IBM SPSS Analytic Server
Versión 3.2.2

Guía del administrador

Nota

Antes de utilizar esta información y el producto al que da soporte, lea la información del apartado "Avisos" en la página 31.

Información sobre el producto

Esta edición se aplica a la versión 3, release 2, modificación 2 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. Gestión de inquilinos

Los inquilinos proporcionan una división de alto nivel de usuarios, proyectos y orígenes de datos, de modo que los objetos no se pueden compartir entre inquilinos. Cada usuario accede al sistema en el contexto del inquilino que le ha sido asignado.

En la consola de Analytic Server puede gestionar los inquilinos y asignar usuarios a los inquilinos. La vista de la página Inquilinos depende del rol del usuario que ha iniciado sesión en la consola:

- El administrador "superusuario" que se configura durante la instalación es el administrador de los inquilinos. Este usuario es el único que puede crear inquilinos y editar las propiedades de cualquiera de ellos.
- Los usuarios con el rol de administrador pueden editar las propiedades del inquilino con el que hayan iniciado sesión.
- Los usuarios con el rol de usuario no pueden editar las propiedades de ningún inquilino. La página Inquilinos se les oculta.
- Los usuarios con el rol de lector no pueden editar los orígenes de datos ni incluso iniciar sesión en la consola de Analytic Server.

Los administradores pueden acceder a las páginas Proyectos y Orígenes de datos, y gestionar los proyectos y orígenes de datos para administrarlos o realizar una limpieza. Consulte la *Guía del usuario de IBM SPSS Analytic Server* para obtener más información.

Lista de inquilinos

La página principal Inquilinos muestra los inquilinos existentes en una tabla. Sólo el administrador "superusuario" puede realizar ediciones en esta página.

- Pulse el nombre de un inquilino para visualizar sus detalles y editar sus propiedades.
- Pulse el URL de un inquilino para abrir la consola en el contexto de dicho inquilino.

  Nota: Se finalizará su sesión y deberá iniciar sesión con las credenciales válidas para dicho inquilino.
- Escriba en el área de búsqueda para filtrar la lista de forma que solo se muestren los inquilinos que tienen la serie de búsqueda en su nombre.
- Pulse **Nuevo**para crear un inquilino con el nombre que especifique en el diálogo **Añadir inquilino nuevo**. Consulte la sección “Reglas de denominación ” en la página 2 para obtener información acerca de los nombres que puede asignar a los inquilinos.
- Pulse **Suprimir** para eliminar el inquilino.
- Pulse **Renovar** para actualizar la lista.

Detalles del inquilino individual

El área de contenido se divide en varias secciones contraíbles.

Detalles

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

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

URL
Es el URL que se facilita a los usuarios para que inicien sesión en el inquilino a través de la consola de Analytic Server, y para configurar el servidor de SPSS Modeler. Consulte la *Guía de instalación y configuración de IBM SPSS Analytic Server* para obtener detalles acerca de cómo configurar SPSS Modeler.

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 obtenidos del proveedor de seguridad configurado durante la instalación. Se pueden añadir principales a un inquilino como Administradores, Usuarios 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 **Administrador**, **Usuario** o **Lector** en la lista desplegable para asignar su rol en el inquilino. Pulse **Añadir participante** para añadirlo a la lista de autores.
- Para eliminar un participante, seleccione un usuario o grupo en la lista de miembros y pulse **Eliminar participante**.

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 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. Iniciación de usuarios

Indique a los usuarios que vayan a `http://<host>:<puerto>/<raíz-contexto>/admin/<inquilino>` y escriban su nombre de usuario y contraseña para iniciar sesión en la consola de Analytic Server.

Nota: El nombre de usuario que se especifica durante la solicitud de inicio de sesión en la consola de Analytic Server se especifica sin el sufijo de nombre de dominio. Como resultado, cuando se definen varios dominios, se presenta a los usuarios una lista desplegable de **Dominios** que les permite seleccionar el dominio adecuado. Cuando sólo hay un dominio definido, no se presenta a los usuarios ninguna lista despegable **Dominios** al iniciar sesión en Analytic Server.

