio: Windows 7 x 64 SP1; Windows 8 x 64
- Servidores: La familia Windows Server 2008 x 64 SP2; Familia de Windows Server 2008 R2 SP1; Familia Windows Server 2012
- Solaris en SPARC: actualización de la versión 10 9
- Solaris x 64 (x 64-86): actualización de la versión 10 9; Versión 11

Nivel de cliente

- Microsoft Windows (32 bits): Windows 7 x 88-64; Windows 8 x 88-64
- Microsoft Windows (64 bits): Windows 7 x 64 SP1; Windows 8 x 64

HERRAMIENTAS PARA EXPLORAR Y VISUALIZAR DATOS CIENTÍFICOS

En este apartado SAS encaja las siguientes herramientas:

- **SAS Visual Analytics.** Permite explorar visualmente los datos, descubrir nuevos patrones y publicar informes en la web y dispositivos móviles.
- **SAS In-Memory Statistics for Hadoop.** Permite trabajar el análisis de datos en Hadoop con un entorno que se mueve rápidamente a través de cada fase del ciclo de vida analítico.

La Figura 8-19 muestra algunas herramientas SAS para trabajar en la memoria en Hadoop.

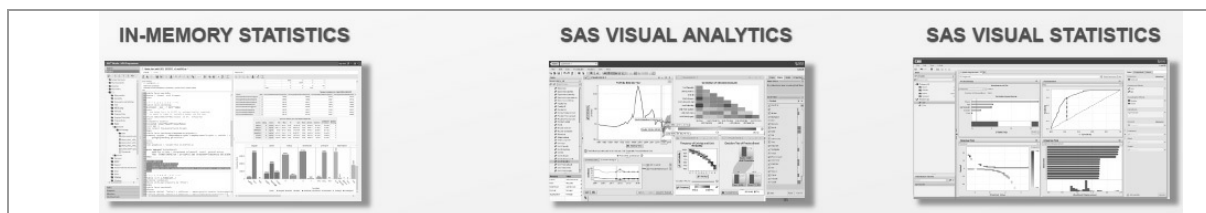


Figura 8-19

SAS VISUAL ANALYTICS

Con Visual Analytics se facilita la analítica automatizada. El procesamiento en memoria hace que el trabajo sea rápido. Se pueden obtener respuestas en solo minutos o segundos con el software de visualización de datos de SAS.

Sin importar el tamaño de la organización, o los datos, con SAS Visual Analytics se pueden explorar todos los datos relevantes rápida y fácilmente. Se puede ver más opciones de las habituales, descubrir oportunidades ocultas, identificar relaciones clave y tomar decisiones más precisas y más rápidas que nunca antes (Figura 8-20).

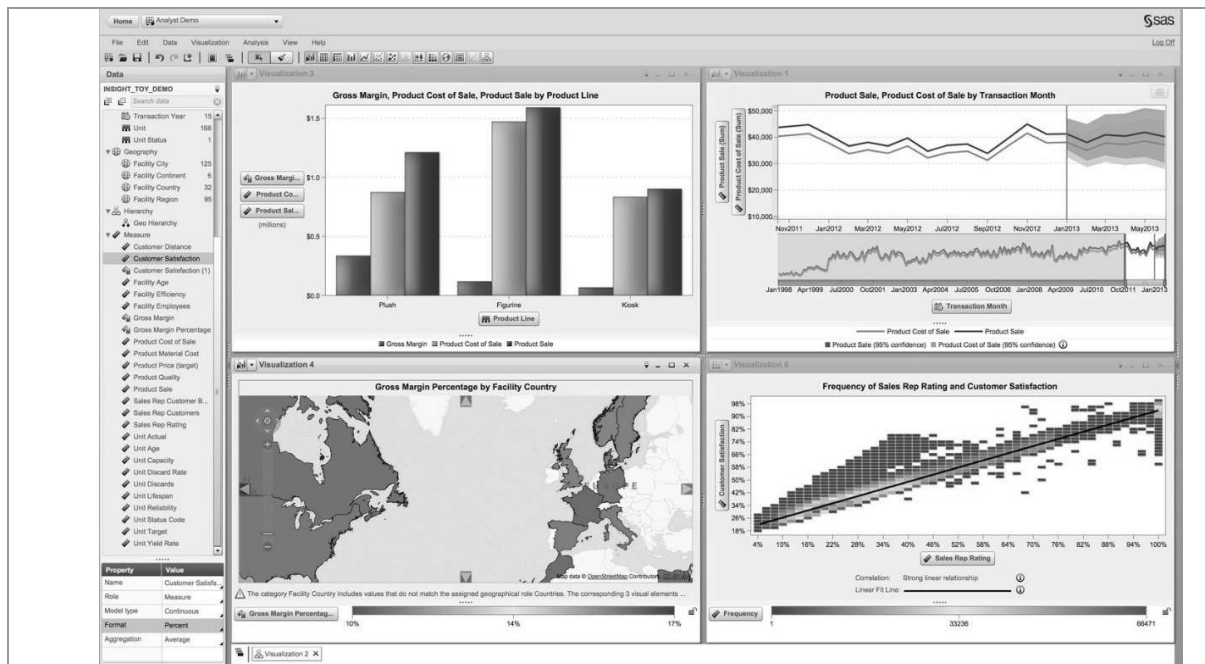


Figura 8-20

Esta herramienta proporciona análisis de gran alcance a usuarios de negocios con limitadas capacidades técnicas, a estadístico o a científico de datos. Análisis sofisticados, incluyendo los árboles de decisión (Figura 8-21), diagramas de red, predicción y análisis de escenarios, se han integrado perfectamente con las características de facilidad de uso como las capacidades de arrastrar y colocar, autocargado y otras. Cualquier persona puede entender y beneficiarse del conocimiento escondido en datos complejos.

