IBM SPSS Analytic Server
versión 1

Guía del usuario

Nota
Antes de utilizar esta información y el producto que soporta, lea la información de "Avisos" en la página 17.

Información sobre el producto

Esta edición se aplica a la versión 1, release 0, modificación 0 de IBM SPSS Analytic Server y a todos los releases y modificaciones posteriores mientras no se indique lo contrario en nuevas ediciones.

Contenido

Capítulo 1. 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

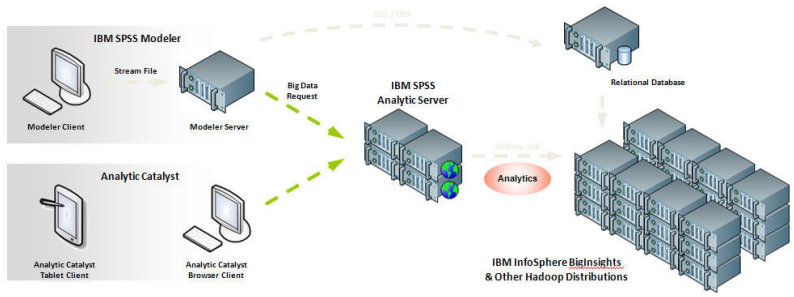

Figura 1. Arquitectura

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, las aplicaciones cliente son IBM SPSS Modeler y IBM SPSS Analytic Catalyst.
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.

Capítulo 2. Integración de SPSS Modeler

SPSS Modeler es un entorno de trabajo de minería de datos con un enfoque visual al análisis. Cada acción individual de un trabajo, desde el acceso a un origen de datos hasta la fusión de registros, pasando por la generación de un nuevo archivo o de un modelo, se representa mediante un nodo en el lienzo. Dichas acciones se enlazan entre sí para formar una secuencia analítica.

Para construir una secuencia de SPSS Modeler que pueda ejecutarse en HDFS, se empieza con un nodo "Origen de Analytic Server". SPSS Modeler inyectará tanta secuencia como sea posible a Analytic Server y luego, en caso de ser necesario, extraerá un subconjunto de los registros de HDFS para terminar de ejecutar la secuencia "en local" en el servidor de SPSS Modeler. El número máximo de registros que se descarga SPSS Modeler puede configurarse en las "Propiedades de una secuencia de Analytic Server" en la página 4.

Si el análisis finaliza con registros que se han vuelto a escribir en HDFS, termine la secuencia con un nodo "Exportación de Analytic Server" en la página 4.

Origen de Analytic Server

El origen de Analytic Server permite ejecutar una secuencia en un sistema de archivos distribuido de Hadoop (HDFS en sus siglas inglesas). La información de un origen de datos de Analytic Server puede proceder de diversos lugares, entre los que se incluyen:

- archivos de texto en HDFS
- bases de datos
- HCatalog

Normalmente, una secuencia con un origen de Analytic Server se ejecuta en HDFS. Sin embargo, si una secuencia contiene un nodo cuya ejecución no esté soportada en HDFS, se "inyectará de vuelta" tanta secuencia como sea posible a Analytic Server, y entonces SPSS Modeler Server intentará procesar el resto de la secuencia. Los bloques de datos muy grandes tendrán que dividirse en submuestras colocando, por ejemplo, un nodo Muestra en la secuencia.

Origen de datos. Suponiendo que el administrador de SPSS Modeler Server haya establecido una conexión, seleccione el origen de datos que contenga los datos que desee utilizar. Un origen de datos contiene los archivos y los metadatos asociados a ese origen. Pulse **Seleccionar** para visualizar una lista de los orígenes de datos disponibles. Consulte el tema "Selección de un origen de datos en Analytic Server" para obtener información adicional.

Si necesita crear un origen de datos o editar uno existente, pulse **Lanzar editor de orígenes de datos...**. Consulte "Orígenes de datos" en la página 9 para obtener detalles relativos a la creación y edición de orígenes de datos.

Selección de un origen de datos en Analytic Server

La tabla Orígenes de datos muestra una lista de los orígenes de datos disponibles. Seleccione el origen que desee y pulse **Aceptar**.

Pulse **Mostrar propietario** para visualizar el propietario del origen de datos.

Filtrar por permite filtrar el listado de orígenes de datos por **Palabra clave**, aplicando los criterios de filtro al nombre del origen de datos, a su descripción o al **Propietario**. Como criterio de búsqueda puede

especificarse una combinación de cadenas, valores numéricos y caracteres comodín (%). En la cadena de búsqueda se distingue entre mayúsculas y minúsculas. Pulse **Renovar** para actualizar la tabla Orígenes de datos.

Exportación de Analytic Server

La exportación de Analytic Server permite escribir datos desde un análisis a un origen de datos de Analytic Server existente. Este origen puede ser, por ejemplo, archivos de texto en un sistema de archivos distribuido de Hadoop (HDFS) o una base de datos.

una secuencia con un nodo Exportar de Analytic Server suele comenzar también con nodos Origen de Analytic Server, y se envía a Analytic Server y se ejecuta en HDFS. De forma alternativa, una secuencia con orígenes de datos "locales" puede terminar con un nodo Exportar de Analytic Server para cargar bloques de datos relativamente pequeños (de no más de 100.000 registros) para usarlos en Analytic Server.