<host>
La dirección del host de Analytic Server.

<puerto>
El puerto donde escucha Analytic Server. De forma predeterminada, este valor es 9080.

<raíz_contexto>
La raíz de contexto de Analytic Server. De forma predeterminada es analyticserver.

<inquilino>
En un entorno de varios inquilinos, el inquilino al que pertenece. En un entorno de un solo inquilino, el inquilino predeterminado es **`ibm`**.

Por ejemplo, si la máquina host tiene la dirección IP 9.86.44.232, ha creado un inquilino "mycompany" y ha añadido usuarios a este último y los demás valores se han dejado en sus valores predeterminados, los usuarios deben navegar a `http://9.86.44.232:9080/analyticserver/admin/mycompany` para acceder a la consola de Analytic Server.

Capítulo 3. Nombres de trabajos de Analytic Server

Analytic Server genera trabajos de reducción de correlaciones y Spark que se pueden supervisar mediante la interfaz de usuario de gestión de recursos de clúster de Hadoop.

El nombre del trabajo de reducción de correlaciones tiene la estructura siguiente.

```
AS/{nombre inquilino}/{nombre usuario}/{nombre algoritmo}
```

{nombre inquilino}
Es el nombre del inquilino bajo el que se ejecuta el trabajo.

{nombre usuario}
Es el usuario que ha solicitado el trabajo.

{nombre algoritmo}
Es el algoritmo principal del trabajo. Tenga en cuenta que una sola secuencia puede generar varios trabajos de reducción de correlaciones. Del mismo modo, si hay varias operaciones en una secuencia, éstas pueden estar contenidas en un solo trabajo de reducción de correlaciones.

Todos los trabajos de reducción de correlaciones se muestran en la interfaz de usuario del gestor de recursos. Se inicia una aplicación Spark independiente para cada Analytic Server. Abra al interfaz de usuario de la aplicación Spark para supervisar los trabajos Spark (los nombres de los trabajos se muestran en la columna **Descripción**).

Capítulo 4. Propiedades personalizadas de Analytic Server

Hay las siguientes propiedades personalizadas definidas, o se pueden configurar, en el archivo `analytic.cfg` de Analytic Server.

Tabla 1. Propiedades personalizadas de Analytic Server

Nombre de propiedad	Tipo	Valor predeterminado	Valores permitidos	Descripción
admin.username	Cadena			Define el nombre del usuario administrador de Analytic Server.
ae.cluster.ha.cascade.failure.protection	Booleano	TRUE		Cuando está habilitado (true), el gestor de trabajos del clúster impide que los trabajos que han fallado en varios miembros del clúster bloqueen todos los servidores del clúster. Esto se hace suspendiendo permanentemente el trabajo que da problemas o asignando que solo se pueda ejecutar en un servidor en cuarentena designado.
ae.cluster.heapdump_onexit.filename	Cadena			Si se ha especificado un nombre de archivo. la JVM de Analytic Server escribe un volcado de almacenamiento dinámico (al archivo especificado) cuando se producen varios sucesos **CPU_Starvation**.
ae.cluster.job.artifacts.cleanup.delay.minutes	Entero	5		El tiempo que Analytic Server espera después de que se complete un trabajo antes de limpiar los artefactos relacionados con el trabajo en zookeeper.
ae.cluster.quarantine.server.name	Cadena			En entornos en clúster, identifica el servidor de Analytic Server que se utiliza para ejecutar trabajos en cuarentena (trabajos que superan el umbral de número de errores).