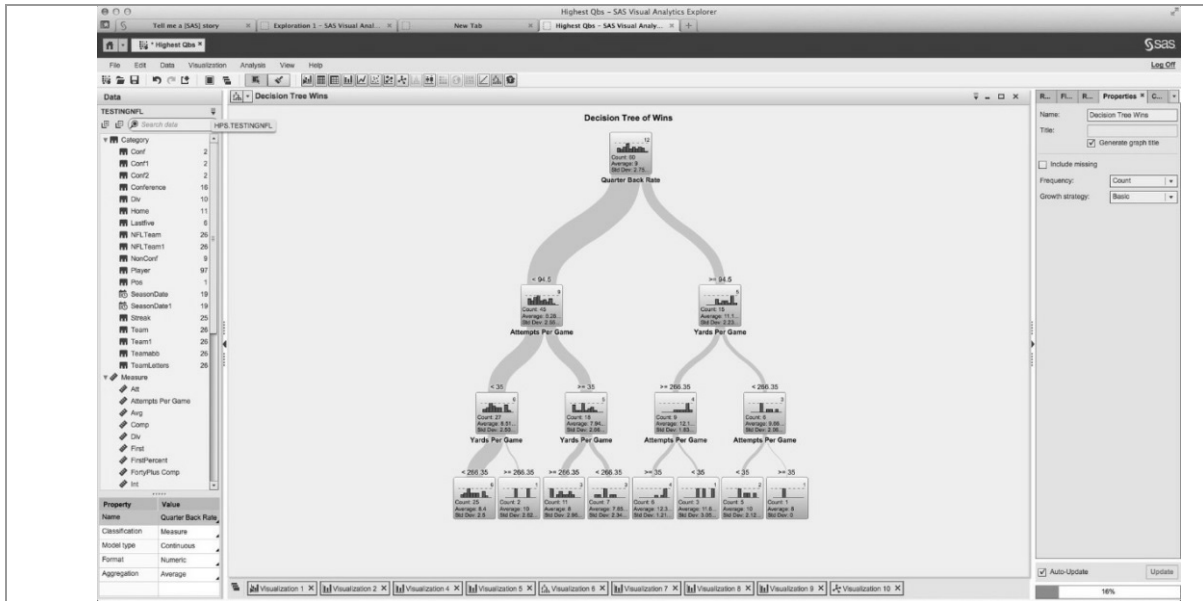


Figura 8-21

Visual Analytics permite un rápido diseño de informes que son atractivos, interactivos y significativos. Se distribuyen fácilmente vía web y son integrables en las aplicaciones de Microsoft o dispositivos móviles. Es posible incluso crear informes que permiten a los beneficiarios explorar los datos por cuenta propia (Figura 8-22).

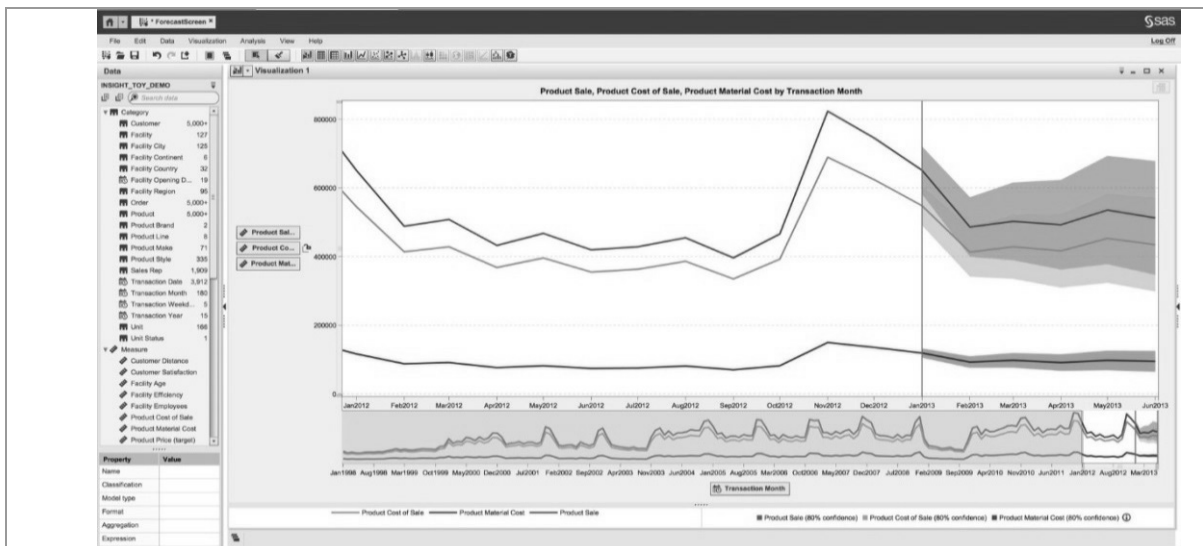


Figura 8-22

SAS Visual Analytics ofrece capacidades de inteligencia de negocio móvil permitiendo a los ejecutivos un fácil acceso y exploración de informes y cuadros de mando desde sus dispositivos móviles, en cualquier momento y en cualquier lugar (Figura 8-23). La exploración de informes es accesible incluso cuando no hay ninguna conexión a Internet.



Figura 8-23

SAS Visual Analytics proporciona lo último en flexibilidad de implementación. Si desea implementar SAS Visual Analytics in situ usando su hardware (entorno simple o distribuido), en su propia nube privada, en una nube pública como Amazon o en el entorno Cloud SAS, existen todo tipo de opciones para satisfacer las necesidades organizacionales (Figura 8-24).

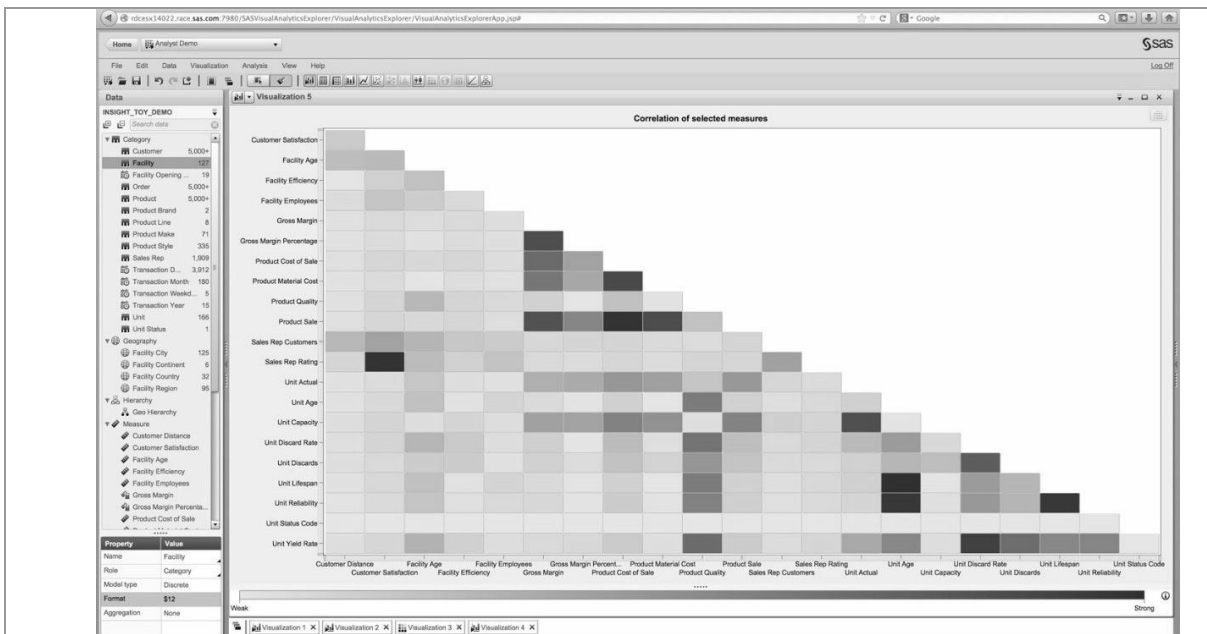


Figura 8-24

Más de 400 millones de tweets son enviados cada día. ¿No sería maravilloso saber que lo que se dice acerca de su empresa o productos? Ahora es posible Importar datos de Twlter a SAS Visual Analytics y analizarlos. Aplicando análisis del texto a los datos de Twitter, al cliente puede obtener Información rápida en los “temas candentes” del día (Figura 8-25). También se pueden realizar análisis para ver qué palabras se utilizan con mayor frecuencia para poder determinar qué temas son los más Importantes y garantizar mayor exploración.



