IBM SPSS Analytic Server
Versión 3.1.0

Guía del usuario

Nota
Antes de utilizar esta información y el producto al que da soporte, lea la información del apartado "Avisos" en la página 33.

Información sobre el producto

Esta edición se aplica a la versión 3, release 1, modificación 0 de IBM SPSS Analytic Server y a todos los releases y modificaciones posteriores hasta que se indique lo contrario en nuevas ediciones.

Contenido

Capítulo 1. Consola de Analytic Server

Analytic Server proporciona una interfaz de cliente ligero para gestionar orígenes de datos y proyectos.

Iniciar la sesión

1. Escriba el URL de Analytic Server en la barra de direcciones del navegador. El URL puede obtenerse 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.

Después del inicio de sesión, se visualiza la pantalla de inicio de la consola.

Navegación en la consola

- La cabecera muestra el nombre de producto, el nombre del usuario conectado actualmente y el enlace al sistema de ayuda. El nombre del usuario que actualmente ha iniciado la sesión es el primero de una lista desplegable que incluye el enlace de cierre de sesión.
- El área de contenidos visualiza las acciones que puede realizar desde la pantalla inicial de la consola.

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 relacional, HCatalog o geoespacial. 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.

Listado de orígenes de datos

La página principal de Orígenes de datos proporciona una lista de los orígenes de datos de los que el usuario actual es miembro.

- Pulse el nombre de un origen de datos para visualizar sus detalles y editar sus propiedades.
- Escriba en el área de búsqueda para filtrar el listado a fin de visualizar solo orígenes de datos con la serie de búsqueda en su nombre.
- Pulse **Nuevo** para crear un nuevo origen de datos con el nombre y el tipo de contenido especificado en el diálogo **Añadir nuevo origen de datos**.
  - Consulte “Reglas de denominación” en la página 23 para conocer las restricciones sobre los nombres que puede dar a los orígenes de datos.
  - Los tipos de contenido disponibles son Archivo, Base de datos, HCatalog y Geospatial.

    Notas:
    - La opción HCatalog solo está disponible si se ha configurado Analytic Server para trabajar con esos orígenes de datos.
    - Una vez seleccionado, el tipo de contenido no podrá editarse.
    - Puede importar/exportar varios orígenes de datos en una sola acción.
- Pulse **Suprimir** para eliminar el origen de datos. Esta acción no afecta en modo alguno a los archivos asociados al origen de datos.
- Pulse **Renovar** para actualizar la lista.
- La lista desplegable **Acciones** realiza la acción seleccionada.

1. Seleccione **Exportar** para crear un archivado de los orígenes de datos seleccionados y guarde el archivado en el sistema de archivos local. Este archivo incluye los archivos que se añadieron a los orígenes de datos seleccionados en modalidad **Proyectos** o en modalidad **Origen de datos**.

   Nota: Cuando solamente se selecciona un origen de datos, el nombre de archivo de archivado comparte el nombre de origen de datos seleccionado. Si se selecciona más de un origen de datos, el nombre de archivo de archivado adopta de forma predeterminada el nombre `datasources.zip`.
2. Seleccione **Importar** para importar los archivos creados por la acción Exportar.

   Nota: Los archivos de archivado que contienen información de varios orígenes de datos no se pueden importar. En estos casos, los archivos de orígenes de datos individuales deben extraerse en primer lugar del archivado `datasources.zip`.
3. Seleccione **Duplicar** para crear una copia del origen de datos.

Detalles individuales del origen de datos

El área de contenidos se divide en varias secciones, dependiendo del tipo de contenido del origen de datos.

Detalles

Estos valores son comunes a todos los tipos de contenido.

Nombre
: Campo de texto editable que muestra el nombre del origen de datos.

Nombre de visualización
: Campo de texto editable que muestra el nombre del origen de datos tal como se visualiza en otras aplicaciones. Si está en blanco, se utiliza el Nombre como nombre de visualización.

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 origen de datos (cuando está marcada) o si deben añadirse explícitamente usuarios y grupos como miembros (cuando está sin marcar).

Atributos personalizados
: Las aplicaciones pueden asociar propiedades a los orígenes de datos como, por ejemplo, si el origen de datos es temporal, mediante el uso de atributos personalizados. Estos atributos se exponen en la consola de Analytic Server para proporcionar una visión más detallada del modo en que las aplicaciones utilizan el origen de datos.

Pulse **Guardar** para conservar el estado actual de los valores.

Compartición

Estos valores son comunes a todos los tipos de contenido.

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

- Cuando se escribe en el cuadro de texto, se filtran usuarios y grupos que tengan la serie de búsqueda en el nombre. Seleccione **Autor** o **Lector** de la lista desplegable para asignar su rol en el origen de datos. Pulse **Añadir miembro** para añadirlos a la lista de miembros.
- Para eliminar un participante, seleccione un usuario o grupo en la lista de miembros y pulse **Eliminar miembro**.

Nota: Los usuarios con el rol de **Administrador** tendrán acceso de lectura y escritura a todos los orígenes de datos, independientemente de que aparezcan o no específicamente en la lista de miembros.

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 de proyectos de Analytic Server, **Origen de datos** para ver los archivos almacenados en un origen de datos o **Sistema de archivos** para visualizar el sistema de archivos (normalmente HDFS). Puede examinar la estructura de carpetas, pero HDFS no se puede editar de ningún modo, y en la modalidad **Proyectos**, no puede añadir archivos, crear carpetas ni eliminar elementos en el nivel raíz, sino solo dentro de los proyectos definidos. Para crear, editar o suprimir un proyecto, utilice Proyectos.

- Pulse **Cargar** para cargar un archivo en el origen de datos o en el proyecto/subcarpeta actuales. Puede buscar y seleccionar varios archivos en un único directorio.

  Nota: Los archivos se cargan en el sistema de archivos distribuidos. Puede encontrar los archivos cargados en la estructura de directorios `/analytic-root`, bajo el inquilino, el origen de datos o proyecto (en función de la modalidad elegida) y la subcarpeta correspondientes. Por ejemplo, si:

  1. Inicia la sesión en el inquilino `ibm`
  2. Crea un origen de datos denominado `fraudDetection`
  3. Selecciona una modalidad de **Origen de datos**
  4. Crea una subcarpeta denominada `historicalData`
  5. Carga un archivo `charges2015.csv`

  En este caso, el archivo se encontraría en el sistema de archivos distribuidos en `/analytic-root/ibm/.datasource/fraudDetection/historicalData/charges2015.csv`. Si, en cambio:

  1. Inicia la sesión en el inquilino `ibm`
  2. Crea un origen de datos denominado `fraudDetection`
  3. Selecciona la modalidad **Proyecto**
  4. Selecciona un proyecto existente denominado `creditProcessing`
  5. Crea una subcarpeta denominada `historicalData`
  6. Carga un archivo `charges2015.csv`

  En este caso, el archivo se encontraría en el sistema de archivos distribuidos en `/analytic-root/ibm/creditProcessing/historicalData/charges2015.csv`.
- Pulse **Nueva carpeta** para crear una carpeta bajo la carpeta actual, con el nombre especificado en el diálogo Nombre de la nueva carpeta.
- Pulse **Descargar** para descargar los archivos seleccionados en el sistema de archivos local.
- Pulse **Suprimir** para eliminar los archivos y carpetas seleccionados.

Archivos incluidos en la definición del origen de datos

Utilice el botón de mover para añadir los archivos y carpetas seleccionados al origen de datos, o para eliminarlos de él. Por cada archivo o carpeta seleccionado en el origen de datos, pulse Valores para definir las especificaciones de lectura del archivo.