Origen de datos. Seleccione el origen de datos que contenga los datos que desee. Un origen de datos contiene los archivos y los metadatos asociados a ese origen. Pulse **Seleccionar** para visualizar una lista de los orígenes de datos disponibles. Consulte el tema “Selección de un origen de datos en Analytic Server” en la página 3 para obtener información adicional.

Si necesita crear un origen de datos o editar uno existente, pulse **Lanzar editor de orígenes de datos...**. Consulte “Orígenes de datos” en la página 9 para obtener detalles relativos a la creación y edición de orígenes de datos.

Modo. Seleccione **Añadir** para añadir al origen de datos existente, o **Sustituir** para sustituir el contenido del origen de datos.

Generar un nodo Importación para estos datos. Selecciónelo para generar un nodo origen para los datos como se han exportado al origen de datos especificado. Este nodo se añadirá al lienzo de la secuencia.

Propiedades de una secuencia de Analytic Server

Los siguientes valores proporcionan una serie de opciones para trabajar con Analytic Server.

Número máximo de registros que se procesan fuera de Analytic Server
: Especifica el número máximo de registros que se importan en el servidor de SPSS Modeler procedentes de un origen de datos de Analytic Server.

Notificación cuando un nodo no puede procesarse en Analytic Server
: Este valor determina lo que ocurre cuando una secuencia que se va a enviar a Analytic Server contiene un nodo que no puede procesarse en Analytic Server. Especifica si debe emitirse un aviso y continuar procesándose la secuencia, o si se genera un error y se detiene el procesamiento.

Valores de almacenamiento de modelos divididos

Almacenar modelos divididos por referencia en Analytic Server cuando el tamaño (MB) del modelo sobrepase
: Los nuggets (fragmentos) de modelo se almacenan habitualmente como parte de la secuencia. Los modelos divididos en muchas partes pueden dar lugar a nuggets de gran tamaño, y el trasiego de dichos nuggets entre la secuencia y Analytic Server puede tener un impacto negativo en el rendimiento. Para resolver este problema, cuando un modelo dividido sobrepasa el tamaño especificado, se almacena en Analytic Server, y el nugget que está en SPSS Modeler contiene una referencia al modelo.

Carpeta predeterminada para almacenar modelos por referencia en Analytic Server una vez finalizada la ejecución
Especifica la ruta predeterminada donde se almacenan los modelos divididos en Analytic Server. La ruta debe comenzar con un nombre válido de proyecto de Analytic Server.

Carpeta en la que almacenar los modelos ascendidos
Especifica la ruta predeterminada en la que se desean almacenar los modelos "ascendidos". Un modelo ascendido no se limpia cuando termina la sesión de SPSS Modeler.

Nodos soportados

La ejecución de muchos nodos de SPSS Modeler está soportada en HDFS, aunque es posible exista alguna diferencia en la ejecución de determinados nodos, mientras que otros ni siquiera están soportados en la actualidad. Este tema detalla en nivel de soporte actual.

General

- Analytic Server no acepta algunos caracteres que normalmente se aceptan en el interior de un nombre de campo entrecomillado de Modeler.
- Para que una secuencia de Modeler ejecute en Analytic Server, debe empezar con uno o más nodos Origen de Analytic Server y terminar con un único nodo de modelado o de exportación de Analytic Server; se permiten las uniones, pero no las bifurcaciones.
- Se recomienda definir el almacenamiento de destinos continuos como real en lugar de entero. Los modelos de puntuación siempre escriben valores reales en los archivos de datos de salida de los destinos continuos, mientras que el modelo de datos de salida de las puntuaciones se ajusta al almacenamiento del destino. Por tanto, si un destino continuo tiene un almacenamiento entero, se producirá una discordancia entre los valores escritos y el modelo de datos de las puntuaciones, y dicha discordancia provocará errores cuando se intenten leer los datos puntuados.

Origen

- Una secuencia que comience con cualquier cosa que no sea un nodo de origen de Analytic Server ejecutará en local.

Operaciones de registro

Se soportan todas las operaciones de Registro. A continuación se detalla la funcionalidad de estos nodos.

Seleccionar

- Soporta el mismo conjunto de funciones que soporta el nodo Derivar (ver abajo).

Muestrear

- No se soporta el muestreo a nivel de bloque.
- No se soportan los métodos de muestreo complejos.

Agregación

- Las claves contiguas no están soportadas.
- Las estadísticas de orden (mediana, primer cuartil, tercer cuartil) no están soportadas.

Ordenar

- La pestaña Optimización no está soportada.

Fusionar

- La fusión por Orden no está soportada.
- La fusión por Condición no está soportada.
- La pestaña Optimización no está soportada.