Figura 8-25

Características

Visualización de datos

- El modo de exploración de datos interactivo, basado en la web se adapta a las necesidades de todo tipo de usuarios.
- La característica de autográfico (Autocharting) elige automáticamente el gráfico más adecuado para mostrar los datos seleccionados.
- Los mapas geográficos proporcionan vistas para un entendimiento

rápido de datos geoespaciales.

- Los mapas de Choropleth hacen fácil visualizar las variaciones de medida sobre un área geográfica.
- Es posible utilizar capacidades para identificar automáticamente y explicar las relaciones entre variables.
- Las capacidades de exploración están disponibles para las fuentes de datos en la memoria de servidor.
- A los gráficos pueden aplicársele aspecto 3D y efectos de luz.
- Se pueden realizar visualizaciones atractivas incluyendo diagramas de caja, mapas de calor, gráficos de burbujas, gráficos animados y mucho más.
- Existen posibilidades de mostrar las líneas de cuadrícula adecuadamente para permitir ajustar los ejes para optimizar la visualización.
- Las consultas pueden cambiarse seleccionando los datos que se muestran desde una barra lateral o por filtración dinámica y agrupación.
- Existen capacidades de predicción adicionales a los gráficos que aportan capacidades para explorar datos y averiguar lo que podría pasar en el futuro.
- Pueden mostrarse las miniaturas de artículos recientes y favoritos.

Análisis fácil

- El procesamiento en memoria servidor proporciona un análisis rápido de los datos.
- Es posible consultar fácilmente datos de un conjunto de modos de visualización.
- Datos multidimensionales pueden ser filtrados aplicando filtros en cualquier nivel de jerarquía.
- Las jerarquías incluyen niveles extensibles y plegables.
- Una estadística descriptiva visible proporciona una visión general de la materia.

- Nuevas medidas pueden ser calculadas y agregadas.
- Se pueden generar pronósticos sobre la marcha con previsión de intervalos de confianza incluidos.
- El algoritmo de predicción más apropiado para datos específicos se selecciona automáticamente.
- Usando análisis de escenarios, se puede ver cómo afectaría la previsión modificando las distintas variables.
- Existe una capacidad de generar interactivamente los árboles de decisión para representar gráficamente los resultados probables. Un nivel experto permite modificar ciertos parámetros que influyen para la generación del árbol.
- Análisis del texto, incluyendo contenido y categorización.
- Diagramas de red con la capacidad de mostrar las redes a través de un mapa, incluyendo mapas ESRI.
- Integración con la tecnología de mapeo de ESRI.

Informes robustos

- Se proporciona una interfaz basada en la web para la creación de informes interactivos.
- El asistente para adquisición de datos está disponible para previsualizado, filtrado o muestreo de datos antes de crear visualizaciones o informes.
- Los datos se pueden cargar en servidor SAS LASR para una analítica avanzada desde dentro de la interfaz de diseño del informe.
- Un diseño de precisión proporciona flexibilidad en el diseño y presentación del informe.
- Existen plantillas de gráfico personalizado.
- Pueden crearse jerarquías para agregar funcionalidades de desglose a informes y visualizaciones.
- Se pueden seleccionar filtros predefinidos, establecer agrupaciones y clasificaciones y anular formato predeterminado.
- Se crean fácilmente cálculos personalizados y filtros progresivos.

- Se incluye una variedad de gráficos: Barra 3-D con múltiples líneas, pastel, pastel 3-D, línea, dispersión, mapa de calor, burbuja, burbuja animada y azulejo, todo lo cual permite líneas de referencia.
- Integración con la tecnología de mapeo de ESRI.
- Capacidades de filtrado y selección pueden añadirse a los informes con la facilidad de integrar elementos de acción comunes tales como botones, botones de opción, casillas de verificación y deslizadores.
- Se pueden establecer alertas basadas en rangos o basadas en umbrales para suscribirse a informes y recibir avisos cuando cambia un informe.
- Modo de invitado para la visualización de las exploraciones, informes y cuadros de mando sin requerir un ID de usuario o contraseña.
- Integración perfecta con SAS Office Analytics y SAS Add-In para Microsoft Office que permite al usuario abrir informes desde sus aplicaciones de Microsoft Office.
- Capacidad para agregar comentarios consolidados en un informe. Los usuarios móviles, los usuarios de Microsoft Office y usuarios de la web pueden todos compartir y ver los comentarios en una ubicación central.

BI móvil

- Nativa iPad y Android utilizan capacidades como zoom, etc., para optimizar la facilidad de uso.
- Diseños flexibles permiten crear contenidos adaptados a diferentes necesidades.
- Los informes pueden verse correctamente en dispositivos móviles ya sea online u offline.
- Un nuevo soporte de colaboración Incluye la posibilidad de anotar, compartir otros informes por correo electrónico.
- Pueden capturarse Imágenes y comentarlos a compartir con otros.
- Las alertas se envían a los dispositivos móviles cuando se actualizan los informes.

- La integración con terceros proporciona administración de dispositivos móviles.

Servidor analítico de alto rendimiento en memoria

- Utiliza análisis en memoria para llevar a cabo rápidamente exploración y análisis.
- Diseñado para ejecutarse en un servidor único para pequeñas organizaciones y departamentos.
- Optimizado para que los entornos distribuidos utilicen la URL con capacidades de muchos nodos a escala cuando el procesamiento de datos crece.
- Integración con Hadoop para optimización del rendimiento y la escalabilidad.
- Puede utilizarse en aplicaciones de base de datos o en una nube.

Administración de IT

- interfaz fácil de usar basada en web para la administración de IT.
- Los activos de información SAS, incluyendo usuarios, servidores y datos, se manejan fácilmente.
- Las estructuras de datos se traducen fácilmente en términos que todos pueden entender.
- Soporte para esquemas dentro del servidor analítico SAS LASR que pueden utilizarse para la presentación o exploración al igual que otras tablas.
- Soporta autorización de Información y autenticación de usuario en el manejo de datos.
- Los datos se cargan en la memoria de los servidores basados en el volumen, frecuencia de actualizaciones y necesidades de escalabilidad.
- Despliegue flexible.
- Implementaciones tradicionales locales.
- Servidor único con soporte para servidores Windows apropiado para pequeñas y medianas organizaciones.
- El modo distribuido es fácilmente ampliable para apoyar demandas

crecientes de datos (en aplicaciones de base de datos o hardware de los productos básicos).