Cuando se incluyen varios archivos en un origen de datos, deben compartir metadatos comunes; es decir, cada archivo debe tener el mismo número de campos, los campos deben analizarse en el mismo orden en cada uno de los archivos, y cada campo debe

tener el mismo almacenamiento en todos los archivos. Las discrepancias entre los archivos pueden hacer que la consola no pueda crear la Vista previa y metadatos, o que valores válidos se analicen como no válidos (nulos) cuando Analytic Server lee el archivo.

Selecciones de base de datos

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

Base de datos
Seleccione el tipo de base de datos a la que desea conectarse. Elija entre: DB2, Greenplum, Amazon Redshift, MySQL, Netezza, Oracle, SQL Server, Sybase IQ, TeraData, Hive, DashDB o BigSQL. Si el tipo que está buscando no aparece en la lista, solicite al administrador del servidor que configure Analytic Server con el controlador JDBC adecuado.

Nota: Analytic Server permite utilizar bases de datos MySQL situadas en sistemas remotos.

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.

Número máximo de lecturas simultáneas
Especifique el límite en el número de consultas paralelas que se pueden enviar desde Analytic Server a la base de datos para leer desde la tabla especificada en el origen de datos.

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. El filtrado de HCatalog sólo está soportado en claves de partición de Hive de tipo serie.

Nota: los operadores !=, <> y LIKE parecen no funcionar en ciertas distribuciones Hadoop. Se trata de un problema de compatibilidad entre HCatalog y dichas distribuciones.

Correlaciones de campos de HCatalog
Muestra la correlación de un elemento de HCatalog con un campo del origen de datos. Pulse Editar para modificar las correlaciones de campos.

Nota: Después de crear un origen de datos basado en HCatalog que expone los datos de una tabla de Hive, puede que observe que, cuando la tabla de Hive se ha formado a partir de un gran número de archivos de datos, se produce un retardo significativo cada vez que Analytic Server empieza a leer datos del origen de datos. Si observa tales retrasos, vuelva a crear la tabla de Hive utilizando un número más pequeño de archivos de datos más grandes y reduzca el número de archivos a 400 o menos.

Selecciones geoespaciales
Especifique los parámetros para acceder a los datos geográficos.

Tipo geospacial
Los datos geográficos pueden venir de un servicio de correlación en línea o de un archivo de forma.

Si está utilizando un servicio de correlación, especifique el URL del servicio y seleccione la capa de correlación que dese usar.

SI está usando un archivo de forma, seleccione o cargue el archivo de forma. Tenga en cuenta que un archivo de forma realmente es un conjunto de archivos con un nombre de archivo común, almacenado en el mismo directorio. Seleccione el archivo con el sufijo SHP. Analytic Server buscará y utilizará los demás archivos. Debe haber siempre presentes dos archivos adicionales con los sufijos SHX y DBF; en función del archivo de forma, puede haber presentes también un número de archivos adicionales.

Vista previa y metadatos
Una vez especificados los valores de configuración del origen de datos, pulse Vista previa y metadatos para comprobar y confirmar las especificaciones del origen de datos.

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

- Para los orígenes de datos con tipo de contenido de base de datos, elija una tabla de base de datos de salida donde se escriben los datos de salida.
- Para los orígenes de datos con tipo de contenido de archivos:
  1. Elija una carpeta de salida donde se escriben los nuevos archivos.

     Consejo: Utilice una carpeta separada para cada origen de datos de modo que sea más fácil hacer un seguimiento de las asociaciones entre los archivos y los orígenes de datos.
  2. Seleccione un formato de archivo; **CSV** (valor separado por comas) o **Formato binario divisible**.
  3. De forma opcional, seleccione **Crear archivo de secuencia**. Esto es útil si desea crear archivos comprimidos divisibles que se puedan utilizar en trabajos MapReduce en sentido descendente.
  4. Seleccione **Puede especificarse un escape para las nuevas líneas** si la salida es CSV y tiene campos de serie que contienen caracteres de nueva línea o de retorno de carro incluidos. Esto hará que cada nueva línea se escriba como una barra inclinada invertida seguida de la letra “n”, el retorno de carro como una barra inclinada invertida seguida de la letra “r” y la barra inclinada invertida como dos barras inclinadas invertidas consecutivas. Este tipo de datos se debe leer con el mismo valor. Sugerimos que utilice el formato binario divisible al gestionar los datos de serie que contienen caracteres de retorno de carro o de nueva línea.
  5. Seleccione un formato de compresión. La lista incluye todos los formatos que se han configurado para su uso en la instalación de Analytic Server.

     Nota: algunas combinaciones de formato de compresión y formato de archivo hacen que la salida no pueda dividirse y, por lo tanto, no son apropiadas para proceso MapReduce adicional. Analytic Server genera un aviso en la sección de salida cuando se realiza una selección de este tipo.

Valores (orígenes de datos de archivos)

El diálogo Valores permite definir las especificaciones de lectura de datos basados en archivos. Los valores se aplican a todos los archivos seleccionados y todos los archivos de las carpetas seleccionadas que coinciden con los criterios de la pestaña **Carpeta**.

La especificación de valores de analizador incorrectos para un archivo pueden hacer que la consola no pueda crear la Vista previa y metadatos, o que valores válidos se analicen como no válidos (nulos) cuando Analytic Server lee el archivo.

Pestaña Valores

La pestaña Valores permite especificar el tipo de archivo y los valores de analizador específicos del tipo de archivo.

Puede definir orígenes de datos utilizando archivos comprimidos para cualquier formato de archivo soportado. Los formatos de compresión soportados incluyen Gzip, Deflate, Bz2, Snappy e IBM CMX.

Tipo de archivo delimitado

Los archivos delimitados son archivos de texto de campo libre, cuyos registros contienen un número constante de campos pero un número variable de caracteres por campo. Los archivos delimitados tienen generalmente las extensiones de archivo *.csv o *.tab. Consulte “Valores de tipo de archivo delimitado” en la página 7 para obtener información adicional.

Tipo de archivo fijo

Los archivos de texto de campo fijo son archivos cuyos campos no están delimitados, sino que se inician en la misma posición y son de longitud fija. Los archivos de texto de campo fijo suelen tener la extensión de archivo *.dat. Consulte “Valores de tipo de archivo fijo” en la página 9 para obtener información adicional.

Tipo de archivo semiestructurado

Los archivos semiestructurados (como por ejemplo *.log) son archivos de texto que tienen una estructura previsible que puede correlacionarse con los campos por medio de expresiones regulares, pero que no están tan estructurados como los archivos delimitados. Consulte “Valores de tipo de archivo semiestructurado” en la página 9 para obtener información adicional.

Tipo de archivo de análisis de texto

Los archivos de análisis de texto son documentos (como *.doc, *.pdf o *.txt) que se pueden analizar utilizando SPSS Text Analytics.

Omitir líneas vacías
: Especifica si se deben ignorar las líneas vacías en el contenido del texto extraído. El valor predeterminado es **No**.

Separador de líneas
: Especifica la serie que define una línea nueva. El valor predeterminado es el carácter de nueva línea "\n".

Tipo de archivo de SPSS Statistics

Los archivos de SPSS Statistics (*.sav, *.zsav) son archivos binarios que contienen un modelo de datos. No son necesarios más valores de la pestaña Valores para este tipo de archivo.

Tipo de archivo de formato binario divisible

Especifica que el tipo de archivo es un archivo de formato binario divisible (*.asbf). Este tipo de archivo puede representar todos los tipos de campo de Analytic Server (a diferencia de CSV, que no puede representar los campos de lista en absoluto y requiere valores especiales para gestionar las nuevas líneas y retornos de carro incluidos. No son necesarios más valores de la pestaña Valores para este tipo de archivo.

Tipo de archivo de secuencia

Los archivos de secuencia (*.seq) son archivos de texto estructurado como pares de clave/valor. Se utilizan habitualmente como formato de intermediario en los trabajos de MapReduce.