- En la actualidad no se soporta colocar un nodo Ordenar o un nugget de modelo entre un nodo Origen de Analytic Server y un nodo Fusionar. Normalmente es posible especificar un nodo Seleccionar para sustituir la funcionalidad del nodo Muestrear.
- Analytic Server no realiza uniones sobre claves de cadena vacías, es decir, si una de las claves por las que se está efectuando la fusión contiene cadenas vacías, los registros que contengan la cadena vacía se descartarán de la salida fusionada.
- Las operaciones de fusión son relativamente lentas. Si se dispone de espacio suficiente en HDFS, puede ser mucho más rápido fusionar una única vez los orígenes de datos y utilizar el origen fusionado en las secuencias posteriores en lugar de fusionar los orígenes de datos en cada secuencia.

Operaciones de campo

Se soportan los nodos Preparación de datos automática, Tipo, Filtrar, Derivar, Ensamblar, Relleno, Reclasificar, Intervalos, Análisis de RFM, Partición, Marcas, Reestructurar y Reorg. campos. A continuación se detalla la funcionalidad de estos nodos.

Preparación automática de datos

- No se soporta el entrenamiento del nodo. La aplicación de transformaciones en un nodo Preparación automática de datos entrenado a datos nuevos está soportada.

Tipo

- La columna Comprobación no está soportada.
- La pestaña Formato no está soportada.

Derivar

- Se soportan todas las funciones de Derivar, a excepción de las funciones de secuencia.
- Los campos divididos no pueden derivarse en la misma secuencia que los utiliza como divisiones. En tal caso será necesario crear dos secuencias: una que derive el campo dividido y una que utilice el campo como división.
- Un campo de distintivo no puede utilizarse por sí solo en una comparación, es decir, if (campoDistintivo) then ... endif provocará un error. Esto se resuelve utilizando if (campoDistintivo=valorVerdadero) then ... endif
- Cuando se utilice el operador ** para comparar resultados en Modeler, se recomienda especificar el exponente como número real, por ejemplo, x**2.0 en lugar de x**2.

Relleno

- Soporta el mismo conjunto de funciones que soporta el nodo Derivar (ver arriba).

Análisis de RFM

- No se soporta la opción "Mantener en el actual" para manejar empates. La actualidad de RFM, la frecuencia y las puntuaciones monetarias no siempre coincidirán con las que Modeler calcula a partir de los mismos datos. Los rangos de puntuación serán los mismos, pero las asignaciones de puntuación (números de intervalo) pueden diferir en una.

Gráficos

Se soportan todos los nodos Gráfico.

Modelado

Se soporta un número limitado de modos Modelado, incluidos Lineal, Red neuronal, C&RT, Chaid y Quest. A continuación siguen algunos comentarios adicionales sobre la funcionalidad de estos nodos.

Lineal

- No se soporta el entrenamiento continuado de modelos de PSM existentes.
- El objetivo Generación de Modelo estándar solo se recomienda si los campos se definen de modo que el número de registros de cada división no sea demasiado grande, donde

la definición de "demasiado grande" depende de la potencia de los nodos individuales del clúster de Hadoop. Por contra, también debe procurarse que las divisiones no se definan tan pequeñas que contengan demasiados pocos registros como para construir un modelo.

- No se soporta el objetivo de Potenciación.
- No se soporta el objetivo de Agregación autodocimante.
- No se recomienda el objetivo de Conjuntos de datos muy grandes cuando hay pocos registros; con frecuencia no se generará un modelo, o el modelo generado estará degradado. Es posible que también se presenten problemas si los registros de entrada están ordenados de alguna forma sistemática que infrinja los presupuestos de aleatoriedad que hay tras los algoritmos utilizados de creación de modelos de conjuntos.
- No se soporta la preparación de datos automática. Esto puede dar lugar a problemas cuando se intente construir un modelo a partir de datos con muchos valores ausentes; normalmente dichos datos se imputarían como parte de la preparación de datos automática. Una solución consistiría en utilizar un modelo de árbol o una red neuronal con el valor Avanzado seleccionado para imputar los valores ausentes.
- La estadística de precisión no se calcula para los modelos divididos.

Red neuronal

- No se soporta el entrenamiento continuado de modelos de PSM o estándar existentes.
- El objetivo Generación de Modelo estándar solo se recomienda si los campos se definen de modo que el número de registros de cada división no sea demasiado grande, donde la definición de "demasiado grande" depende de la potencia de los nodos individuales del clúster de Hadoop. Por contra, también debe procurarse que las divisiones no se definan tan pequeñas que contengan demasiados pocos registros como para construir un modelo.
- No se soporta el objetivo de Potenciación.
- No se soporta el objetivo de Agregación autodocimante.
- No se recomienda el objetivo de Conjuntos de datos muy grandes cuando hay pocos registros; con frecuencia no se generará un modelo, o el modelo generado estará degradado. Es posible que también se presenten problemas si los registros de entrada están ordenados de alguna forma sistemática que infrinja los presupuestos de aleatoriedad que hay tras los algoritmos utilizados de creación de modelos de conjuntos.
- Cuando falten muchos valores en los datos, utilice el valor Avanzado para imputar los valores que faltan.
- La estadística de precisión no se calcula para los modelos divididos.