Tabla 1. Propiedades personalizadas de Analytic Server (continuación)

Nombre de propiedad	Tipo	Valor predeterminado	Valores permitidos	Descripción
ae.cluster.queue.callback.threadpool.size	Entero	20		Tamaño de **ThreadPool** que utiliza el servicio de clúster al consumir mensajes de clúster interproceso enviados a través de zookeeper.
ae.cluster.thread.scheduler.delay.detector	Entero	30		Detecta problemas de rendimiento locales (por ejemplo, registro de mensajes de inanición de CPU).
ae.cluster.thread.scheduler.detect.error	Entero	10		Detecta problemas de rendimiento locales (por ejemplo, registro de mensajes de inanición de CPU).
as.db.connect.method	Cadena	Basic	Kerberos|Basic	Identifica el método de conexión del origen de datos de la base de datos de Analytic Server.
as.spark.driver.cleanup.delay	Entero	2		El tiempo (en minutos) después de un cierre de sesión que se tarda en finalizar la JVM cliente de Spark.
cleanup.delay	Entero	20		El número de minutos de retardo entre cada ejecución de limpieza en segundo plano (por ejemplo, de archivos de proyecto).
default.project.versions.tokeep	Entero	25		El número de versiones de proyecto a conservar al limpiar versiones de proyecto anteriores.
distrib.fs.root	Cadena	/user/as_user/ analytic-root		La carpeta base de Analytic Server para el sistema de archivos distribuido.
hive.precheckPermission	Booleano	TRUE		Cuando está establecido en **TRUE**, Analytic Server comprueba los permisos de archivos HDFS para validar los derechos de acceso de usuario a la ubicación de los datos de la tabla de base de datos.
hive.sql.check	Booleano	FALSE		Añade el prefijo EXPLAIN a sentencias SQL generadas.

Tabla 1. Propiedades personalizadas de Analytic Server (continuación)

Nombre de propiedad	Tipo	Valor predeterminado	Valores permitidos	Descripción
io.sort.mb	Entero	10		La cantidad total de memoria de almacenamiento intermedio a utilizar (en megabytes) al ordenar archivos. De forma predeterminada, cada secuencia de fusión tiene 1MB asignado, lo que debería minimizar las búsquedas. Para obtener más información, consulte http://hadooptutorial.info/hadoop-performance-tuning/.
java.security.krb5.conf	Cadena			La ubicación del archivo `krb5.conf` de Kerberos.
jndi.aedb.driver	Cadena			La clase de controlador del metastore de Analytic Server.
jndi.aedb.password	Cadena			La contraseña del metastore de Analytic Server.
jndi.aedb.url	Cadena			La cadena de conexión JDBC del repositorio del metastore de Analytic Server.
jndi.aedb.username	Cadena			El nombre de usuario del metastore de Analytic Server.
join.small.data.size	Entero	1048576		La cantidad máxima de datos (en bytes) que el motor analítico intentará unir en un algoritmo de lado de correlación.
mapred.child.java.opts	Cadena	"-server"		Controla los tamaños de almacenamiento dinámico de JVM para las correlaciones y reduce las tareas que se ejecutan en Hadoop. Establezca este parámetro en el valor más grande que los nodos del clúster puedan manejar.
max.asl.size	Entero	20971520		Tamaño máximo permitido (en bytes) para el programa ASL.
max.datamodel.size	Entero	20971520		Tamaño máximo permitido (en bytes) para la cadena XML del modelo de datos.

Tabla 1. Propiedades personalizadas de Analytic Server (continuación)

Nombre de propiedad	Tipo	Valor predeterminado	Valores permitidos	Descripción
mmr.taskparallel.targets.threshold	Entero	100		M3R procesa los trabajos cuando la proporción de destinos/núcleos es menor que este umbral. .
mmr.threads	Entero	4		El número de hebras que se van a utilizar para los trabajos de reducción de correlación en memoria (M3R). (*2)
mmr.upper.bound.threshold	Numérico	100		La cantidad máxima de datos que procesará este M3R. Las cantidades más grandes de datos las procesa Hadoop o Spark.
nested.groups	Cadena	enabled	enabled|disabled| null (no definido)	Significa que se están utilizando grupos anidados en LDAP.
node.max.jobs	Entero	50		El número máximo de trabajos que se pueden ejecutar de forma simultánea en un miembro de clúster de Analytic Server.
orchestrator.thread.pool	Entero	30		El tamaño del orquestador de agrupación de hebras que se utiliza al enviar trabajo de motor analítico (por ejemplo, **EngineCommands**).
orchestrator.thread.pool.fixed	Booleano	TRUE		Especifica si la agrupación de hebras del orquestador es elástica o fija.
preferred.mapreduce	Cadena	spark	m3r|hadoop|spark	Define el motor de reducción de MapReduce preferido a utilizar.
resource.pool.default(*1)	Cadena			Establece el valor de **spark.scheduler.pool** cuando no se encuentra ninguna correlación de consumidor en **resource.pool.mapping**. Para obtener más información, consulte https://spark.apache.org/docs/latest/job-scheduling.html.
resource.pool.enabled	Booleano	FALSE		Habilita el uso de definiciones de correlación de cola YARN (**yarn.queue.mapping**).

