IBM SPSS Analytic Server
Versión 3.2.1

Descripción general

Nota
Antes de utilizar esta información y el producto al que da soporte, lea la información del apartado "Avisos" en la página 5.

Información sobre el producto

Esta edición se aplica a la versión 3, release 2, modificación 1 de IBM SPSS Analytic Server y a todos los releases y modificaciones posteriores hasta que se indique lo contrario en nuevas ediciones.

Contenido

Descripción general

IBM® SPSS Analytic Server es una solución de análisis masivo de datos que combina tecnología de IBM SPSS con sistemas de datos masivos, y que permite trabajar con interfaces de usuario de IBM SPSS conocidas para resolver problemas a una escala antes impensable.

Por qué es importante el análisis masivo de datos

El volumen de datos recopilados por las organizaciones crece de forma exponencial; por ejemplo, las empresas financieras y de venta al por menor tienen todas las transacciones de clientes de un año (o dos años, o diez), los proveedores de telecomunicaciones tienen los registros de datos de llamadas (CDR) y lecturas de sensores de dispositivos, y las empresas de internet tienen los resultados de los rastreos web.

Un análisis masivo de datos es necesario cuando existe:

- Un gran volumen de datos (terabytes, petabytes o exabytes), sobre todo cuando es una mezcla de datos estructurados y no estructurados.
- Datos que cambian/se acumulan con rapidez.

El análisis masivo de datos también es de ayuda cuando:

- Se construye un gran número de modelos (del orden de miles).
- Los modelos se construyen/renuevan con frecuencia.

Retos

Las mismas organizaciones que recopilan grandes volúmenes de datos suelen tener dificultades a la hora de utilizarlos, por una serie de razones:

- La arquitectura de los productos analíticos tradicionales no está pensada para la computación distribuida, y
- los algoritmos estadísticos existentes no están diseñados para trabajar con cantidades masivas de datos (tales algoritmos esperan que los datos les lleguen, pero cuesta mucho mover datos masivos), por tanto
- el análisis de datos masivos con tecnología puntera requiere nuevas habilidades y un conocimiento a fondo de los sistemas de datos masivos. Muy pocos analistas poseen estas habilidades.
- Las soluciones residentes en memoria son aptas para problemas de tamaño medio, pero no escalan bien a datos realmente masivos.

Solución

Analytic Server proporciona:

- Una arquitectura centrada en datos que saca partido de sistemas de datos masivos tales como Hadoop Map/Reduce con datos en HDFS.
- Una interfaz definida para incorporar nuevos algoritmos estadísticos diseñados para ir a los datos.
- Conocidas interfaces de usuario de IBM SPSS que ocultan los detalles de los entornos de datos masivos, de modo que el analista pueda centrarse en el análisis de los datos.
- Una solución escalable a problemas de cualquier tamaño.

Arquitectura

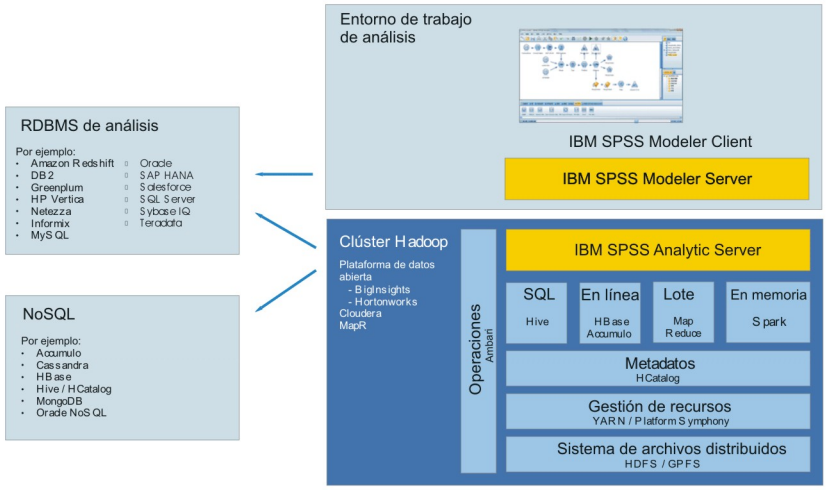

Figura 1. Arquitectura

Analytic Server se sitúa entre una aplicación cliente y una nube Hadoop. Suponiendo que los datos residan en la nube, la forma de trabajar con Analytic Server sería a grandes rasgos:

1. Se definen los orígenes de datos de Analytic Server que dan acceso a los datos de la nube.
2. Se define el análisis que se desea realizar en la aplicación cliente. En el release actual, la aplicación de cliente es IBM SPSS Modeler.
3. Cuando se ejecuta el análisis, la aplicación cliente envía una solicitud de ejecución de Analytic Server.
4. Analytic Server organiza el trabajo para que ejecute en la nube de Hadoop e informa de los resultados a la aplicación cliente.
5. Los resultados pueden utilizarse para definir análisis adicionales, con lo que se repetiría el ciclo.

Spark y Analytic Server

Analytic Server se integra con Apache Spark para aumentar el rendimiento.

Cuándo se utiliza y cuándo no se utiliza Spark

Si Spark está instalado como un servicio Ambari en el clúster de Hadoop, Analytic Server lo utiliza para procesar trabajos de datos masivos. Se aplican las siguientes directrices para determinar cuándo no se utiliza Spark:

1. Si el juego de datos tiene menos de 128 MB, Analytic Server utiliza la función incluida MapReduce en la JVM de Analytic Server y no utiliza Spark o el clúster de Hadoop.
2. Si Spark no está instalado en el clúster, Analytic Server utiliza MapReduce v2.
3. Analytic Server utiliza MapReduce v2 para crear modelos PSM. Cuando un trabajo finaliza con una creación de modelo PSM, Analytic Server utiliza Spark para procesar el trabajo a través de todos los pasos que conducen a la creación del modelo, a continuación, lo graba en disco y finalmente, utiliza MapReduce para crear el modelo PSM. Por ejemplo, si un trabajo incluye una unión seguida de una creación de modelo PSM, la unión se ejecuta en Spark y PSM se ejecuta en los datos unidos en MapReduce.

Cómo se utiliza Spark

Cuando se inicia el servicio Analytic Server y descubre que Spark está disponible, inicializa un "Trabajo Spark Hadoop" que permite establecer la comunicación entre tareas distribuidas a través del clúster. Este trabajo se ejecuta siempre y cuando se ejecute el servicio Analytic Server y se utiliza para todas las ejecuciones de Analytic Server. Este método mejora el rendimiento respecto a la organización de varios trabajos MapReduce Hadoop, pero elimina la sobrecarga que supone volver a cargar todos los componentes de Analytic Server para cada trabajo Hadoop.

Spark es capaz de ejecutar trabajos MapReduce. Esto permite que Analytic Server utilice algoritmos Spark "nativos", como por ejemplo, join, sort y union siempre que estén disponibles. Al mismo tiempo, Analytic Server puede ejecutar algoritmos SPSS Map and Reduce existentes en Spark y sin utilizar directamente la API de Hadoop.

Novedades de la versión 3.2.1

Versión 3.2.1

Plataforma

Soporte para Hadoop Data Platform (HDP) 3.0 y 3.1.

Soporte para Cloudera 6.0 y 6.1.

Función de rango

La función **rank** se utiliza para dividir el conjunto de datos de entrada en particiones separadas y generar un campo nuevo que muestra el rango de cada fila de la partición. La característica es parecida a las funciones Hive **rank()**, **dense_rank()** y **row_number()**.

Integración de UDF de Hive

Se ha introducido nuevas funciones UDF de Hive. Una vez registradas las UDF de Hive en la base de datos de Hive, Analytic Server puede utilizar las nuevas funciones UDF para llevar a cabo su integración.