Árbol C&R, CHAID y Quest

- No se soporta el entrenamiento continuado de modelos de PSM existentes.
- El objetivo Generación de Modelo estándar solo se recomienda si los campos se definen de modo que el número de registros de cada división no sea demasiado grande, donde la definición de "demasiado grande" depende de la potencia de los nodos individuales del clúster de Hadoop. Por contra, también debe procurarse que las divisiones no se definan tan pequeñas que contengan demasiados pocos registros como para construir un modelo.
- No se soporta el objetivo de Potenciación.
- No se soporta el objetivo de Agregación autodocimante.
- No se recomienda el objetivo de Conjuntos de datos muy grandes cuando hay pocos registros; con frecuencia no se generará un modelo, o el modelo generado estará degradado. Es posible que también se presenten problemas si los registros de entrada

están ordenados de alguna forma sistemática que infrinja los presupuestos de aleatoriedad que hay tras los algoritmos utilizados de creación de modelos de conjuntos.

- No se soportan las sesiones interactivas.
- La estadística de precisión no se calcula para los modelos divididos.

Puntuación de modelos

Se soportan los siguientes nuggets de modelo para la puntuación: Red neural, C&RT, Chaid, Quest, Logística, Regresión, Genlin, Clúster bietápico, C5, Redes bayesianas, SVM, R, y Minería de textos.

- No se puntuarán las propensiones brutas ni las ajustadas. A modo de solución alternativa, puede lograrse el mismo efecto calculando manualmente la propensión bruta utilizando un nodo Derivar con la siguiente expresión: if 'valor-pronosticado' == 'valor-de-interés' then 'prob-de-ese-valor' else 1-'prob-de-ese-valor' endif
- Al puntuar un modelo, Analytic Server no comprueba si todos los campos utilizados en el modelo están presentes en el conjunto de datos, así que asegúrese de que sea así antes de ejecutar en Analytic Server

Salida Se soportan los nodos Matriz, Análisis, Auditar datos, Transformar, Estadísticas y Medias.

Exportar

Una secuencia puede comenzar con un nodo origen de Analytic Server y terminar con un nodo de exportación distinto del nodo de exportación de Analytic Server, pero los datos se moverán de HDFS a SPSS Modeler Server y por último a la ubicación de exportación.

Capítulo 3. Consola de Analytic Server

Analytic Server proporciona una interfaz de cliente ligero para gestionar orígenes de datos y proyectos. Puede accederse a ella de dos formas:

- A través de un nodo Origen o Exportación de Analytic Server, que abre la consola en un navegador.
- Directamente a través de un navegador.
  1. Escriba el URL de Analytic Server en la barra de direcciones del navegador. Puede obtenerlo del administrador del servidor.
  2. Escriba el nombre de usuario con el que iniciar sesión en el servidor.
  3. Escriba la contraseña asociada al nombre de usuario especificado.

Una vez iniciada la sesión, el acordeón predeterminado es el acordeón Orígenes de datos.

Navegación por la consola

La consola de Analytic Server tiene cuatro componentes:

- La cabecera muestra el nombre de producto y el nombre del usuario que ha iniciado la sesión actual.
- La columna izquierda muestra los acordeones o agrupaciones funcionales disponibles El acordeón seleccionado determina lo que se muestra en el área de contenidos.
- El área de contenidos. Muestra los controles asociados al acordeón seleccionado en ese momento. Los detalles del contenido de cada acordeón se muestran en las secciones que vienen a continuación.
- El pie de página contiene el enlace de cierre de sesión y muestra la versión instalada de Analytic Server.

Orígenes de datos

Un origen de datos es una colección de registros más un modelo de datos que define un conjunto de datos de análisis. El origen de los registros puede ser un archivo (texto delimitado, texto de ancho fijo, Excel) en HDFS, una base de datos o HCatalog. El modelo de datos de define todos los metadatos (nombres de campo, almacenamiento, nivel de medida, etc.) necesarios para analizar los datos. Los propietarios de un origen de datos pueden otorgar o restringir el acceso a dicho origen de datos.

Columna izquierda

La columna izquierda muestra los orígenes de datos existentes bajo la cabecera del acordeón.

- Seleccione un origen de datos para visualizar sus detalles en el área de contenidos y editar sus propiedades. Cuando se escribe en el área de búsqueda, se filtra el listado mostrándose solo aquellos orígenes de datos que contengan la cadena de búsqueda en el nombre.
- Pulse el botón **Nuevo origen de datos** para crear un origen de datos con el nombre y el tipo especificados en el diálogo **Añadir nuevo origen de datos**.
  - En los nombres de orígenes de datos se distingue entre mayúsculas y minúsculas. Se omiten los espacios en blanco que pueda haber al comienzo o al final. Determinados nombres se rechazarán para ofrecer protección frente a inyecciones SQL.
  - Los tipos de contenido disponibles son Archivo, Base de datos y HCatalog.

    Nota: Una vez seleccionado, el tipo de contenido no podrá editarse.
- Pulse el botón **Suprimir origen de datos** para eliminar el origen de datos. Esta acción no afecta en modo alguno a los archivos asociados al origen de datos.

Área de contenidos

El área de contenidos se divide en varias secciones, dependiendo del tipo de contenido del origen de datos. Una vez especificados los valores de configuración del origen de datos, pulse Vista previa y metadatos para completar el origen de datos.

Propiedades del origen de datos
Valores comunes a todos los tipos de contenido.