Tabla 1. Propiedades personalizadas de Analytic Server (continuación)

Nombre de propiedad	Tipo	Valor predeterminado	Valores permitidos	Descripción
resource.pool.mapping(*1)	Correlación		(a:b,c:d....) donde a y c son nombres de consumidores de AS, b, y d son nombres de agrupaciones de recursos de yarn	Correlaciona nombres de consumidores de Analytic Server con nombres de agrupaciones de recursos YARN.
session.max.inactivity.time	Entero	14400		El valor del tiempo de espera de sesión HTTP (en segundos). El valor predeterminado es 14400 segundos (4 horas).
spark.cache	Booleano	TRUE		Determina si los Spark RDD almacenados en caché se utilizan en la JVM del ¡ cliente Spark. Por motivos de rendimiento, esto se debería dejar con el valor predeterminado **TRUE**.
spark.dependency.exclude.regex	Cadena		Una expresión regular válida para excluir los archivos *.jar de la classpath de la JVM del cliente Spark	Excluye archivos *.jar problemáticos de la classpath de la JVM del cliente Spark.
spark.version	Cadena	2.x		La versión de Spark en la pila HDP/CDH que Analytic Server utiliza para ejecutar trabajos de Spark.
split.sort.mb	Entero	100		Establece el valor **io.sort.mb**. Para obtener más información, consulte https://hadoop.apache.org/docs/r2.4.1/hadoop-mapreduce-client/hadoop-mapreduce-client-core/mapred-default.xml.
yarn.queue.default	Cadena	default		El nombre de cola YARN predeterminado que se utiliza cuando no se encuentra ninguna correlación válida en **yarn.queue.mapping**.
yarn.queue.mapping	Correlación		(a:b,c:d....) donde a y c son nombres de usuario o de consumidor de Analytic Server (según determine **yarn.queue.mode**), b y d son nombres de cola yarn	Correlaciona un nombre (sea de usuario o de consumidor) con una cola YARN.

Tabla 1. Propiedades personalizadas de Analytic Server (continuación)

Nombre de propiedad	Tipo	Valor predeterminado	Valores permitidos	Descripción
yarn.queue.mode	Cadena		user|tenant	Determina si se utiliza **userName** o **consumerName** para **mapreduce.job.queuename**. Para obtener más información, consulte https://hadoop.apache.org/docs/r2.4.1/hadoop-mapreduce-client/hadoop-mapreduce-client-core/mapred-default.xml.

Capítulo 5. Mejores prácticas y recomendaciones de IBM SPSS Analytic Server

Las siguientes secciones proporcionan las mejores prácticas y recomendaciones de Analytic Server en relación a los orígenes de datos, la configuración de clústeres y las secuencias de IBM SPSS Modeler.

Orígenes de datos

Analytic Server da soporte a los siguientes tipos de datos:

- Orígenes de datos basados en archivo, por ejemplo, archivos delimitados, de texto fijo o Microsoft Excel.
- Bases de datos relacionales, como Db2, Oracle, Microsoft SQL Server, Teradata, Postgres, Netezza, MySQL, y Amazon Redshift.
- Orígenes de datos Hive/HCatalog que incluyen todos los tipos de datos incorporados (por ejemplo, ORC y Parquet), así como cualquier tipo de datos personalizados para los que hay disponible una implementación de Hive Serializer-Deserializer adecuada. Además, Analytic Server puede configurarse para acceder a bases de datos no SQL, como HBase, MongoDB, Cassandra Accumulo, Oracle NoSQL, y otras bases para las que hay disponible una implementación del manejador de almacenamiento Hive adecuada.