Para obtener la información más actualizada sobre los requisitos del sistema, utilice los informes detallados de requisitos del sistema en el sitio de soporte técnico de IBM: http://publib.boulder.ibm.com/infocenter/prodguid/v1r0/clarity/softwareReqsForProduct.html. En esta página:

1. Especifique SPSS Analytic Server como nombre de producto y pulse **Search**.
2. Seleccione la versión deseada y el ámbito del informe y, a continuación, haga clic en **Submit**.

Avisos

Esta información se ha desarrollado para productos y servicios que se comercializan en los EE.UU. Este material puede estar disponible en IBM en otros idiomas. Sin embargo, es probable que sea necesario que disponga de una copia del producto o versión del producto en dicho idioma para tener acceso.

Es posible que IBM no ofrezca en otros países los productos, servicios o características que se describen en este documento. Póngase en contacto con el representante local de IBM, que le informará sobre los productos y servicios disponibles actualmente en su área. Las referencias a programas, productos o servicios de IBM no pretenden establecer ni implicar que sólo puedan utilizarse dichos productos, programas o servicios de IBM. En su lugar, se puede utilizar cualquier producto, programa o servicio equivalente que no infrinja ninguno de los derechos de propiedad intelectual de IBM. No obstante, es responsabilidad del usuario evaluar y verificar el funcionamiento de cualquier producto, programa o servicio que no sea de IBM.

IBM puede tener patentes o solicitudes de patente pendientes que cubran la materia descrita en este documento. El suministro de este documento no le otorga ninguna licencia sobre dichas patentes. Puede enviar consultas sobre licencias, por escrito, a:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
EE.UU.

Si tiene consultas sobre licencias relacionadas con información DBCS (de doble byte), póngase en contacto con el Departamento de propiedad intelectual de IBM en su país o envíelas, por escrito, a:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokio 103-8510, Japón

INTERNATIONAL BUSINESS MACHINES CORPORATION PROPORCIONA ESTA PUBLICACIÓN "TAL CUAL" SIN GARANTÍAS DE NINGÚN TIPO, NI EXPLÍCITAS NI IMPLÍCITAS, INCLUIDAS, AUNQUE SIN LIMITARSE A, LAS GARANTÍAS DE NO CONTRAVENCIÓN, COMERCIALIZACIÓN O ADECUACIÓN A UN PROPÓSITO DETERMINADO. Algunas jurisdicciones no permiten la renuncia a las garantías explícitas o implícitas en determinadas transacciones; por lo tanto, es posible que esta declaración no sea aplicable en su caso.

Es posible que esta información contenga imprecisiones técnicas o errores tipográficos. Periódicamente se realizan cambios en la información que aquí se presenta; estos cambios se incorporarán en las nuevas ediciones de la publicación. IBM puede realizar en cualquier momento mejoras o cambios en los productos o programas descritos en esta publicación sin previo aviso.

Las referencias hechas en esta publicación a sitios web que no son de IBM se proporcionan sólo para la comodidad del usuario y no constituyen un aval de esos sitios web. Los materiales de dichos sitios web no forman parte del material de este producto de IBM y el usuario es el único responsable del uso que haga de ellos.

IBM puede utilizar o distribuir la información que se le proporcione del modo que considere adecuado sin incurrir por ello en ninguna obligación con el remitente.

Los titulares de licencias de este programa que deseen obtener información sobre el mismo con el fin de permitir: (i) el intercambio de información entre programas creados independientemente y otros programas (incluido éste) y (ii) el uso mutuo de la información que se ha intercambiado, deben ponerse en contacto con:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
EE.UU.

Dicha información puede estar disponible, sujeta a los términos y condiciones correspondientes, incluyendo, en algunos casos, el pago de una tarifa.

El programa bajo licencia que se describe en este documento y todo el material bajo licencia disponible los proporciona IBM bajo los términos de las Condiciones Generales de IBM, Acuerdo Internacional de Programas Bajo Licencia de IBM o cualquier acuerdo equivalente entre las partes.