Nombre
Campo de texto editable que muestra el nombre del origen de datos. En los nombres de orígenes de datos se distingue entre mayúsculas y minúsculas. Se omiten los espacios en blanco que pueda haber al comienzo o al final. Puede que determinados nombres se rechacen para ofrecer protección frente a inyecciones SQL.

Descripción
Campo de texto editable que proporciona un texto descriptivo del origen de datos.

Es público
Casilla de verificación que indica si cualquiera puede ver el proyecto (cuando está marcada) o si deben añadirse explícitamente usuarios y grupos a la lista de propietarios (cuando está sin marcar).

Compartición
La propiedad de un origen de datos puede compartirse añadiendo usuarios y grupos en calidad de autores.

- Cuando se escribe en el cuadro de texto, se filtran usuarios y grupos que tengan la cadena de búsqueda en el nombre. Pulse el botón **Añadir participante** para añadirlos a la lista de autores.
- Para eliminar un autor, seleccione un usuario o grupo en la lista de autores y pulse el botón **Eliminar psrticipante**.

Nota: Los administradores tendrán acceso de lectura y escritura en todos los orígenes de datos, independientemente de que aparezcan o no en la lista de autores.

Entrada de archivo
Valores propios de la definición de orígenes de datos con tipo de contenido archivo.

Visor de archivos

Muestra los archivos disponibles para su inclusión en el origen de datos. Seleccione el modo **Proyectos** para visualizar archivos dentro de la estructura del proyecto de Analytic Server, o **HDFS** para visualizar el resto del sistema de archivos distribuido de Hadoop. Puede examinarse cualquiera de las dos estructuras de carpetas, pero HDFS no es editable en modo alguno, y la estructura de carpetas de Analytic Server solo es editable dentro de los proyectos definidos. Es decir, no se pueden añadir archivos, ni crear carpetas ni suprimir elementos en el nivel raíz del modo **Proyectos**. Para crear, editar o suprimir un proyecto, utilice el Acordeón Proyectos.

- Cuando se pulsa el botón **Cargar archivo en HDFS**, se carga un archivo en el proyecto/subdirectorio actual.
- Cuando se pulsa el botón **Crear carpeta**, se crea una carpeta bajo la carpeta actual, con el nombre especificado en el diálogo Nombre de la nueva carpeta.
- Cuando se pulsa el botón **Descargar archivo al sistema de archivos local**, se descargan los archivos seleccionados al sistema de archivos local.
- Cuando se pulsa el botón **Suprimir archivo(s) seleccionado(s)**, se eliminan los archivos/carpetas seleccionados.

Archivos incluidos en la definición del origen de datos
Utilice el botón de mover para añadir los archivos seleccionados al origen de datos, o

para eliminarlos de él. Por cada archivo seleccionado en el origen de datos, pulse Valores para definir las especificaciones de lectura del archivo.

Salida de archivo
A los orígenes de datos con tipo de contenido de archivo se les puede añadir la salida de secuencias ejecutadas en Analytic Server. Seleccione **Hacer modificable** para habilitar la adición y seleccione la carpeta de salida donde se escriben los nuevos archivos.

Selecciones de base de datos

Especifique los parámetros de conexión de la base de datos que contiene el contenido de registros.

Base de datos
Selecciona el tipo de base de datos a la que se desea conectar. Se elige entre: DB2, Oracle, SQL Server, TeraData o Netezza.

Dirección del servidor
Especifica el URL del servidor en el que se aloja la base de datos.

Puerto del servidor
El número de puerto por el que escucha la base de datos.

Nombre de la base de datos.
Nombre de la base de datos a la que desea conectarse.

Nombre de usuario
Si la base de datos está protegida mediante contraseña, especifica el nombre de usuario.

Contraseña
Si la base de datos está protegida mediante contraseña, especifica la contraseña.

Nombre de la tabla
Especifica el nombre de la tabla de base de datos que desee utilizar.

Máximo número de lecturas simultáneas

Salida de base de datos
A los orígenes de datos con tipo de contenido de base de datos se les puede añadir la salida de secuencias ejecutadas en Analytic Server. Seleccione **Hacer modificable** para habilitar la adición y seleccione la tabla de base de datos de salida donde se escriben los datos de salida.

Selecciones de HCatalog
Especifica los parámetros de acceso a los datos gestionados en Apache HCatalog.

Base de datos
Nombre de la base de datos de HCatalog.

Nombre de la tabla
Especifica el nombre de la tabla de base de datos que desee utilizar.

Filtro El filtro de partición de la tabla, si la tabla se ha creado como tabla particionada.

Esquema de HCatalog
Muestra la estructura de la tabla especificada. HCatalog puede soportar un conjunto de datos altamente estructurado. Para definir un origen de datos de Analytic Server sobre dichos datos, deberá aplanarse la estructura en filas y columnas simples. Seleccione un elemento del esquema y pulse el botón de mover para correlacionarlo con un campo de análisis. No todos los nodos de árbol pueden correlacionarse. Por ejemplo, un vector o una correlación de tipos complejos se considera "padre" y no puede correlacionarse. Tales nodos se identifican en el árbol mediante una etiqueta terminada con ...:array:struct o ...:map:struct.