Tipo de archivo Excel

Especifica que el tipo de archivo es un archivo Microsoft Excel (*.xls, *.xlsx). Consulte “Valores de tipo de archivo Excel” en la página 10 para obtener información adicional.

Valores de tipo de archivo delimitado:

Puede especificar los valores siguientes para los tipos de archivo delimitado.

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**.

Delimitadores de campo
Uno o más caracteres que marcan los límites de un campo. Cada carácter se toma como un delimitador independiente. Por ejemplo, si selecciona **Coma** y **Tabulador** (o selecciona **Otro** y especifica `,\t`), significa que una coma o un tabulador marca los límites de 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 "," si los campos no están delimitados por caracteres de control; de lo contrario, el valor predeterminado será la serie vacía.

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 0.

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 **Sí**.

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 serie 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 serie 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).

Descartar
Descarta secuencias de bytes no válidas.

Sustituir con
Sustituye cada secuencia de bytes no válida por el carácter único proporcionado.

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 series 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 series encerradas entre comillas dobles.

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

Puede especificarse un escape para las nuevas líneas

Indica si el analizador interpreta una barra inclinada invertida seguida de la letra "n", la letra "r" u otra barra inclinada invertida como un carácter de nueva línea, retorno de carro o barra inclinada invertida, respectivamente. Si no se procesan con escape los caracteres de nueva línea,

esas secuencias de caracteres se leen literalmente como una barra inclinada invertida seguida de la letra "n" y así sucesivamente. El valor predeterminado es **No**.

Valores de tipo de archivo fijo:

Puede especificar los valores siguientes para los tipos de archivo fijo.

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**.

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).

Descartar
Descarta secuencias de bytes no válidas.

Sustituir con
Sustituye cada secuencia de bytes no válida por el carácter único proporcionado.

Longitud de registro
Indica cómo se definen los registros. Si es **Delimitado por nueva línea**, los registros se definen (delimitan) mediante nuevas líneas, comienzo de archivo o fin de archivo. Si es **Longitud específica**, los registros se definen mediante una longitud de registro en bytes. Especifique un valor positivo.

Registros iniciales omitidos
El número de registros que se omiten al comienzo del archivo. Especifique un entero no negativo. El valor predeterminado es 0.

Campos
Esta sección define los campos del archivo. Pulse **Añadir campo** y especifique el nombre del campo, la columna en la que se inician los valores de campo y la longitud de los valores de campo. Las columnas de un archivo se numeran a partir de 0.

Valores de tipo de archivo semiestructurado:

Los valores de los archivos semiestructurados están formados por reglas de correlación del contenido del archivo con los campos.

Tabla de reglas
Las reglas individuales extraen información de un registro para crear un campo;, todas ellas juntas en la tabla de reglas, definen todos los campos que se pueden extraer de cada registro de un origen de datos.

Las reglas de la tabla se aplican por orden a cada registro; si todas las reglas de la tabla coinciden con el registro, no son necesarias otras tablas de reglas para procesar el registro, y se procesa el registro siguiente. Si ninguna regla de la tabla no coincide, todos los valores de campo extraídos por las reglas anteriores de la tabla se descartan; si hay otra tabla de reglas, las reglas de esa tabla se aplican al registro. Si ninguna tabla coincide con el registro, se aplica la regla de no coincidencia.

No coincidencia
Puede elegir **Omitir** los registros que no coincidan con ninguna de las tablas de reglas, o establecer el valor de todos los campos del registro en **Faltante** (nulo).

Exportar reglas
Puede guardar la tabla de reglas actualmente visible para su reutilización. La tabla exportada se guarda en el servidor.

Importar reglas
Puede importar una tabla de reglas guardada en la tabla de reglas visible actualmente. Esto sobrescribe las reglas que haya definido para dicha tabla, por lo que es mejor crear una tabla nueva y, a continuación, importar una tabla de reglas.

Editor de reglas

El editor de reglas permite crear una regla de extracción para un campo único.

Grupo de captura anónimo
Una regla de captura de campo normalmente empieza a extraer datos de un registro en la posición donde se ha detenido la regla anterior. Cuando existe información superflua entre dos campos de un origen de datos semiestructurado, puede ser útil definir un grupo de captura anónimo que sitúe el analizador en el punto donde empieza el campo siguiente. Al seleccionar **Grupo de captura anónimo**, los controles para denominar y etiquetar el grupo de captura quedan inhabilitados, pero el resto del diálogo funciona con normalidad.

Nombre de campo
Especifique un nombre para el campo. Se utiliza para definir los metadatos del origen de datos. Los nombres de campo deben ser exclusivos dentro de una tabla de reglas.

Nombre de regla
Opcionalmente, especifique una etiqueta descriptiva para la regla.

Descripción
Opcionalmente, especifique una descripción más larga para la regla.

Definición de una regla
Existen dos métodos para definir reglas.

Utilizar controles para reglas de extracción
Esto simplifica la creación de reglas de extracción.

1. Especifique el punto donde debe iniciarse la extracción de datos de campo; **Posición actual** empezará donde se ha detenido la regla anterior, y **Omitir hasta** empezará al principio del registro e ignorará todos los caracteres hasta llegar al especificado en el recuadro de texto. Seleccione **Incluir** si desea que los datos del campo incluyan el carácter situado en la posición inicial.
2. Seleccione un grupo de captura de campos en la lista desplegable **Captura**.
3. Opcionalmente, seleccione el punto donde debe detenerse la extracción de datos del campo; **Espacio en blanco** se detendrá cuando se encuentre algún carácter de espacio en blanco (tales como espacios o tabuladores), y **En carácter(s)** se detendrá en la serie especificada. Seleccione **Incluir** si desea que los datos del campo incluyan el carácter situado en la posición de detención.

Definir manualmente reglas regexp
Seleccione esta opción si prefiere escribir sintaxis de expresión regular. Especifique una expresión regular en el recuadro de texto **Regexp**.

Añadir grupo de captura de campos
Permite guardar la expresión regular para su uso posterior. El grupo de captura guardado aparece en la lista desplegable **Captura**.

El editor de reglas muestra una vista previa de los datos extraídos del primer registro por esta regla, una vez aplicadas todas las reglas anteriores de la tabla de reglas.

Valores de tipo de archivo Excel:

Puede especificar los valores siguientes para los archivos Excel.

Selección de hoja de cálculo
Selecciona la hoja de cálculo Excel que se va a utilizar como origen de datos. Especifique un índice numérico (el índice de la primera hoja de cálculo es 0) o el nombre de la hoja de cálculo. El valor predeterminado es utilizar la primera hoja de cálculo.

Selección de rango de datos para la importación.
Puede importar datos que comiencen por la primera fila que no esté en blanco o con un rango de casillas explícito.

- **El rango comienza en la primera fila no en blanco**. Localiza la primera casilla que no está en blanco y la utiliza como el ángulo superior izquierdo del rango de datos.
- También puede especificar un rango de casillas explícito por fila y columna. Por ejemplo, para especificar el rango de Excel A1:D5, puede especificar A1 en el primer campo y D5 en el segundo (o, como alternativa, R1C1 y R5C4). Se devolverán todas las filas del rango especificado, incluidas las filas en blanco.

La primera fila contiene los nombres de los campos
Especifica si la primera fila del rango de celdas seleccionado contiene los nombres de campo. El valor predeterminado es **No**.

Detener lectura después de encontrar filas en blanco
Especifica si se debe detener la lectura de registros después de encontrar más de una fila vacía, o continuar leyendo todos los datos hasta el final de la hoja de cálculo, incluidas las filas en blanco. El valor predeterminado es **No**.

Formatos

La pestaña Formatos permite definir información de formato para los campos analizados.