  Nota: El soporte para Parquet se limita a la lectura y la adición de tablas Hive. Cuando se necesita sobrescribir información de la tabla, se crea una nueva tabla porque el proceso de sobrescribirla puede hacer que se cambien valores de los datos, así como el modelo de datos.
- Orígenes de datos de tipo geoespacial (basados en shape y basados en servicios de correlación).

Limitaciones de Analytic Server en orígenes de datos de Hive/HCatalog

- Si es necesaria la retrotracción de Hive para el nodo Seleccionar de SPSS Modeler, la expresión de filtrado puede hacer referencia únicamente a columnas particionadas del tipo STRING. A partir de Analytic Server 3.0, se ha añadido el soporte del tipo de datos para las siguientes columnas particionadas: TINYINT, SMALLINT, INT, BIGINT. La expresión de filtro estático que se especifica para el origen de datos de Hive puede tener expresiones de filtrado para columnas particionadas de cualquier tipo de datos.
- Analytic Server no da soporte a orígenes de datos HCatalog basados en vistas Hive. Todos los demás tipos de origen de datos (como por ejemplo Hive SQL) basados en vistas Hive están soportados.

Orígenes de datos Hive con grandes volúmenes de datos en Cloudera

Se recomienda habilitar Apache Spark para acelerar el proceso de orígenes de datos Hive con grandes volúmenes de datos en Cloudera.

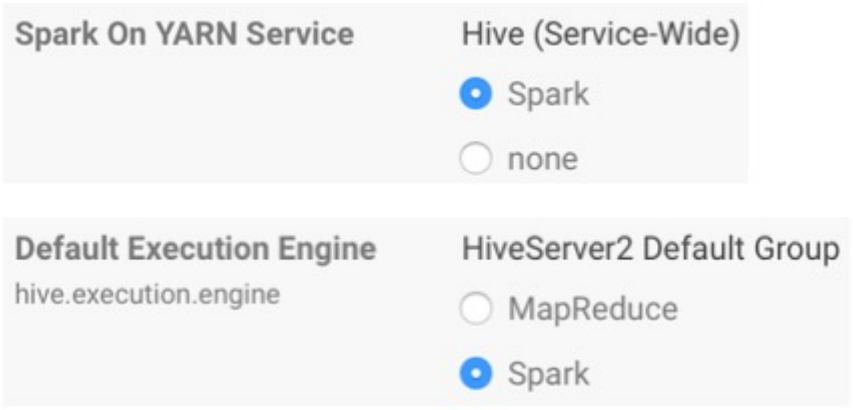

Figura 1. Valores de Spark

Configuración de clúster - seguridad

Suplantación de Kerberos

Antes de la versión 3.0.1, las instancias Analytic Server utilizaban un nombre de principal de usuario en el archivo de tablas de clave de Analytic Server para autenticar operaciones HDFS cuando la seguridad de Kerberos está habilitada. A partir de la versión 3.0.1, Analytic Server utiliza un nombre de principal de servicio en la tabla de claves de Analytic Server junto con el nombre de usuario solicitante (del usuario que realiza la solicitud Rest) para autenticar operaciones HDFS que utilizan la suplantación de Kerberos. Es necesario Analytic Server 3.0.1, o superior, para añadir atributos de configuración de suplantación a HDFS (o las configuraciones de servicio Hive) cuando se ejecuta en un clúster habilitado para Kerberos. En el caso de HDFS, deben añadirse las siguientes propiedades al archivo HDFS `core-site.xml`:

```
hadoop.proxyuser.<nombre_principal_servicio_analytic_server> .hosts = *
hadoop.proxyuser.<nombre_principal_servicio_analytic_server> .groups = *
```

donde `<nombre_principal_servicio_analytic_server>` es el valor predeterminado de `as_user` que está especificado en el campo `Analytic_Server_User` de la configuración de Analytic Server.