- Implementaciones de nube:
 - Proveedores de nube pública, incluyendo servicios web de Amazon EC2.
 - Su propia nube privada o data centers virtualizados.
 - The SAS Cloud.
- SAS Solutions OnDemand:
 - Proporciona una experiencia personalizada con soluciones SAS bajo demanda para los servicios de nube privada.

Requisitos del sistema. Entorno de servidor

Operating Systems

- Red Hat Enterprise Linux 6.
- SuSE Linux Enterprise Server 11.
- Oracle Linux 6.1.
- Windows (solamente para los despliegues): Windows Server 2008 R2 Enterprise SP1, Windows Server 2008 R2 Datacenter SP1, Windows Server 2012 Standard, Windows Server 2012 Datacenter.

Hardware

- HP y Dell (con opciones de empaquetado de software y hardware preconfigurado).
- Hardware de base de datos Teradata y Pivotal (previamente Greenplum).
- SAS también trabajará con los clientes que deseen utilizar hardware de otros proveedores, incluyendo IBM, Cisco y mucho más.

Middle Tier

- Servidor de aplicaciones Web SAS (incluido).

Requisitos del sistema. Entorno del cliente

Navegadores

- Internet Explorer 9 y superiores (modo nativo).
- Firefox 6 y más.
- Cromo 15 y hacia arriba.

Flash Player

- Se requiere Adobe Flash Player 11.1 o más reciente.

Software necesario

- Puede ser requerido software adicional, tal como varios motores de acceso/SAS para acceder a fuentes de datos nativas. Tenga en cuenta que una opción de interfaz SAS/ACCESS a un origen de datos se incluye con SAS Visual Analytics.
- Para integrar SAS Visual Analytics y SAS Office Analytics, las versiones siguientes son necesarias: SAS Visual Analytics 6.4 y SAS Office Analytics 6.1M1 (hotfix 1).

Soporte al cliente de BI móvil SAS

iOS

- SAS BI móvil para iPad es una aplicación gratuita para iOS disponible en la App Store de iTunes.
- Plataforma/sistema operativo: Apple iOS v6.0 o superior.
- Dispositivos: iPad 2, 3, 4, Mini y iPad aire.

Android

- SAS BI móvil para Android es una aplicación gratuita para Android disponible desde Google Play:
- Android 4.4:

- Google Nexus 7 (2012 and 2013 editions).
- Google Nexus 10.1.
- Android 4.1, 4.2, 4.3:
- Samsung Galaxy Tab 2 10.1.
- Samsung Galaxy Note 10.1.
- Samsung Galaxy Note 8.
- Google Nexus 7 (2012 and 2013 editions).
- Google Nexus 10.1.
- Android 4.0.3/4.0.4:
- Samsung Galaxy Tab 10.1.
- ASUS EEE Pad Transformer TF101.

SAS IN-MEMORY STATISTICS FOR HADOOP

Obtener ideas fuera de Hadoop de manera oportuna requiere un enfoque diferente. Se necesita analítica en memoria y preparación interactiva de datos analíticos, exploración, modelado e implementación. Todo ello para obtener respuestas precisas al instante.

Se aplican algoritmos estadísticos probados y técnicas de aprendizaje máquina para encontrar las mejores respuestas. Puede explorar y utilizar múltiples enfoques analíticos para revelar ideas y tomar decisiones de alto impacto.

Hasta ahora, estadísticos y datos científicos debían reunir los distintos lenguajes de programación o productos de acceso, preparar, modelizar y anotar datos en Hadoop. Y cuando llegó el momento de poner en funcionamiento los modelos, el software no pudo escalar. De la preparación de datos y la exploración a la construcción de modelos y despliegue, la solución está probada y puede escalar a su entorno de producción.

Varios usuarios pueden analizar simultánea e interactivamente grandes datos en Hadoop usando un lenguaje de programación analítico rápido en memoria. Preparar, manipular, transformar, explorar, modelizar y evaluar datos se realiza todo dentro de Hadoop.

La infraestructura en memoria funciona sobre Hadoop, elimina el movimiento de datos costosos y permite datos en la memoria de durante toda la sesión analítica. Esto reduce significativamente la latencia de datos y proporciona un análisis rápido.

Con SAS In-Memory Statistics para hadoop es posible un rápido análisis interactivo (Figura 8-26).

También es posible construir bosques de decisión aleatoria (Figura 8-27) y sistemas de recomendaciones (Figura 8-28)