Valores de conversión de campo

Recortar espacios en blanco
Elimina los espacios en blanco del comienzo y/o final de los campos de serie. 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 serie.

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

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

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

idioma
Un código válido de dos letras en minúscula definido por ISO-639.

país Un código válido de dos letras en mayúscula definido por ISO-3166.

variante
Código específico de navegador o proveedor.

Separador de decimales
Establece el carácter utilizado como signo decimal. Toma como valor predeterminado el valor específico del entorno local.

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

Formato de fecha predeterminado
Define un formato de fecha predeterminado. Todos los patrones de formato definidos por la especificación de lenguaje de códigos de datos de entorno local (LDML) unicode están soportados.

Formato de hora predeterminado
Define un formato de hora predeterminado.

Indicación de fecha y hora predeterminada
Define un formato de indicación de fecha y hora predeterminado.

Huso horario predeterminado
Establece el huso horario. El valor predeterminado es UTC. El valor se aplica a los campos de hora y de indicación de fecha y hora de que no tienen un huso horario especificado explícitamente.

Alteraciones temporales de campo

Esta sección permite asignar instrucciones de formato a campos individuales. Seleccione un campo del modelo de datos o especifique un nombre de campo y haga clic en **Añadir** para añadirlo a la lista de campos con instrucciones individuales. Pulse **Eliminar** para eliminarlo de la lista. Para un campo seleccionado en la lista, puede establecer las siguientes propiedades del campo.

Almacenamiento
Establezca el almacenamiento del campo.

Separador de decimales
Para campos con almacenamiento Real, establece el carácter utilizado como signo decimal. Toma como valor predeterminado el valor específico del entorno local.

Separadores de miles
Para los campos con almacenamiento Entero o real, establece si se utiliza el carácter específico del entorno local que representa el separador de millares.

Formatos
Para los campos con almacenamiento de Fecha, Hora o Indicación de fecha y hora, establece el formato. Elija un formato en la lista desplegable.

Pestaña Orden de campos

Para los tipos de archivo delimitado y Excel, la pestaña Orden de campos permite definir el orden analizado de los campos del archivo. Esto es importante cuando hay varios archivos en un origen de datos, porque el orden real de los campos puede ser diferente en los archivos, pero el orden analizado de los campos debe ser el mismo para crear un modelo de datos coherente.

Para los tipos de archivo fijo y semiestructurado, el orden se define en la pestaña Valores.

Cuando hay un solo archivo en el origen de datos o todos los archivos tienen el mismo orden de campos, puede utilizar el valor predeterminado de **El orden de los campos coincide con el modelo de datos**. Si hay varios archivos en el origen de datos y el orden de los campos del archivo no coincide, defina un **Orden de campos específico** para analizar el archivo.

1. Para añadir un campo a la lista ordenada, especifique el nombre del campo o selecciónelo en la lista proporcionada por el modelo de datos. Puede añadir todos los campos del modelo de datos a la vez pulsando **Añadir todo**. Los nombres de campo sólo se añadirán una vez a la lista ordenada.
2. Use los botones de flecha para ordenar los campos como desee.

Cuando se utiliza **Orden de campos específico**, los campos que no se han añadido a la lista no forman parte del conjunto de resultados de este archivo. Si hay campos del modelo de datos que no aparecen en este diálogo, los valores son nulos en el conjunto de resultados.

Pestaña Carpeta

Al especificar valores de analizador para una carpeta, la pestaña Carpeta le permite seleccionar los archivos de la carpeta que se incluyen en el origen de datos.

Comparar todos los archivos de la carpeta seleccionada
El origen de datos incluye todos los archivos del nivel superior de la carpeta; los archivos de subcarpetas no se incluyen.

Comparar archivos utilizando una expresión regular
El origen de datos incluye todos los archivos del nivel superior de la carpeta que coinciden con la expresión regular especificada; los archivos de subcarpetas no se incluyen.

Comparar archivos utilizando una expresión de globalización Unix (potencialmente recursiva)
El origen de datos incluye todos los archivos que coinciden con la expresión de globalización Unix especificada; la expresión puede incluir archivos que están en subcarpetas de la carpeta seleccionada.

Correlaciones de campos de HCatalog

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, una matriz o una correlación de tipos complejos se considera "padre" y no puede correlacionarse directamente; cada elemento simple de una matriz o correlación de HCatalog debe añadirse por separado. Tales nodos se identifican en el árbol mediante una etiqueta terminada con ...:array:struct o ...:map:struct.

Por ejemplo:

- Para una matriz de enteros, puede asignar un campo a un valor de la matriz: bigintarray[45], pero no a la propia matriz: bigintarray
- Para una correlación, puede asignar un campo a un valor de la matriz: datamap["key"], pero no a la propia matriz: datamap
- Para una matriz de enteros, puede asignar un campo a un valor bigintarrayarray[45][2], pero no a la propia matriz, bigintarrayarray[45]

  Por lo tanto, cuando se asigna un campo a un elemento de matriz o correlación, la definición del elemento debe incluir el índice o clave: bigintarray[index] o bigintmap["key"].

El usuario actual solo puede ver las tablas a las que tiene acceso. Puesto que el directorio HDFS es el único directorio con permiso de lectura y ejecución (el archivo interior tiene permiso de lectura, que el usuario puede ver), los usuarios no pueden ver las tablas a las que no tienen acceso. La limitación está en vigor para proteger las tablas Hive gestionadas, tablas Hive externas y directorios particionados.

Correlaciones de 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 serie entrecomillada que se corresponda con la clave que desee correlacionar con un campo.

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.

Datos en bruto
Visualiza 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.

Nota: Cualquier filtrado especificado en las Selecciones de HCatalog se aplica a la vista de los datos en bruto.

Utilización de orígenes de datos de HCatalog

Analytic Server proporciona soporte para orígenes de datos de HCatalog. Esta sección describe cómo habilitar varias bases de datos NoSQL subyacentes.

En la mayoría de los casos, debe consultar la documentación del proveedor para la integración de Hive.

Apache Accumulo
https://cwiki.apache.org/confluence/display/Hive/AccumuloIntegration

Apache Cassandra
"Apache Cassandra"

Apache HBase
https://cwiki.apache.org/confluence/display/Hive/HBaseIntegration

MongoDB
https://github.com/mongodb/mongo-hadoop/wiki/Hive-Usage

Oracle NoSQL
https://docs.oracle.com/cd/E57371_01/doc.41/e57351/bigsql.htm#BIGUG21115

Orígenes de datos XML
"Orígenes de datos XML" en la página 16

Apache Cassandra

Analytic Server proporciona soporte para orígenes de datos HCatalog que tienen contenido subyacente en Apache Cassandra.

Cassandra proporciona un almacén estructurado de claves-valores. Las claves se correlacionan con varios valores, que se agrupan en familias de columnas. Las familias de columnas son fijas cuando se crea una base de datos, pero pueden añadirse columnas a una familia en cualquier momento. Además, sólo se añaden columnas a las claves especificadas, de modo que claves diferentes pueden tener números de columnas diferentes de cualquier familia determinada. Los valores de una familia columnas para cada clave se almacenan juntos.

Hay dos formas de definir tablas Cassandra: utilizando la interfaz de línea de mandatos de Cassandra de herencia (cassandra-cli) y la nueva shell CQL (csqlsh).