Correlaciones de campos de HCatalog
Muestra la correlación de un elemento de HCatalog con un campo del origen de datos.

Pulse **Previsualización de los datos en bruto** para ver los registros tal y como están almacenados en HCatalog; esto podrá ser de ayuda a la hora de determinar cómo correlacionar el esquema de HCatalog con los campos.

Elemento de HCatalog
Efectúe una doble pulsación sobre una celda para editarla. La celda deberá editarse cuando el elemento de HCatalog sea un vector o una correlación. En el caso de los vectores, especifique el entero que corresponda con el miembro del vector que desee correlacionar con un campo. En el caso de las correlaciones, especifique una cadena entrecomillada que se corresponda con la clave que desee correlacionar con un campo. Consulte Figura 2 para obtener un ejemplo de cómo puede utilizarse la vista previa de datos en bruto para determinar la cadena que corresponde al índice de la correlación.

Campo de correlación
El campo tal y como aparece en el origen de datos de Analytic Server. Efectúe una doble pulsación sobre una celda para editarla. Los valores duplicados en la columna Campo de correlación no están permitidos y dan lugar a un error.

Almacenamiento
El almacenamiento del campo. El almacenamiento se deriva de HCatalog y no puede editarse.

Nota: Cuando se pulsa Vista previa y metadatos para terminar un origen de datos de HCatalog, no hay opciones de edición.

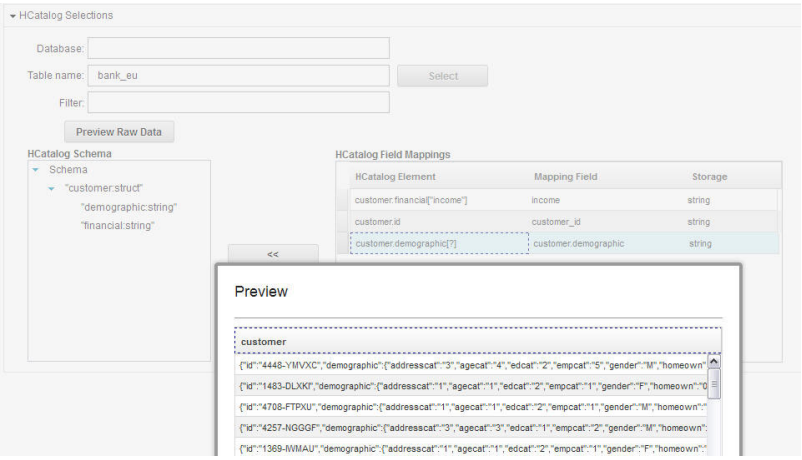

Figura 2. Acordeón de orígenes de datos, definición de un origen de datos de HCatalog

Valores (orígenes de datos de archivos)

Codificación de juego de caracteres
La codificación de caracteres del archivo. Seleccione o especifique un nombre de juego de caracteres Java como, por ejemplo, "UTF-8", "ISO-8859-2" o "GB18030". El valor predeterminado es "UTF-8".

Entorno local
Define un entorno local. Opcional. Toma como valor predeterminado el entorno local del servidor. La cadena del entorno local debe especificarse como: <idioma>[_país[_variante]], donde:

idioma
Un código válido de dos letras en minúscula tal y como se define en ISO-639. Es obligatorio.

país Un código válido de dos letras en mayúscula tal y como se define en ISO-3166. Es optativo.

variante
Código específico de navegador o proveedor. Es opcional.

Recortar espacios en blanco
Elimina los espacios en blanco del comienzo y/o final de los campos de cadena. El valor predeterminado es Ninguno. Los valores soportados son los siguientes:

Ninguno
No elimina los espacios en blanco.

Izquierda
Elimina los espacios en blanco que hay al comienzo de la cadena.

Derecha
Elimina los espacios en blanco que hay al final de la cadena.

Ambos
Elimina los espacios en blanco al comienzo y al final de la cadena.

Separadores de miles
Determina si se utiliza el carácter específico del entorno local que representa el separador de millares.

Delimitadores de campo
Uno o más caracteres que marcan los límites de un campo. Cada carácter se toma como un delimitador independiente. Por ejemplo, ",\t" significa que tanto una coma como un tabulador marcan los límites del campo. Si los campos están delimitados por caracteres de control, los caracteres especificados aquí se tratarán como delimitadores además de los caracteres de control. El valor predeterminado será "," cuando los campos no están delimitados por caracteres de control, y una cadena vacía en los demás casos.

Los caracteres de control delimitan campos
Determina si los caracteres de control ASCII, salvo LF y CR, se tratan como delimitadores de campo. El valor predeterminado es **No**.

La primera fila contiene los nombres de los campos
Determina si la primera fila se utiliza para especificar los nombres de campo. El valor predeterminado es **No**.

Número de caracteres iniciales omitidos
El número de caracteres que se omiten al comienzo del archivo. Es un entero no negativo. El valor predeterminado es cero.

Fusionar espacios en blanco
Determina si varias apariciones de espacios y/o tabuladores se tratan como un único delimitador de campo. No tiene efecto si ni el espacio en blanco ni el tabulador son delimitadores de campo. El valor predeterminado es 'true'.