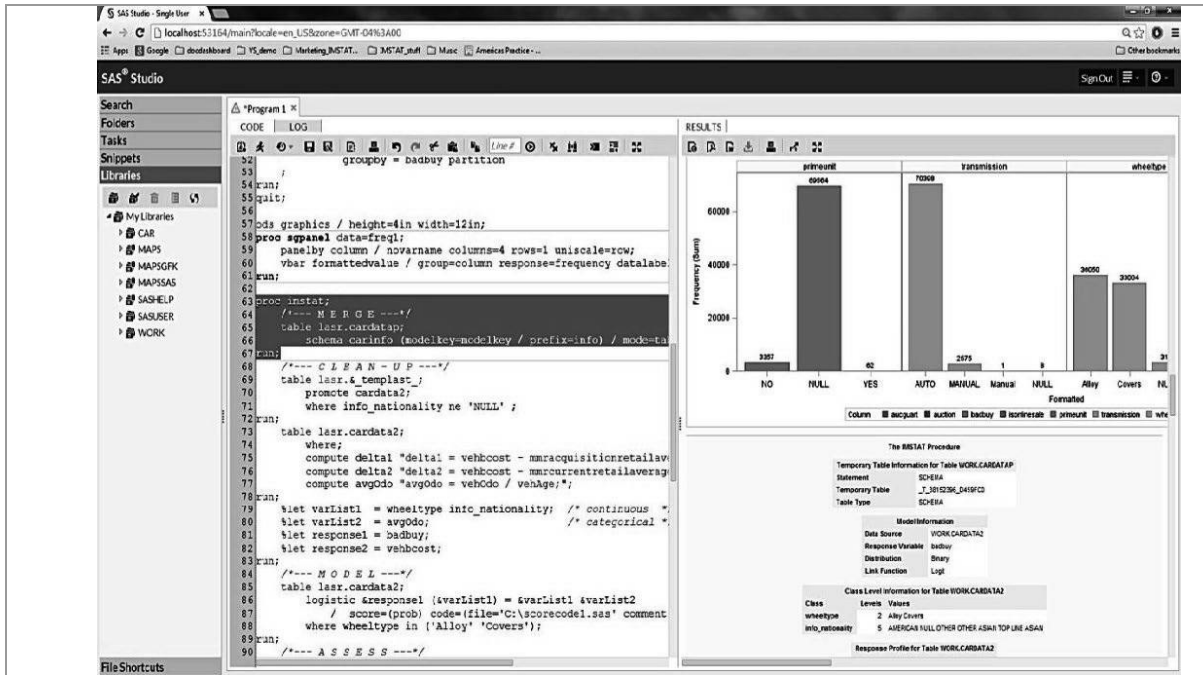


Figura 8-26

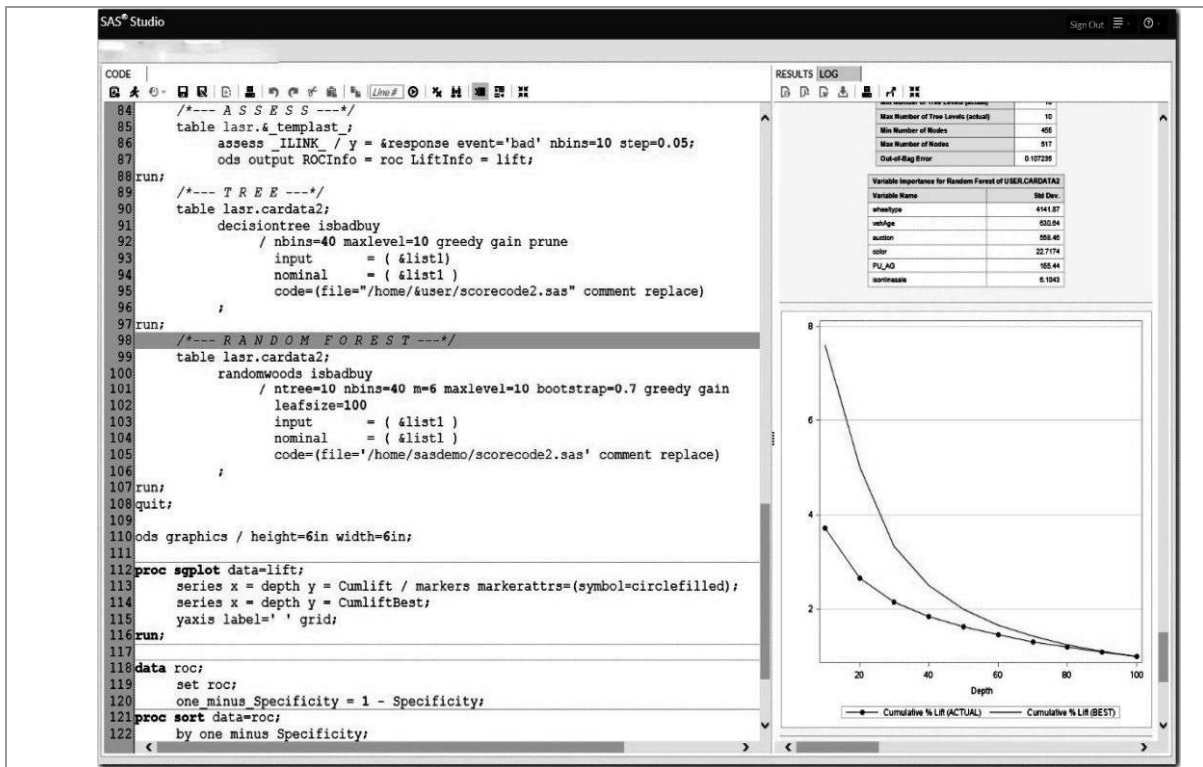


Figura 8-27

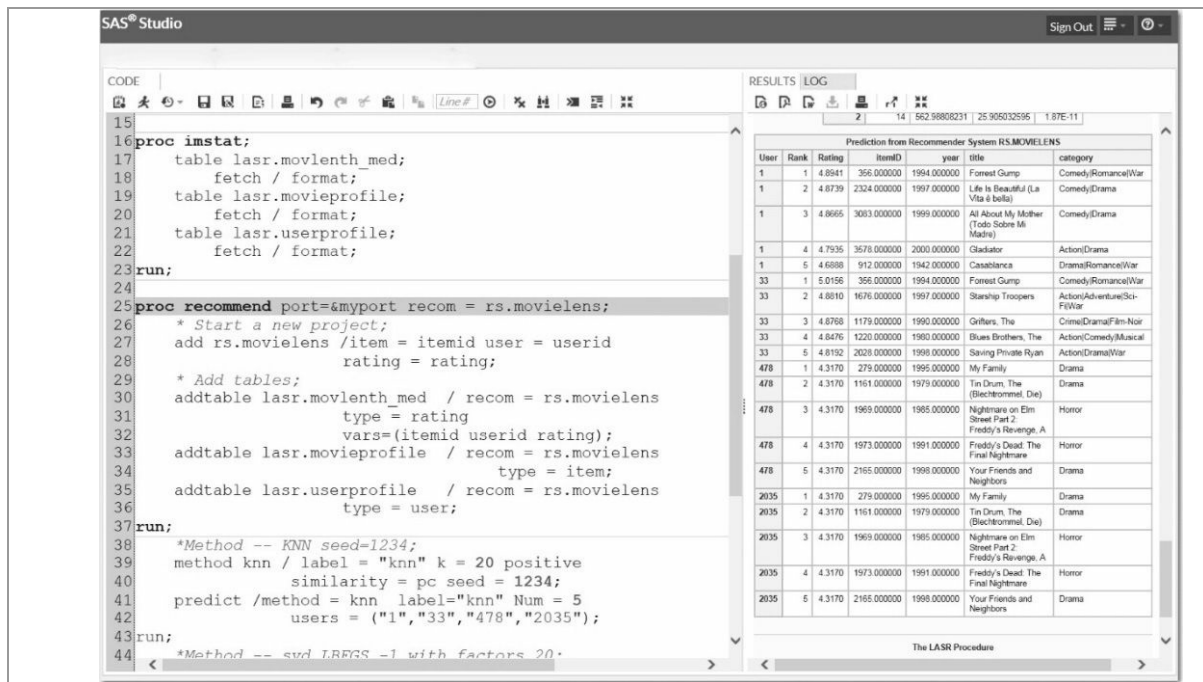


Figura 8-28

Características

Programación interactiva en memoria

- Realiza todos los cálculos matemáticos en memoria.
- Utiliza un grupo de procesamiento de operación dinámica para calcular y procesar resultados para cada grupo, división o segmento sin tener que ordenar los datos del índice de cada vez.
- Ofrece una nueva interfaz basada en web, SAS Studio, para programadores de SAS.
- El lenguaje de programación interactiva soporta enviar, recuperar resultados y luego someter más declaraciones sobre la marcha.
- Encadenar tareas analíticas como un solo trabajo en memoria sin tener que cargar los datos o escribir resultados intermedios a los discos.
- Puede actualizar tablas de origen con nuevas transformaciones de columna, filtrar filas y realizar procesamiento de agrupar por.