Utilice la sintaxis siguiente para crear una tabla Apache Cassandra externa en Hive si la tabla se ha creado utilizando la CLI de herencia.

```
CREATE EXTERNAL TABLE <nombre_tabla_hive>  (<especificaciones de columna>)
STORED BY 'org.apache.hadoop.hive.cassandra.CassandraStorageHandler'
WITH SERDEPROPERTIES("cassandra.cf.name" = "<familia_columnas_cassandra>",
"cassandra.host"="<host_cassandra>","cassandra.port" = "<puerto_cassandra>")
TBLPROPERTIES ("cassandra.ks.name" = "<espacio_claves_cassandra>");
```

Por ejemplo, para la definición de tabla de CLI siguiente:

```
create keyspace test
with placement_strategy = 'org.apache.cassandra.locator.SimpleStrategy'
and strategy_options = [{replication_factor:1}];

create column family users with comparator = UTF8Type;

update column family users with
        column_metadata =
        [
        {column_name: first, validation_class: UTF8Type},
        {column_name: last, validation_class: UTF8Type},
        {column_name: age, validation_class: UTF8Type, index_type: KEYS}
        ];

assume users keys as utf8;

set users['jsmith']['first'] = 'John';
set users['jsmith']['last'] = 'Smith';
set users['jsmith']['age'] = '38';
set users['jdoe']['first'] = 'John';
set users['jdoe']['last'] = 'Dow';
set users['jdoe']['age'] = '42';

get users['jdoe'];
```

... el DDL de tabla Hive sería el siguiente:

```
CREATE EXTERNAL TABLE cassandra_users  (key string, first string, last string, age string)
STORED BY 'org.apache.hadoop.hive.cassandra.CassandraStorageHandler'
WITH SERDEPROPERTIES("cassandra.cf.name" = "users",
"cassandra.host"="<host_cassandra>","cassandra.port" = "9160")
TBLPROPERTIES ("cassandra.ks.name" = "test");
```

Utilice la sintaxis siguiente para crear una tabla Apache Cassandra externa en Hive si la tabla se ha creado utilizando CQL.

```
CREATE EXTERNAL TABLE <nombre_tabla_hive>  (<especificaciones de columna>)
STORED BY 'org.apache.hadoop.hive.cassandra.cql.CassandraCqlStorageHandler'
WITH SERDEPROPERTIES("cassandra.cf.name" = "<familia_columnas_cassandra>",
"cassandra.host"="<host_cassandra>","cassandra.port" = "<puerto_cassandra>")
TBLPROPERTIES ("cassandra.ks.name" = "<espacio_claves_cassandra>");
```

Por ejemplo, para la definición de tabla de CQL3 siguiente:

```
CREATE KEYSPACE TEST WITH REPLICATION = { 'class' : 'SimpleStrategy', 'replication_factor' : 2 };
USE TEST;

CREATE TABLE bankloan_10(
  row int,
  age int,
  ed int,
  employ int,
  address int,
  income int,
  debtinc double,
  creddebt double,
  othdebt double,
  default int,
  PRIMARY KEY(row)
);
```

```
INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (1,41,3,17,12,176,9.3,11.359392,5.008608,1);
INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (2,27,1,10,6,31,17.3,1.362202,4.000798,0);
INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (3,40,1,15,14,55,5.5,0.856075,2.168925,0);
INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (4,41,1,15,14,120,2.9,2.65872,0.82128,0);
INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (5,24,2,2,0,28,17.3,1.787436,3.056564,1);
INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (6,41,2,5,5,25,10.2,0.3927,2.1573,0);
INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (7,39,1,20,9,67,30.6,3.833874,16.668126,0);
INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (8,43,1,12,11,38,3.6,0.128592,1.239408,0);
INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (9,24,1,3,4,19,24.4,1.358348,3.277652,1);
INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (10,36,1,0,13,25,19.7,2.7777,2.1473,0);
```

... el DDL de tabla Hive sería el siguiente:

```
CREATE EXTERNAL TABLE cassandra_bankloan_10 (row int, age int,ed int,employ int,address int,
   income int,debtinc double,creddebt double,othdebt double,default int)
STORED BY 'org.apache.hadoop.hive.cassandra.cql.CassandraCqlStorageHandler'
WITH SERDEPROPERTIES("cassandra.cf.name" = "bankloan_10","cassandra.host"="<host_cassandra>",
   "cassandra.port" = "9160")
TBLPROPERTIES ("cassandra.ks.name" = "test");
```

Orígenes de datos XML

Analytic Server proporciona soporte para datos XML a través de HCatalog.

Ejemplo

1. Correlacione el esquema XML con los tipos de datos Hive a través del lenguaje de definición de datos (DDL) de Hive, de acuerdo con las reglas siguientes.

```
CREATE [EXTERNAL] TABLE <nombre_tabla> (<especificaciones_columna>)
ROW FORMAT SERDE "com.ibm.spss.hive.serde2.xml.XmlSerDe"
WITH SERDEPROPERTIES (
    ["xml.processor.class"="<nombre_clase_procesador_xml_>",]
    "column.xpath.<nombre_columna>"="<consulta_xpath>",
     ...
    ["xml.map.specification.<nombre_elemento>"="<especificación_correlación>"
     ...
    ]
)
STORED AS
    INPUTFORMAT "com.ibm.spss.hive.serde2.xml.XmlInputFormat"
    OUTPUTFORMAT "org.apache.hadoop.hive.ql.io.IgnoreKeyTextOutputFormat"
[LOCATION "<ubicación_datos>"]
TBLPROPERTIES (
    "xmlinput.start"="<etiqueta_inicio ",
    "xmlinput.end"="<etiqueta_final>"
);
```

Nota: si los archivos XML están comprimidos por compresión Bz2, `INPUTFORMAT` debe establecerse en `com.ibm.spss.hive.serde2.xml.SplittableXmlInputFormat`. Si se comprimen con compresión CMX, debe establecerse en `com.ibm.spss.hive.serde2.xml.CmxXmlInputFormat`.

Por ejemplo, el XML siguiente...

```
<records>
  <record customer_id="0000-JTALA">
    <demographics>
      <gender>F</gender>
      <agecat>1</agecat>
      <edcat>1</edcat>
      <jobcat>2</jobcat>
      <empcat>2</empcat>
      <retire>0</retire>
      <jobsat>1</jobsat>
```

```
      <marital>1</marital>
      <spousedcat>1</spousedcat>
      <residecat>4</residecat>
      <homeown>0</homeown>
      <hometype>2</hometype>
      <addresscat>2</addresscat>
    </demographics>
    <financial>
      <income>18</income>
      <creddebt>1.003392</creddebt>
      <othdebt>2.740608</othdebt>
      <default>0</default>
    </financial>
  </record>
</records>
```

...se representaría mediante el siguiente DDL Hive.

```
CREATE TABLE xml_bank(customer_id STRING, demographics map<string,string>, financial map<string,string>)
ROW FORMAT SERDE 'com.ibm.spss.hive.serde2.xml.XmlSerDe'
WITH SERDEPROPERTIES (
    "column.xpath.customer_id"="/record/@customer_id",
    "column.xpath.demographics"="/record/demographics/*",
    "column.xpath.financial"="/record/financial/*"
)
STORED AS
    INPUTFORMAT 'com.ibm.spss.hive.serde2.xml.XmlInputFormat'
    OUTPUTFORMAT 'org.apache.hadoop.hive.ql.io.IgnoreKeyTextOutputFormat'
TBLPROPERTIES (
    "xmlinput.start"="<record customer",
    "xmlinput.end"="</record>"
);
```

Consulte “Correlación de tipos de datos XML a Hive” para obtener información adicional.

2. Cree un origen de datos Analytic Server con tipo de contenido HCatalog en la consola de Analytic Server.

Limitaciones

- Actualmente, sólo está soportada la especificación XPath 1.0.
- La parte local de los nombres calificados para los elementos y atributos se utiliza al manejar los nombres de campo de Hive. Los prefijos de espacio de nombres se ignoran.

Correlación de tipos de datos XML a Hive: Los datos modelados en XML pueden transformarse en los tipos de datos Hive utilizando los convenios documentados a continuación.