Las siguientes propiedades también deben añadirse al archivo HDFS `core-site.xml` en casos en los que el acceso a los datos se realiza desde HDFS vía Hive/HCatalog:

```
hadoop.proxyuser.hive.hosts = *
hadoop.proxyuser.hive.groups = *
```

Autenticación entre varios dominios Kerberos

Analytic Server da soporte a la autenticación entre varios dominios Kerberos. Para habilitar esta característica, primero debe asegurarse de que la autenticación entre varios dominios KDC está habilitada y, a continuación, añada el siguiente valor a la sección de configuración de Ambari de Analytic Server en **Custom analytics.cfg**:

```
kerberos.user.realm.trim = true
```

Configuración del clúster - valores y resultados del ajuste de rendimiento

Configuración de Spark

Analytic Server utiliza la modalidad `yarn-client` para interactuar con YARN y ejecutar trabajos Spark en el clúster de Hadoop.

Configuración personalizada de Analytic Server:

- Los valores de Ambari están definidos en la sección **Custom analytics.cfg** de la configuración de Ambari de Analytic Server.

- Los valores de Cloudera están ubicados en la sección **Analytic Server Advanced Configuration Snippet (Safety Valve) for analyticserver-conf/config.properties** de Cloudera Manager.

1. Contemple la posibilidad de incrementar el valor de la configuración **spark.driver.memory** añadiendo un elemento de configuración en la configuración personalizada de Analytic Server (cuando no se establezca explícitamente, el valor predeterminado es 1g). Por ejemplo:

```
spark.driver.memory=2g
```

2. Seleccione de entre uno de los siguientes Analytic Server con opciones de uso de recursos de Spark.

   - **Opción A: Configuración de asignación de recursos estáticos**

     Existen 3 parámetros que se deben configurar en la configuración personalizada de Analytic Server:

```
spark.executor.instances
spark.executor.cores
spark.executor.memory
```

     Los pasos siguientes describen cómo determinar los valores de parámetro.

     a. Establezca el porcentaje, en términos de CPU y memoria, que Analytic Server puede asignar de forma permanente para Spark. El resultado es un número específico de núcleos (C) y una cantidad fija de memoria que puede utilizarse en cada máquina (M).

     b. Establezca el número de ejecutores (E) que pueden ejecutarse en cada máquina. Estos autores se ejecutan como distintos contenedores de Hadoop (procesos) en cada nodo de clúster. Normalmente, un valor superior a 2 es adecuado, pero el valor debe ser inferior al número total de núcleos. La memoria que se asigna a Spark se divide entre estos ejecutores, de manera que seleccionar un valor elevado para este parámetro reducirá la cantidad de memoria que se asigna a cada contenedor.

     c. Establezca el número de núcleos que se utilizan para cada ejecutor (CE). Normalmente este valor es C/E (el número de núcleos de cada máquina que se asignan a la aplicación Spark divididos por el número total de ejecutores).

     d. Establezca la cantidad de memoria que se utiliza para cada ejecutor (ME). Esto suele ser M/E.

        Nota: El número de ejecutores y núcleos que se utilizan debe equilibrarse de manera que la cantidad de memoria de cada ejecutor debe ser mayor que 3G * CE. Cada núcleo de cada ejecutor debe estar asignado al menos a 3G de memoria que se utilizará como almacenamiento o memoria de cálculo.

```
spark.executor.instances = <E>*N /<E> // value established in step b where N is the number of compute nodes
spark.executor.cores = <CE> // value established in step c
spark.executor.memory = <ME> // value established in step d
```

     Figura 2. Valores Spark de analytics.cfg personalizados

   - **Opción B: Configuración de asignación de recursos dinámica**

     Al utilizar esta opción, todos los ejecutores asignados por YARN se incrementan/disminuyen de forma dinámica según los recursos reales disponibles del clúster completo.