Preparación de datos analíticos

- Sistema de archivo distribuido colocado en Hadoop.

- Particionamiento inteligente por una variable (s) en el cluster para acceso de datos más eficiente. Puede crear particiones de los datos en cualquier momento.
- Derivar nuevas tablas temporales y promoverlas para su uso por otros analistas.
- Manipulación de datos para subconjuntos.
- Las fuentes de datos se pueden definir mediante la actualización, anexo y filtrado. Se derivan declaraciones de agregado.
- Exportar tablas de resultados ODS para el desarrollo gráfico del cliente.

Estadística descriptiva

- Estadísticos de centralización y dispersión, incluyendo valores atípicos para una o más variables.
- Correlaciones para medir el coeficiente de correlación de Pearson para un conjunto de variables.
- Tabulaciones cruzadas.
- Tablas de contingencia, incluyendo medidas de asociaciones.
- Procesamiento en paralelo por grupo.
- Histogramas con opciones para controlar umbrales de valor máximo y otras, características.
- Resúmenes multidimensionales en una sola pasada de los datos.
- Percentiles para una o más variables.
- Obtención de estadísticas de resumen como el número de observaciones, el número de valores faltantes, la suma de los valores nonmissing, media, desviación estándar, errores estándar, sumas de cuadrados sin corregir y corregidas, min y max y el coeficiente de variación.
- Estimación kernel de funciones de densidad usando normal, tri-cubo y funciones cuadráticas del núcleo.

Algoritmos estadísticos y técnicas de aprendizaje máquina

- Árboles de clasificación y regresión.
- Basado en el algoritmo C4.5.
- El control sobre el criterio de separación (ganancia de información, ratio de ganancia de información).
 - Establecer la profundidad del árbol, máximo de ramas, tamaño de la hoja, poda y mucho más.
- Pronóstico.
- Identificar automáticamente las características estadísticas de una serie temporal y seleccionar los modelos apropiados.
 - Manejar valores atípicos, cambios estructurales, datos desigualmente espaciados y eventos del calendario.
- Agregar las series temporales mediante la suma o promedio.
- Especificar y limitar la longitud del horizonte del pronóstico.
- Modelos lineales generales.
- Respuestas continuas.
- Soporte para efectos de intervalo y clase.
- Especificar términos de interacción.
- Selección del modelo paso a paso.
- Variables peso y FREQ.
- Modelo lineal generalizado.
- Modelos lineales generalizados y clase exponencial de modelos.
- Las distribuciones soportadas incluyen Beta, exponencial, Gamma, Poisson, inversa gaussiana, Binomial Negativa, T de Student y Weibull.
 - Completo conjunto de funciones de enlace, incluyendo la identidad, Logit, Probit, LogLog Log, Cloglog, recíproco, de alimentación-2 y power.
 - Técnicas de optimización incluyendo gradiente conjugado, doble-dog-leg, Nelder-Mead, Newton-Raphson con y sin rebordes, quasi-Newton y la región de confianza.

- Frecuencia y ponderación de las variables (FREQ y peso).
- Regresión logística.
- Modelos para datos binarios o binomiales logit, log-log y log-log complementarlo enlazan funciones.
- Múltiples técnicas de optimización.
- Soporte variable compensada.
- Conveniente para el montaje de modelos polinómicos lineales, cuadráticos o cúbicos para cada par de variables numéricas.
- La mejor opción para volver a los modelos de regresión con el mayor coeficiente de determinación.
 - El control sobre la orden polinomial.
 - Extenso diagnóstico residual, Influencia y aprovechamiento del análisis.
- Decisión al azar y los bosques de regresión.
- Apoya el desarrollo de modelos de conjunto al azar en el bosque.
- Rápido crecimiento de los árboles en un bosque en paralelo.
- Árboles de conjunto por votación mayoritaria.
- Definir el tamaño de arranque; muestreo de apoyo con y sin reemplazo.
 - Cálculo basado en un conjunto de árboles de Importancia variable.
 - El control sobre el bosque incluyendo número de árboles, número de variables a evaluar en cada punto de partida, criterio de división y mucho más.
 - El control sobre los árboles incluyendo el tamaño de la hoja, ramas máximas, número de compartimientos para evaluar y más.
- Clustering.
- k-means clustering.
- Clustering espacial basado en la densidad.
 - El control sobre el tamaño del cluster, semilla, criterio de convergencia, número de Iteraciones y más.

- La distancia soportadas incluyen medidas cuadrado euclídiano, Manhattan, máximo, coseno, Jaccard y Hamming. K-means utiliza distancia euclídiana.

Modelo de evaluación

- Soportes para generar resúmenes de comparación de modelos tales como curvas ROC, tablas estadísticas y clasificación de concordancia, para uno o más modelos.

Puntuación del modelo

- Generación de código de paso para los datos de SAS.
- Declaración de puntuación para aplicar puntuación lógica a la formación, la retención y nuevos datos.

Análisis de texto

- Análisis sintáctico y derivados.
- Iniciar y detener las listas.
- Frecuencia de plazo y documento.
- Factorización de matriz (descomposición de valor singular).
- Extracción de la entidad y resolución.
- Proyecciones del tema del documento.

Sistema de recomendación

- Procedimiento de recomendar interactivo (todos los algoritmos pueden ejecutarse interactivamente en memoria).
- Aplicar un filtro interactivamente para desarrollar recomendaciones para poblaciones específicas.
- Basado en proyectos para soportar carga de usuario, artículos y tablas de clasificación en la memoria.
- k- vecino más cercano, incluyendo coseno, coseno ajustado y correlación de Pearson.
- Factorización de matriz con opciones para las funciones de pérdida, factores de regularización, métodos de optimización y mucho más.

- Agrupamiento de los usuarios y/o elementos utilizando otros atributos, incluyendo el término frecuencia e inverso documento frecuencia pesos.
- Modelos híbridos o conjunto.
- Capacidad de definir un conjunto de retención de los usuarios y las calificaciones para la evaluación del entrenamiento y validación.
- Capacidad para predecir la acción para marcar uno o más usuarios nuevos o una tabla.

Requisitos del sistema

Entorno de servidor

Sistemas operativos

- Red Hat Enterprise Linux 6.
- SuSE Linux Enterprise Server 11.

Hardware

- HP y Dell (con opciones de empaquetado de software y hardware preconfigurado).
- SAS también trabajará con los clientes que deseen utilizar hardware de otros proveedores, incluyendo IBM, Cisco y mucho más.