Estructuras

El elemento XML puede correlacionarse directamente con el tipo de estructura de Hive, de modo que todos los atributos se conviertan en los miembros de datos. El contenido del elemento se convierte en un miembro adicional de tipo primitivo o complejo.

datos XML

```
<result name="ID_DATUM">03.06.2009</result>
```

Datos en bruto y DDL de Hive

```
struct<name:string,result:string>
{"name":"ID_DATUM", "result":"0.3.06.2009"}
```

Matrices

Las secuencias de elementos XML pueden representarse como matrices de Hive de tipo primitivo o complejo. El ejemplo siguiente muestra cómo definir una matriz de series utilizando el contenido del elemento XML <result>.

datos XML

```
<result>03.06.2009</result>
<result>03.06.2010</result>
<result>03.06.2011</result>
```

Datos en bruto y DDL de Hive

```
result array<string>
{"result":["03.06.2009","03.06.2010",...]}
```

Correlaciones

El esquema XML no proporciona soporte nativo para correlaciones. Hay tres enfoques comunes para modelar correlaciones en XML. Para dar cabida a los diferentes enfoques, se utiliza la sintaxis siguiente:

```
"xml.map.specification.<nombre_elemento>"="<clave>-><valor>"
```

donde

nombre_elemento
El nombre del elemento XML que debe considerarse como una entrada de correlación

clave El nodo XML de clave de entrada de correlación

valor El nodo XML de valor de entrada de correlación

La especificación de correlación para el elemento XML indicado debe definirse en la sección SERDEPROPERTIES del DDL de creación de tablas de Hive. Las claves y los valores pueden definirse utilizando la sintaxis siguiente:

@attribute
La especificación @attribute permite al usuario utilizar el valor del atributo como clave o valor de la correlación.

element
El nombre de elemento puede utilizarse como valor o clave.

#content
El contenido del elemento puede utilizarse como valor o clave. Dado que las claves de correlación sólo pueden ser de tipo primitivo, el contenido complejo se convertirá a serie.

Los métodos para representar las correlaciones en XML y sus correspondientes datos en bruto y DDL de Hive son los siguientes.

Nombre de elemento a contenido
El nombre del elemento se utiliza como clave y el contenido como valor. Esta es una de las técnicas comunes y se utiliza de forma predeterminada al correlacionar XML con tipos de correlación de Hive. La limitación evidente de este enfoque es que la clave de correlación sólo puede ser de tipo serie.

datos XML

```
<entry1>value1</entry1>
<entry2>value2</entry2>
<entry3>value3</entry3>
```

Correlación, DDL Hive y datos en bruto

En este caso, no es necesario especificar una correlación debido a que el nombre del elemento se utiliza como clave y el contenido como valor de forma predeterminada.

```
result map<string,string>
{"result":{"entry1": "value1", "entry2": "value2", "entry3": "value3"}}
```

Atributo a contenido de elemento
Se utiliza un valor de atributo como clave y el contenido del elemento como valor.

datos XML

```
<entry name="key1">value1</entry>
<entry name="key2">value2</entry>
<entry name="key3">value3</entry>
```

Correlación, DDL Hive y datos en bruto

```
"xml.map.specification.entry"="@name->#content"

result map<string,string>

{"result":{"key1": "value1", "key2": "value2", "key3": "value3"}}
```

Atributo a atributo

datos XML

```
<entry name="key1" value="value1"/>
<entry name="key2" value="value2"/>
<entry name="key3" value="value3"/>
```

Correlación, DDL Hive y datos en bruto

```
"xml.map.specification.entry"="@name->@value"

result map<string,string>

{"result":{"key1": "value1", "key2": "value2", "key3": "value3"}}
```

Contenido complejo

El contenido complejo utilizado como tipo primitivo se convertía a una serie XML válida añadiendo un elemento raíz llamado <string>. Examine el XML siguiente:

```
<dataset>
 <value>10</value>
 <value>20</value>
 <value>30</value>
</dataset>
```

La expresión XPath /dataset/* dará como resultado la devolución de una serie de nodos XML <value>. Si el campo de destino es de tipo primitivo, la implementación transformará el resultado de la consulta en el XML válido añadiendo el nodo raíz <string>.

```
<string>
 <value>10</value>
 <value>20</value>
 <value>30</value>
</string>
```

Nota: la implementación no añadirá un elemento raíz <string> si el resultado de la consulta es un sólo elemento XML.

Contenido de texto

El contenido textual sólo de espacio en blanco de un elemento XML se ignora.

Vista previa y metadatos (orígenes de datos)

Cuando se pulsa **Vista previa y metadatos**, se visualiza una muestra de registros y el modelo de datos del origen de datos. Esto da la posibilidad de revisar 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 visualiza los metadatos de campo básicos. Para los orígenes de datos con tipo de contenido de archivos, el modelo de datos se genera a partir de una pequeña muestra de registros, y puede editar manualmente los metadatos de campo en esta pestaña. Para los orígenes de datos con tipo de contenido de HCatalog, el modelo de datos se genera en función de las correlaciones de campos de HCatalog y no puede editar el almacenamiento de campo en esta pestaña.

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. La frecuencia especifica que los valores de un campo deben utilizarse como ponderación de frecuencia para cada registro. El ID de registro se utiliza para identificar un registro en la salida.

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.

Valores
Muestra los valores individuales para los campos con medidas categóricas, o el rango de valores para los campos con medición continua.

Estructura
Indica si los registros del campo contienen un solo valor (primitivo) o una lista de valores.

Profundidad
Indica la profundidad de una lista; 0 es una lista de primitivas, 1 es una lista de listas, etc.

Explorar todos los valores de datos
Permite iniciar y cancelar la exploración de los valores de datos del origen de datos para determinar los valores de categoría y los límites de rango. Si una exploración está en curso, haga clic en el botón para **Cancelar exploración de datos**. La exploración de todos los valores de datos garantiza que los metadatos son correctos, pero puede tardar algún tiempo si el origen de datos tiene muchos campos y registros.

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.

Listado de proyectos

La página principal de Proyectos proporciona una lista de los proyectos de los que el usuario actual es miembro.

- Pulse el nombre de un proyecto para visualizar sus detalles y editar sus propiedades.
- Escriba en el área de búsqueda para filtrar el listado a fin de visualizar solo proyectos con la serie de búsqueda en su nombre.
- Pulse **Nuevo** para crear un proyecto con el nombre especificado en el diálogo **Añadir nuevo proyecto**. Consulte “Reglas de denominación” en la página 23 para conocer las restricciones sobre los nombres que puede dar a los proyectos.

- Pulse **Suprimir** para eliminar los proyectos seleccionados. Esta acción elimina el proyecto y suprime todos los datos asociados al proyecto del HDFS.
- Pulse **Renovar** para actualizar la lista.

Detalles de proyectos individuales

El área de contenidos se divide en las secciones contraíbles **Detalles**, **Compartición**, **Archivos** y **Versiones**.

Detalles

Nombre
Campo de texto editable que muestra el nombre del proyecto.

Nombre de visualización
Campo de texto editable que muestra el nombre del proyecto tal como se visualiza en otras aplicaciones. Si está en blanco, se utiliza el Nombre como nombre de visualización.

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

Versiones a conservar
Suprime de forma automática la versión de proyecto más antigua cuando el número de versiones sobrepasa el número especificado. El valor predeterminado es 25.

Nota: el proceso de limpieza no es inmediato, sino que se ejecuta en segundo plano cada 20 minutos.

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 miembros (cuando está sin marcar).

Pulse **Guardar** para conservar el estado actual de los valores.

Compartición

Puede compartir un proyecto añadiendo usuarios y grupos como autores o visores.