Los ejemplos de datos de rendimiento y de clientes citados se presentan solamente a efectos ilustrativos. Los resultados de rendimiento reales pueden variar en función de las configuraciones específicas y de las condiciones de funcionamiento.

La información relativa a productos que no son de IBM se ha obtenido de los proveedores de dichos productos, de los anuncios publicados y de otras fuentes de información pública. IBM no ha comprobado estos productos y no puede confirmar la precisión de su rendimiento, compatibilidad ni contemplar ninguna otra reclamación relacionada con los productos que no son de IBM. Las preguntas relacionadas con las prestaciones de productos que no son de IBM deben dirigirse a los proveedores de dichos productos.

Las declaraciones relativas a la dirección o intenciones futuras de IBM están sujetas a cambio o retirada sin previo aviso y representan únicamente objetivos y metas.

Todos los precios de IBM que se muestran son precios actuales recomendados por IBM de venta al público y están sujetos a cambios sin notificación previa. Los precios en los distribuidores pueden variar.

Esta información es sólo para fines de planificación. Dicha información está sujeta a cambios antes de que los productos descritos estén disponibles.

Esta información contiene ejemplos de datos e informes utilizados en operaciones empresariales diarias. Para ilustrarlas lo mejor posible, los ejemplos contienen nombres de personas, compañías, marcas y productos. Todos estos nombres son ficticios y cualquier parecido con personas o empresas comerciales reales es pura coincidencia.

LICENCIA DE DERECHOS DE AUTOR:

Esta información contiene ejemplos de datos e informes utilizados en operaciones empresariales diarias. Para ilustrarlas lo mejor posible, los ejemplos contienen nombres de personas, compañías, marcas y productos. Todos estos nombres son ficticios y cualquier parecido con personas o empresas comerciales reales es pura coincidencia.

Cada copia o cada parte de estos programas de ejemplo, o trabajos derivados, debe incluir un aviso de copyright como se indica a continuación:

Marcas registradas

IBM, el logotipo de IBM e ibm.com son marcas registradas o marcas comerciales registradas de International Business Machines Corp., registrada en muchas jurisdicciones en todo el mundo. Otros nombres de productos y servicios podrían ser marcas registradas de IBM u otras compañías. En Internet hay disponible una lista actualizada con las marcas registradas de IBM, en "Copyright and trademark information", en la dirección www.ibm.com/legal/copytrade.shtml.

Adobe, el logotipo de Adobe, PostScript y el logotipo de PostScript son marcas registradas o marcas comerciales de Adobe Systems Incorporated en los Estados Unidos y/o en otros países.

IT Infrastructure Library es una marca registrada de la Agencia central de informática y telecomunicaciones que ahora es parte de la Cámara de Comercio.

Intel, el logotipo de Intel, Intel Inside, el logotipo de Intel Inside, Intel Centrino, el logotipo de Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium y Pentium son marcas registradas de Intel Corporation o de sus subsidiarias en EE.UU. y en otros países.

Linux es una marca registrada de Linus Torvalds en Estados Unidos y/o en otros países.

Microsoft, Windows, Windows NT y el logotipo de Windows son marcas registradas de Microsoft Corporation en los Estados Unidos, otros países o ambos.

ITIL es una marca registrada, y una marca de comunidad registrada de The Minister for the Cabinet Office, y está registrada en U.S. Patent and Trademark Office.

UNIX es una marca registrada de The Open Group en Estados Unidos y en otros países.

Cell Broadband Engine es una marca comercial de Sony Computer Entertainment, Inc. en Estados Unidos, otros países o ambos y se utiliza bajo licencia.

Linear Tape-Open, LTO, el logotipo de LTO, Ultrium y el logotipo de Ultrium son marcas comerciales de HP, IBM Corp. y Quantum en Estados Unidos y otros países.

Impreso en España