Caracteres de comentarios de fin de línea
Son uno o más caracteres que marcan los comentarios de fin de línea. Se hará caso omiso del carácter y de todo lo que lo que vaya a continuación. Cada carácter se toma como un marcador independiente de comentario. Por ejemplo, "/*" significa que tanto una barra inclinada como un asterisco dan comienzo a un comentario. No se pueden definir marcadores de comentario de varios caracteres como "//". Una cadena vacía indica que no se ha definido ningún carácter de comentario. Cuando están definidos, los caracteres de comentario se comprueban antes de procesarse las comillas o de omitirse los caracteres que deban omitirse. El valor predeterminado es la cadena vacía.

Caracteres no válidos
Determina el modo en que se tratan los caracteres no válidos (secuencias de bytes que no se

corresponden con ningún carácter de la codificación). Una cadena vacía indica que deben descartarse. Una cadena no vacía (normalmente un único carácter) indica que deben sustituirse por el contenido de dicha cadena. El valor predeterminado es la cadena vacía.

Comillas simples
Especifica el tratamiento que reciben las comillas simples (apóstrofos). El valor predeterminado es **Mantener**.

Mantener
Las comillas simples carecen de significado especial y se tratan como cualquier otro carácter.

Descartar
Las comillas simples se suprimen a menos que vayan entrecomilladas.

Par Las comillas simples se tratan como caracteres de entrecomillado, de modo que los caracteres situados entre un par de comillas simples pierden cualquier significado que pudieran tener (se consideran entrecomillados). El valor **Las comillas pueden ir entrecomilladas mediante duplicación** determina si las propias comillas simples pueden aparecer en cadenas encerradas entre comillas simples.

Comillas dobles
Especifica el tratamiento que reciben las comillas dobles. El valor predeterminado es **Par**.

Mantener
Las comillas dobles carecen de significado especial y se tratan como cualquier otro carácter.

Descartar
Las comillas dobles se suprimen a menos que vayan entrecomilladas.

Par Las comillas dobles se tratan como caracteres de entrecomillado, de modo que los caracteres situados entre un par de comillas dobles pierden cualquier significado que pudieran tener (se consideran entrecomillados). El valor **Las comillas dobles pueden ir entrecomilladas mediante duplicación** determina si las propias comillas dobles pueden aparecer en cadenas encerradas entre comillas dobles.

Las comillas pueden ir entrecomilladas mediante duplicación
Indica si las comillas dobles pueden aparecer en cadenas encerradas entre comillas dobles, y si las comillas simples pueden aparecer en cadenas encerradas entre comillas simples cuando se han establecido a **Par**. Si tiene el valor **Sí**, las comillas dobles se escapan dentro de las cadenas encerradas entre comillas dobles duplicándolas, y las comillas simples se escapan en las cadenas encerradas entre comillas simples duplicándolas también. Si tiene el valor **No**, no será posible colocar una comilla doble dentro de una cadena encerrada entre comillas dobles, ni tampoco colocar una comilla simple dentro de una cadena encerrada entre comillas simples. El valor predeterminado es **Sí**.

Vista previa y metadatos (orígenes de datos)

Cuando se pulsa **Vista previa y metadatos**, se genera el modelo de datos del origen de datos a partir de una muestra de los registros. Esto da la posibilidad de editar la información básica de los metadatos.

Vista previa
La pestaña Vista previa ofrece una pequeña muestra de registros y sus valores de campo.

Edición

La pestaña Edición permite corregir los metadatos de campo básicos.

Campo
Efectúe una doble pulsación en el nombre del campo para editarlo.

Medida

Este es el nivel de medida que se utiliza para describir las características de los datos en un campo determinado.

Rol Se utiliza para indicar a los nodos de modelado si los campos serán de Entrada (campos predictores) o de Salida (campos predichos) para un proceso de aprendizaje automático. Ambos y Ninguno son asimismo roles, junto con Partición, que indica un campo que se utiliza para particionar registros en muestras independientes a efectos de formación, pruebas y validación. El valor División indica que se construirá un modelo aparte por cada posible valor del campo.

Almacenamiento

El almacenamiento describe la forma en que los datos se almacenan en un campo. Por ejemplo, un campo con valores 1 y 0 almacena datos enteros. Esto es distinto del nivel de medida, que describe el uso de los datos y no afecta al almacenamiento. Por ejemplo, puede que le interese establecer el nivel de medida de un campo entero con los valores de 1 y 0 a Indicador. Esto suele indicar que 1 = Verdadero y 0 = Falso.

Proyectos

Los proyectos son espacios de trabajo para almacenar las entradas de los trabajos y acceder a las salidas de los mismos. Proporcionan una estructura organizativa de nivel superior para contener archivos y carpetas. Los proyectos pueden compartirse con usuarios individuales y grupos.

Columna izquierda

La columna izquierda muestra los proyectos existentes bajo la cabecera del acordeón.