- Cuando se escribe en el cuadro de texto, se filtran usuarios y grupos que tengan la serie de búsqueda en el nombre. Seleccione el nivel de compartición y pulse **Añadir miembro** para añadir a la lista de miembros.
  - Los autores son miembros de pleno derecho de un proyecto, y pueden modificar el proyecto, así como las carpetas y los archivos contenidos en ellas. Estos usuarios, y los miembros de estos grupos, tienen acceso de escritura (nodo Exportación de Analytic Server) a este proyecto cuando se conecta a Analytic Server a través de IBM® SPSS Modeler.
  - Los visores pueden ver las carpetas y los archivos de un proyecto y definir orígenes de datos sobre los objetos de un proyecto, pero no pueden modificar el proyecto.
- Para eliminar un autor, seleccione un usuario o grupo en la lista de autores y pulse **Eliminar miembro**.

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

Nota: los cambios efectuados en la Compartición se aplican de forma inmediata y automática.

Archivos

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.

Visor de archivos

Muestra la estructura de carpetas del proyecto actual. La estructura de carpetas 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 o suprimir un proyecto, vuelva a la lista de proyectos.

- Pulse **Cargar archivo en HDFS** para cargar un archivo en el proyecto/subcarpeta actual.
- Pulse **Crear carpeta** para crear una carpeta bajo la carpeta actual, con el nombre especificado en el diálogo **Nombre de la nueva carpeta**.
- Pulse **Descargar archivo al sistema de archivos local** para descargar los archivos seleccionados en el sistema de archivos local.
- Pulse **Suprimir archivo(s) seleccionado(s)** para eliminar los archivos o carpetas seleccionados.

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.

Gestión de usuarios

Los administradores pueden gestionar los roles de usuarios y grupos a través de la página Usuarios.

El área de contenidos se divide en las secciones contraíbles **Detalles** y **Principales**.

Detalles

Nombre
Campo de texto no editable que muestra el nombre del inquilino.

Descripción
Campo de texto editable que permite facilitar un texto explicativo sobre el inquilino.

URL Este es el URL que debe suministrarse a los usuarios para iniciar la sesión en el inquilino a través de la consola de Analytic Server.

Estado

Hay inquilinos **activos** actualmente en uso. Cambiar un usuario a **Inactivo** impide que los usuarios inicien sesión en ese inquilino, pero no suprime ningún dato subyacente.

Principales

Los principales son usuarios y grupos diseñados desde el proveedor de seguridad configurado durante la configuración. Puede cambiar el rol de principales para que sean Administradores, Usuarios o Lectores.

Métricas

Le permite configurar los límites de recursos para un inquilino. Informa acerca del espacio de disco que utiliza actualmente el inquilino.

- Puede establecer una cuota máxima de espacio de disco para el inquilino. Cuando se alcanza este límite, no se puede grabar nada más en disco para este inquilino hasta que se haya borrado el espacio de disco suficiente y el uso de espacio de disco por parte del inquilino se encuentre por debajo de la cuota.
- Puede establecer una cuota máxima de espacio de disco para el inquilino. Cuando se supera este límite, los principales de este inquilino no pueden enviar ningún trabajo analítico hasta que se haya borrado el espacio de disco suficiente y el uso de espacio de disco por parte del inquilino se encuentre por debajo de la cuota.
- Puede establecer una cuota máxima de trabajos paralelos que pueden ejecutarse en este inquilino de una sola vez. Cuando se supera esta cuota, los principales no pueden enviar ningún trabajo analítico en este inquilino hasta que se haya completado el trabajo que está en ejecución.
- Puede establecer el número máximo de campos que puede tener un origen de datos. El límite se comprueba cada vez que se crea o actualiza un origen de datos.
- Puede establecer el número máximo de registros que puede tener un origen de datos. El límite se comprueba siempre que se crea o actualiza un origen de datos; por ejemplo, al añadir un archivo nuevo o cambiar los valores de un archivo.
- Puede establecer el tamaño máximo del archivo en MB. El límite se comprueba al cargar un archivo.

Configuración del proveedor de seguridad

Le permite especificar el proveedor de autenticación de usuarios. **Valor predeterminado** utiliza el proveedor del inquilino predeterminado, el cual se establece durante la instalación y configuración. **LDAP** permite autenticar usuarios con un servidor LDAP externo, tal como Active Directory u OpenLDAP. Especifique uno de los valores para el proveedor y, opcionalmente, especifique valores de filtro para controlar los usuarios y grupos disponibles en la sección Principales.

Reglas de denominación

Para cualquier elemento que pueda recibir un nombre exclusivo en Analytic Server, como por ejemplo orígenes de datos y proyectos, dichos nombres están sujetos a las normas siguientes.

- En un único inquilino, los nombres deben ser exclusivos en objetos del mismo tipo. Por ejemplo, dos orígenes de datos no pueden denominarse insuranceClaims, pero un origen de datos y un proyecto podrían llamarse insuranceClaims.
- Los nombres son sensibles a las mayúsculas y minúsculas. Por ejemplo, insuranceClaims e InsuranceClaims se consideran nombres exclusivos.
- Los nombres ignoran los espacios en blanco iniciales y finales.
- Los caracteres siguientes no son válidos en los nombres.

  `~, #, %, &, *, {, }, \\, :, <, >, ?, /, |, ", \t, \r, \n`

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 analítica que se ejecute con Analytic Server:

1. La secuencia debe empezar con un nodo de origen de Analytic Server.
2. Construya el centro de la secuencia en la interfaz de Modeler como haría normalmente, seleccione nodos de proceso (operaciones de Campo o Registro) soportados por Analytic Server. Hay un panel de Analytic Server en la paleta de Modeler que muestra los nodos soportados.
3. Hay un par de opciones para finalizar la secuencia.
   - Seleccione un nodo de terminal (Salida, Gráfico, Exportación o Modelado) que esté soportado por Analytic Server. En este caso, Modeler incorpora la secuencia entera a Analytic Server. Analytic Server orquestra los trabajos necesarios en el clúster de Hadoop y pone los resultados a disposición de Modeler. Modeler toma los resultados y los presenta al usuario, de la misma manera que si la secuencia se procesara localmente.
   - Si selecciona un nodo de terminal que no está soportado por Analytic Server, Modeler incorpora a Analytic Server tanta parte de la secuencia como sea posible y, a continuación, empieza a extraer los registros de Hadoop. Tenga en cuenta que Analytic Server puede puntuar algunos modelos que no pueden construirse actualmente con Analytic Server. Esto significa que puede estructurar una secuencia para que tome un sub-ejemplo válido estadísticamente de datos masivos con Analytic Server y, a continuación, construir un modelo "localmente" en Modeler. El nugget de modelo resultante podrá incluirse a continuación en una secuencia de puntuación que se ejecute por completo en Analytic Server.

   Nota: El número máximo de registros que SPSS Modeler descargará de Hadoop puede configurarse en las propiedades de secuencia de Analytic Server.

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.

Nota: Consulte la documentación SPSS Modeler para obtener información sobre el funcionamiento ordinario de estos nodos.

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 se 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 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

Todas las operaciones de registro están soportadas, con la excepción de los nodos Resolución de TS y Cuadros de espacio-tiempo. A continuación se detalla la funcionalidad de nodo soportada.

Seleccionar

- Soporta el mismo conjunto de funciones soportado por el Nodo de derivación.

Muestrear

- No se soporta el muestreo a nivel de bloque.
- No se soportan los métodos de muestreo complejos.
- El primer muestro `n` con `"Descartar muestreo"` no está soportado.
- El primer muestreo `n` con n `N>20000` no está soportado.
- El muestreo 1 de n no está soportado cuando el `"Tamaño máximo de muestreo"` no está establecido.
- El muestreo 1 de n no está soportado cuando `N * "Tamaño máximo de muestreo" > 20000`.
- El muestreo a nivel de bloque de % aleatorio no está soportado.
- Actualmente, el % aleatorio permite suministrar un valor de inicio.