Hadoop distribuciones compatibles

- CDH Cloudera 4.4 y arriba.
- Hortonworks HDP 2.0.

Nivel intermedio

- Servidor de aplicaciones web SAS (incluido).

Entorno de cliente

Nivel de cliente

- Microsoft Windows (32 bits): Windows 7 x 88-64; Windows 8 x 88-64
- Microsoft Windows (64 bits): Windows 7 x 64 SP1; Windows 8 x

Nivel intermedio

- HP/UX en Itanium
- IBM AIX en energía
- Linux x 64 (x 88-64)
- Microsoft Windows x 64 (x 88-64)
- Solaris (SPARC y x 64)

Navegadores

- Internet Explorer 9 y superiores (modo nativo).
- Firefox 6 y más.
 - Crom.

INDICE ANALITICO

A

- Abrir un informe existente de Power View. 253
- Acceso a un blob usando Azure PowerShell 171
- Administración de Hdinsightcon Powershell 138
- Advanced Anaiytics 87
- Algoritmos de cifrado de datos estándar de la industria 312
- Almacenamiento de blobs de Azure.. 101,103,107, 112,119,129,130,135,137,138,143,145, 150,159,160,162,165,167,177,178,185, 194,201, 207,221,224,238
- Análisis de datos semiestructurados, estructurados y no estructurados 99
- Analysis Services 268,274
- Analytical Decisión Management 28,29,30,31
- Anaiytics....26,27,28,29,31, 73, 75,80,81,85,86, 87,89,100,283, 285,311,313, 314,315,316, 317,320, 323
- Anaiytics con Power System 27
- Apache Hadoop.32, 73, 92, 95, 97,98,99,102,103, 107,150,160,240,246
- Aplicaciones de desarrollo en Hadoop 14
- Aplicaciones típicas del Big Data. 4
- Aprovisionamiento de un cluster de HDInsight.. 108, 138

aprovisionar un cluster 139
Archivos de dirección en almacenamiento de blobs 170
Arquitectura de almacenamiento de HDInsight. 165
Avro 15, 75, 76
AzCopy 130
Azure Explorer 130,132
Azure HDInsight. 102
Azure PowerShell104,108,113,115,121,122,124, 130,131,132,136,137,138,145,169,171, 175,178,181,183,190,193,194,200,201, 207,216, 221,224,225, 226, 228,234,235, 246

B

BigDataaonHenzrnientasdeOrade. 69
Big Data en el Cali Center 9
BIG DATA en SAS 279
Big datasolution with InfoSphere BigInsights and Streams 32
Big Data y el campo de LA investigación 20
Big Data y el Sector Energía. 8
Big Data y la nube 246
Big Data y soluciones Hadoop de SAS. 282
Big Transaction Data. 2
Bhmetrics 2
Business Intelligence .26,28,30,33,34, 71, 79,80, 81,82,87,92,94,104,106,118,236,245,247,

256,274,275,277 C

Carga de datos en HDInsight. 129

Cassandra 15

Ch

Chukwa 15

C

Cloud computing 26

Cloud Computing en Power Systems 34

cluster de HDInsight. 107,108,109,110,111,113,
114,116,118,121,122,123,125,130,134,

138,140,142,168,169,175,177,179,183,

185,193,194,196,200,201, 207, 214,224,

229,230,232, 234,235,236, 237, 240,242

dusters locales de Hadoop y la nube 100

Collaboration and Decisión Support 28

Componentes de una Plataformia de Big Data. 11

Concepto de Big Data 3

Concesión/Revocación del acceso a los servicios
de HTTP 142

Conexión con el almacenamiento de blobs de
Azure 159

Conexión de Excel a HDInsight con Microsoft Hive ODBC Driver 240

Consultas de Hive usando PowerShell 178

Contenedor de almacenamiento de Azure 139

Contenedor de blobs para HDInsight. 168

Contenedor para el almacenamiento de blobs... 168
Contenedor usando Azure PowerShell 169
Crear informe de Power View 252
CRM Analytics 87
Cuenta de Almacenamiento de Azure 109,138

D

DB2 Advanced Enterprise Edition 28,29
Definición, necesidad y características del Big Data
Desarrollo de programas MapReduce 216
Distribución de clusters de HDInsight142,144

E

Ecosistema Hadoop en Azure 104
Ejecución de HDInsight PowerShell 164
Ejecución de proyecto de Oozie 207
Ejecución de trabajos de Pig 157
Ejecutar trabajos de pig. 157
Eliminación de un clúster 141
Emulador de almacenamiento 145,159,160
Emulador de HDInsight 145
Enterprise Performance Management. 84
Enterprise Reporting. 83
Enumeración y visualización de clústeres 141
Enviar un trabajo de MapReduce 142
Envío de trabajos de Hive 144
Envío de trabajos de MapReduce 142

Escenarios de datos de gran tamaño en
HDInsight 106
Excel para visualizar datos de Hadoop 100
Explorador de almacenamiento de Azure. ..132,133
Exportar a PowerPoint desde Power View 254
Extracción y transformación (ELT) 296
Extracción, transformación y carga (ETL) 296

F

Flash Player. 323
Fujo de trabajo de Oozie 194

G

Gráficos circulares 262
Gráficos de barras 264
Gráficos de columnas 264
Gráficos de dispersión y de burbujas 262
Gráficos de líneas 263,264
Gráficos de líneas, barras y columnas 263
Gráficos y otras visualizaciones en Power View..259

H

Hadoop 11, 94, 97, 279,325
Hadoop Common 12,14
Hadoop Distributed File System 12,73,102
Hadoop MapReduce 12,13,147,224, 294

Hadoop Streaming 19

Hadoop Y Big Data en SAS 279

HBase 15,16,74,101

HDFS 12,13,16,19,20, 70, 73, 74, 75,101,102,
103,104,105,106,129,130,136,137,145,
146,147,148,150,151,152,154,157,164,
165,167,168,171,180,191,201,221,222,
223, 224,280,287,294,308,310

HDInsight. 98,103

Hdinsightcon Powershell. 138

HDInsight de Azure 107

HDInsight de Microsoft Azure 98

HDInsight en la nube 100

HDInsight y Hbase 101

Herramientas avanzadas en la nube 38

herramientas de Big Data 6,9, 70

Herramientas de big data en SAS 279

Herramientas de Business Intelligence y
conectores 106

Herramientas de inteligencia empresan'al de
Microsoft 118

Hive. 16, 70, 74,93,94, 95,103,104,105,106,118,
120,136,144,145,146,150,154,155,156,
157,160,168,175,176,177,178,179,180,
181,182,192,193,194,195,200, 202, 208,
215, 235,240,241, 242,243,246, 280, 282,
284, 287,288,294, 311