- Seleccione un proyecto para visualizar sus detalles en el área de contenidos y editar sus propiedades. Cuando se escribe en el área de búsqueda, se filtra el listado mostrándose solo aquellos proyectos que contengan la cadena de búsqueda en el nombre.
- Pulse **Nuevo proyecto** para crear un proyecto con el nombre especificado en el diálogo Añadir nuevo proyecto. En los nombres se distingue entre mayúsculas y minúsculas, se ignoran los espacios en blanco de principio y fin, y se ofrece protección frente a inyecciones SQL.
- Pulse **Suprimir proyecto** para eliminar el proyecto. Esta acción no afecta en modo alguno a los archivos asociados al origen de datos.

Área de contenidos

El área de contenidos se divide en las pestañas **Valores**, **Propietarios** y **Versión**.

Valores

Descripción del proyecto

Campo de texto editable que proporciona un texto descriptivo del proyecto.

Es público

Casilla de verificación que indica si cualquiera puede ver el proyecto (cuando está marcada) o si deben añadirse explícitamente usuarios y grupos como propietarios (cuando está sin marcar).

Al pulsar **Guardar** se guarda el estado de los valores en ese momento.

Orígenes de datos del proyecto

Es un área no editable en la que se listan todos los orígenes de datos asociados al proyecto.

Panel de estructura del proyecto

El panel derecho muestra la estructura del proyecto/carpeta del proyecto seleccionado en ese momento. Puede examinarse la estructura de carpetas, pero solo podrá editarse mediante los botones.

- Pulse **Descargar archivo a sistema de archivos local** para descargar un archivo seleccionado al sistema de archivos local.
- Pulse **Suprimir archivo(s) seleccionado(s)** para eliminar el archivo o la carpeta seleccionados.

Propietarios

Los propietarios son miembros de pleno derecho de un proyecto, y pueden modificar el proyecto, así como las carpetas y los archivos contenidos en ellas.

La lista Usuarios y grupos disponibles muestra los usuarios y grupos del inquilino activo que no están asociados en ese momento al proyecto.

- Cuando se escribe en el panel de búsqueda, se filtran usuarios y grupos que tengan la cadena de búsqueda en el nombre.
- Cuando se selecciona el icono de usuarios que aparece encima de la lista, se muestran los usuarios disponibles. Si se deselecciona el icono, se ocultan los usuarios. Este icono está seleccionado de forma predeterminada.
- Cuando se selecciona el icono de grupos que aparece encima de la lista, se muestran los grupos disponibles. Si se deselecciona el icono, se ocultan los grupos. Este icono está seleccionado de forma predeterminada.

Usuarios y grupos pueden moverse a los usuarios y grupos del proyecto mediante el botón de mover. Estos usuarios, y los miembros de estos grupos, tienen acceso de lectura (nodo Origen de Analytic Server) y escritura (nodo Exportación de Analytic Server) a este proyecto cuando se conecta a Analytic Server a través de IBM SPSS Modeler.

Nota: Los cambios efectuados en la pestaña Propietarios se aplican de forma inmediata y automática.

Nota: Los administradores siempre tienen acceso de lectura y escritura a todos los proyectos, independientemente de que aparezcan listados como propietarios.

Versiones

Los proyectos se versionan en función de los cambios efectuados al contenido de archivos y carpetas. Los cambios efectuados en los atributos de un proyecto como, por ejemplo, la descripción, si es público, y con quién se comparte, no requieren una nueva versión. La adición, modificación o supresión de archivos o carpetas sí requieren una nueva versión.

Tabla de control de versiones de proyectos

La tabla muestra las versiones de proyecto existentes, sus fechas de creación y confirmación, los usuarios responsables de cada versión y la versión padre. La versión padre es la versión en la que se basa la versión seleccionada.

- Pulse **Bloquear** para efectuar cambios en el contenido de la versión seleccionada.
- Pulse **Confirmar** para guardar todos los cambios efectuados a un proyecto y hacer que esta versión sea el estado visible actual del proyecto.
- Pulse **Descartar** para descartar todos los cambios efectuados a un proyecto bloqueado y que el estado visible del proyecto vuelva a la versión confirmada más reciente.
- Pulse **Suprimir** para eliminar la versión seleccionada.

Limpiar automáticamente cuando se supera el número de versiones

Suprime de forma automática la versión de proyecto más antigua cuando el número de versiones sobrepasa el número especificado.

Avisos

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

For license inquiries regarding double-byte (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502 Japan

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
U.S.A.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

All statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

All IBM prices shown are IBM's suggested retail prices, are current and are subject to change without notice. Dealer prices may vary.

This information is for planning purposes only. The information herein is subject to change before the products described become available.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

Each copy or any portion of these sample programs or any derivative work, must include a copyright notice as follows:

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

Each copy or any portion of these sample programs or any derivative work, must include a copyright notice as follows:

© your company name) (year). Portions of this code are derived from IBM Corp. Sample Programs.

If you are viewing this information softcopy, the photographs and color illustrations may not appear.

Marcas registradas

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at www.ibm.com/legal/copytrade.shtml.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency which is now part of the Office of Government Commerce.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

ITIL is a registered trademark, and a registered community trademark of The Minister for the Cabinet Office, and is registered in the U.S. Patent and Trademark Office.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.

Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.

Índice

Impreso en España