Agregación

- Las claves contiguas no están soportadas. Si va a reutilizar una secuencia existente configurada para ordenar los datos y luego utilizar este valor en el nodo Agregación, cambie la secuencia para eliminar el nodo Ordenar.
- Las estadísticas de orden (mediana, primer cuartil, tercer cuartil) se calculan de forma aproximada y están soportadas a través de la pestaña Optimización.

Ordenar

- La pestaña Optimización no está soportada.

En un entorno distribuido, hay un número limitado de operaciones que conservan el orden de registros establecido por el nodo Ordenar.

- Un nodo Ordenar seguido de uno Exportar produce un origen de datos ordenado.
- Un nodo Ordenar seguido de uno Muestrear con muestreo de **Primer** registro devuelve los *N* primeros registros.

En general, debe colocarse un nodo Ordenar lo más cerca posible de las operaciones que necesitan los registros ordenados.

Fusionar

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

Transformación R

La sintaxis R en el nodo debe constar de operaciones record-at-a-time (registro cada vez).

Operaciones de campo

Todas las operaciones de campo están soportadas, con la excepción de los nodos Anonimizar, Transponer, Intervalos de tiempo e Historial. A continuación se detalla la funcionalidad de nodo soportada.

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.

Derivar

- Se soportan todas las funciones de Derivar, a excepción de las funciones de secuencia.
- Derivar un nuevo campo como un Recuento es esencialmente una operación de secuencia y por lo tanto, no está soportada.
- 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.

Relleno

- Soporta el mismo conjunto de funciones soportado por el Nodo de derivación.

Intervalos

La siguiente funcionalidad no está soportada.

- Intervalos óptimos.
- Rangos.
- Cuantiles -> Creación de cuantiles: suma de valores.
- Cuantiles -> Empates: mantener en actual y asignar aleatoriamente.
- Cuantiles ->N personalizados: valores por encima de 100 y cualquier valor N donde el 100% de N no sea igual a cero.

Análisis de RFM

- No se soporta la opción Mantener en actual para tratar los 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

Los siguientes nodos de Modelado están soportados: Time Series, TCM, árbol-AS, árbol C&R, Quest, CHAID, Lineal, Lineal-AS, Red neuronal, GLE, LSVM, TwoStep-AS, Árboles aleatorios, STP y Reglas de asociación. A continuación se detalla la funcionalidad de estos nodos.

Lineal Al crear modelos basados en datos masivos, lo normal es cambiar el objetivo a Conjuntos de datos muy grandes o bien especificar divisiones.

- 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.
- 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

Al crear modelos basados en datos masivos, lo normal es cambiar el objetivo a Conjuntos de datos muy grandes o bien especificar divisiones.

- 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.
- 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

Al crear modelos basados en datos masivos, lo normal es cambiar el objetivo a Conjuntos de datos muy grandes o bien especificar divisiones.

- 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.
- No se soportan las sesiones interactivas.
- La estadística de precisión no se calcula para los modelos divididos.
- Cuando está presente un campo de división, los tres modelos creados localmente en Modeler son ligeramente distintos de los tres modelos creados mediante Analytic Server, y de este modo producen puntuaciones distintas. Los algoritmos en los dos casos son válidos; los algoritmos que Analytic Server utiliza son simplemente más recientes. Dado el hecho de que los algoritmos de árbol tienden a tener muchas reglas heurísticas, la diferencia entre los dos componentes es normal.

Puntuación de modelos

Todos los modelos soportados para el modelado también están soportados para la puntuación. Además, los nuggets de modelo creados localmente para los nodos siguientes están soportados para la puntuación: C&RT, Quest, CHAID, Lineal y Red neuronal (independientemente de que el modelo sea estándar, "boosting y bagging" o para conjuntos de datos de tamaño muy grande), Regresión, C5.0, Logística, Genlin, GLMM, Cox, SVM, Red bayesiana, Bietápico, KNN, Lista de decisiones, Discriminante, Autoaprendizaje, Detección de anomalías, Apriori, Carma, K-medias, Kohonen, 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

R La sintaxis R en el nugget debe constar de operaciones record-at-a-time (registro cada vez).

Salida Los nodos Matriz, Análisis, Auditoría de datos, Transformar, Establecer globales, Medias y Tabla están soportados. A continuación se detalla la funcionalidad de nodo soportada.

Auditoría de datos

El nodo Auditoría de datos no puede producir la modalidad de campos continuos.

Medias

El nodo Medias no puede producir un error estándar o un intervalo de confianza del 95 %.

Tabla El nodo Tabla está soportado mediante la escritura de un origen de datos Analytic Server temporal que contiene los resultados de operaciones anteriores. A continuación, el nodo Tabla pagina el contenido de dicho origen de datos.

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.

Mejores prácticas

Retrotracción a HCatalog/Hive

Cuando se trabaja con datos en una tabla Hive particionada, puede estructurar la secuencia de Modeler para poder retrotraer la selección de las particiones deseadas en Hive.

1. Inicie la secuencia con un nodo de origen de Analytic Server que haga referencia al origen de datos de HCatalog/Hive.
2. Conecte a un nodo Seleccionar que seleccione registros SOLO para campos que se utilicen como campos de partición en la tabla Hive. Si en la expresión de este nodo Seleccionar se hace referencia a campos que no se utilizan como campos de partición, la secuencia no se retrotraerá a HCatalog/Hive.
3. Conecte a otros nodos como haría normalmente.

Capítulo 3. Resolución de problemas

En esta sección se describen algunos problemas de uso comunes y cómo arreglarlos.

Orígenes de datos

Los filtros definidos en columnas particionadas de orígenes de datos de HCatalog no se respetan

Este es un problema que ha aparecido en algunas versiones de Hive y puede aparecer en las situaciones siguientes.

- Si define un origen de datos de HCatalog y especifica un filtro en la definición de origen de datos.
- Si crea una secuencia de Modeler con un nodo Filtrar que hace referencia a la columna de tabla particionada.

La solución temporal es añadir un nodo Derivar a la secuencia de Modeler que crea un nuevo campo con valores iguales a la columna particionada. El nodo Filtrar debe hacer referencia a este nuevo campo.

Oracle NoSQL

Se encuentran errores "Execution failed" (la ejecución ha fallado) al conectar con un origen de datos Oracle NoSQL

El problema es el resultado del manejador de almacenamiento HiveKVStorageHandler.jar no actualizado. Debe utilizarse un manejador de almacenamiento actualizado. El archivo actualizado lo puede encontrar en https://github.com/dvasilen/HiveKVStorageHandler3/raw/HADOOP_2.6-HIVE-1.2.0-KV-3.3.4/release/hive-kv-storage-handler-1.2.0-3.3.4.jar

`hive-kv-storage-handler-1.2.0-3.3.4.jar`

1. Copie el archivo JAR en el directorio de Hive `{HIVE_HOME}/auxlib` y el directorio de Analytic Server `{AS_ROOT}/ae_wlpserver/usr/servers/aeserver/apps/AE_BOOT.war/WEB-INF/lib`.
2. Ejecute `{AS_ROOT}/bin/hdfsUpdate.sh` para propagar lo cambios a HDFS.
3. Reinicie el Analytic Server para que los cambios entren en vigor.

Nota: Se recomienda el manejador de clase de almacenamiento `oracle.kv.hadoop.hive.table.TableStorageHandler` cuando se utiliza la base de datos Oracle NoSQL 3.0. La clase requiere que los usuarios organicen datos con una metáfora de tabla.

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:

© el nombre de su empresa) (año). Partes de este código se derivan de IBM Corp. Sample Programs.

© Copyright IBM Corp. _especifique el año o años_. Reservados todos los derechos.

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