Hive con HDInsight 175

Human Generated. 3

I

IBM AIX Solution Edition para Cognos 30

IBMAIXSolution Editions para Cognos y SPSS. 29

IBM BLU Acceleration Solution Power Systems
Edition 29

IBM Cognos Business Intelligence 28

IBM DB2 Advanced Workgroup Edition 28,29

IBM DB2 Web Query for i 33

IBM i para Business Intelligence 33

IBM InfoSphere DataStage 29

IBM InfoSphere Information Server 28

IBM POWER SYSTEMS 23

IBM Power Systems Solution Edition for Cloud 37

IBM Power Systems Solution Edition forScale Out
Cloud. 37

IBM PureData System for Operational Analytics.. 31

IBM SmartCloud Entry for Power Systems 37

IBM Solution for Analytics Power Systems Edition27 IBM Solution
for Hadoop Power Systems Edition. 27

IBM SPSS Modeler 28,31,38,48, 67

IBM SPSS Modeler 38,48, 67

IBM SPSS STATISTICS 67

Importación de datos a Excel desde un clúster de
HDInsight. 242

Importación de datos a HDFS. 136
Importación de datos de HDInsight a Excel 237
Importación de datos desde un clúster de
HDInsight 237
Information Discovery 71, 87
informe en Power View para SharePoint Server. 252 Informes de Power
View basados en modelos de
datos 256
Informes de Reporting Services 269
Integration Services 268

J

Jaql 17
Java para HDInsight 216

L

Lucene 17

M

Machine-to-Machine 2
Mapas de Power View 265
MapReduce 11,12,13,14,16,17,18,20, 75, 76,
102,103,104,105,107,108,111,113,114,
115,116,117,118,121,122,123,124,125,
126,129,131,136,142,143,146,147,148,
149,150,152,153,157,158,168,175,176,
183,184,192, 216,217,220, 221, 222,223,

224,229,230, 233,235,288, 308, 310
MapReduce con HDInsight 121
Matrices 267
Message Queue Server 297, 299
MICROSOFT BI SQL Server 247
Microsoft Hive ODBC Driver 241
Microsoft Power Query para Excel. 104,118,236, 237,240
MICROSOFT Y EL BIG DATA 91
Middle Tier 323
modelo multidimensional de Power View. 257

N

NoSQL 15,16, 70, 71, 72, 74, 76, 77, 78,89,101

O

Oozie 17,142,168,192,194,195,200,201,203,
204,207,208,209,211,212,213,214
Oozie con HDInsight 192
Oracle Argus Analytics 86
Oracle BI Mobile App Designer 82,83
Oracle BI Mobile HD 82
Oracle Big Data Appliance 71
Oracle Big Data Connectors 73
Oracle Business Analytics 80, 89
Oracle Business Intelligence Foundation Suite 80
Oracle Communications Data Model 86
Oracle Data Integrator Application Adapter for

Hadoop 73, 74

Oracle ERP 87

Oracle Exadata Dotábase 78

Oracle Exadata Database Machine 78, 79

Oracle Exalytics 71, 79, 87

Oracle exalytics in-memory machine 79

Oracle Financial Analytics 85

Oracle Financial Cióse and Reporting 84

Oracle Health Sciences Clinical Development Analytics 86

Oracle Human Resources Analytics 85

Oracle Linux 72,287,300, 306, 313,322

Oracle Loaderfor Hadoop 73, 74

Oracle Manufacturing Analytics 86

Oracle Marketing Analytics 85

Oracle NoSQL Dotábase 76

Oracle Planning, Budgeting, and Forecasting.. 84

Oracle Procurement & Spend Analytics 85

Oracle Profitability and Cost Management 84

Oracle Project Analytics 85

Oracle R Advanced Analytics for Hadoop 75

Oracle Real-Time Decisions 89

Oracle Retail Customer Analytics 86

Oracle Retail Merchandising Analytics 86

Oracle Sales Analytics 85

Oracle Service Analytics 85

Oracle Social Cloud 89,90

Oracle Social Marketing Cloud Service 90

Oracle Social NetWork. 90
Oracle Social Relationship Management 90
Oracle Strategy Management 84
Oracle Student Information Analytics 86
Oracle Supply Chain and Order Management
Analytics 85
Oracle Tax Analytics 86
Oracle XQueryfor Hadoop 73, 75
ORACLE Y EL BIG DATA 69
Orígenes de Hadoop 97

P

Patrón de detección defraude 5, 7
Patrones de detección del fraude. 4
Patrones de modelado y gestión de riesgo 7
Patrones de Social Media 6
Permisos para Power View 254
PigS, 18,94,95,103,105,146,150,157,158,183,
184,186,189,190,191,192, 246,280,288,
294,308,310
Pig con HDInsight 183
Pig Latín 186
Pig Latín usando PowerShell 189
Plataforma de código abierto HADOOP 11
Power Query para Excel 237, 238
Power View 252
Power View para SharePoint Server 252

PowerPivot para SharePoint 2010 250
PowerPivot para SharePoint 2013 247,248,249
Programa de MapReduce en HDInsight de Azure
229
Programa de recuento de palabras de MapReduce
126
Programa en el emulador 221
Proyecto de Oozie 201

R

Ráele Enterprise Asset Management Analytics 86 Recuento de
palabras 121,122,126,142,147,148,
149,216,217,220,221,223,224,228 Red Hat Enterprise Unix 6..
300,306,313,322,332 Fteporting Services 268

S

SAS Base.... 280,282,288,307,308,310,311,312
SAS Data Management. 282,288,289
SAS DATA MANAGEMENT. 288
SAS Federation Server 282
SAS High-Performance Data Mining 283
SAS High-Performance Optimization 284
SAS High-Performance Statistics 283
SAS High-Performance Text Mining 283
SAS In-Memory Statistics for Hadoop.. 283,313
Sas In-Memory Statistics for Hadoop 325
SAS Scoring Accelerator. 284

SAS Servidor de Federación 302
SAS Visual Analytics 314,317,323
SAS Visual Statistics 283
SAS, Hadoop y EL PROCESO ANALÍTICO 281
SAS/ACCESS Interface to Hadoop 282,284
Servidor analítico de alto rendimiento en memoria 321
Situación de datos en Hadoop 19
Software SAS BYXSE 307
Software System 26
Solución Big Data de Microsoft 92
Soluciones de datos rápidos de Oracle 88
Soluciones Hadoop de SAS 282
SQLServer 96
SQL Server 2014 y EL BIG DATA 245
SQL Server Integration Services 271
Sqoop. 104,105,106,136,192,193,194,200,201, 208,236, 246
SuSE Linux Enterprise Server 11 322

T

Trabajos de MapReduce 152
U
Uso de Hive 176
Uso de Pig 183
Virtualization Foundation Solutions 36
Vistas en un informe de Power View 255

W

Web and Social Media 2

WordCount de MapReduce 111

X

Xplorador de almacenamiento de Azure 130

Z

ZooKeeper 18