     La configuración mínima es:

```
spark.dynamicAllocation.enabled = true
spark.shuffle.service.enabled = true
```

     Una configuración típica es:

```
spark.default.emitter.class = com.spss.ae.spark.ListEmitter
spark.default.emitter.compressed = false
spark.dynamicAllocation.enabled = true
spark.executor.cores = 4
spark.executor.memory = 16g
spark.io.compression.codec = snappy
spark.rdd.compress = true
spark.shuffle.service.enabled = true
```

Notas:

- `spark.executor.instances = <E>` no debe utilizarse; si no, se empleará la asignación de recursos estáticos.
- Las consideraciones sobre los núcleos ejecutores y los valores de memoria son los mismos que la Opción A.

3. Puede inhabilitar la memoria caché Spark en la configuración personalizada de Analytic Server utilizando los valores siguientes:

```
spark.cache=false
spark.storage.memoryFraction = 0.3
```

Figura 3. Valores de memoria caché Spark de analytics.cfg personalizados

La memoria caché Spark no debe inhabilitarse cuando se utilizan secuencias de IBM SPSS Modeler grandes. Inhabilitar la memoria caché en esta instancia genera secuencias de ejecución más lenta, pero impide que haya condiciones de falta de memoria que pueden aparecer cuando la cantidad de memoria especificada por ejecutor es pequeña.

Configuración de JVM

Valores de Ambari:

1. En la configuración Ambari de Analytic Server, establezca la cantidad de memoria que el servidor puede utilizar para el proceso local. El valor predeterminado (2 GB) puede utilizarse con seguridad para secuencias pequeñas y medianas, pero para secuencias más grandes debe utilizarse un valor de tamaño de almacenamiento dinámico más alto (por ejemplo, 10 GB).

   Analytic Server > **Configuración** > **Advanced analytic-jvm-options**

2. Sustituya `-Xmx2048M` por `-Xmx10G`, guarde la configuración y reinicie Analytic Server.

Figura 4. Valores de Advanced analytic-jvm-options

Valores de Cloudera:

1. En Cloudera Manager, vaya a la pestaña **Configuración** del servicio de Analytic Server y actualice el control `jvm-options` para establecer la cantidad de memoria que el servidor puede utilizar para el proceso local. El valor predeterminado (2 GB) puede utilizarse con seguridad para secuencias pequeñas y medianas, pero para secuencias más grandes debe utilizarse un valor de tamaño de almacenamiento dinámico más alto (por ejemplo, 10 GB).

   Analytic Server service > **Configuración** > **jvm-options**

2. Sustituya `-Xmx2048M` por `-Xmx10G`, guarde la configuración y reinicie Analytic Server.

Configuración de YARN MapReduce2:

- Si debe ejecutar trabajos MapReduce en paralelo con trabajos Spark para la ejecución de Analytic Server, el clúster de YARN debe configurarse para que tenga al menos 4 GB de memoria por cada contenedor de YARN.

Configuración de Zookeeper:

- Cloudera le requiere que actualice manualmente la configuración de Zookeeper. Para obtener más información, consulte https://www.cloudera.com/documentation/enterprise/5-4-x/topics/cdh_ig_zookeeper_server_maintain.html.
- Si utiliza secuencias de SPSS Modeler complejas o datos amplios (con un gran número de campos), puede experimentar problemas con trabajos fallidos debido a que hay una conexión de Analytic Server–Zookeeper interrumpida. El problema es el resultado del gran tamaño programa que SPSS Modeler Server envía a Analytic Server. El problema es menos probable que se produzca en Analytic Server 3.0 (o superior). Utilice los pasos siguientes para solucionar el problema:

  1. En la consola de Ambari, vaya al servicio Zookeeper en la pestaña **Configs**, añada la línea siguiente a la plantilla `zookeeper-env` en **Advanced zookeeper-env** y luego reinicie el servicio Zookeeper.

     ```
     export JVMFLAGS="-Xmx2048m -Djute.maxbuffer=2097152"
     ```

     

     Figura 5. Valores de plantilla zookeeper-env

  2. En la consola de Ambari, vaya a la pestaña **Configs** del servicio de Analytic Server, añada lo siguiente a **Advanced analytics-jvm-options** y reinicie a continuación el servicio Analytic Server.

     ```
     -Djute.maxbuffer=2097152
     ```

     

     Figura 6. Valores de Advanced analytics-jvm-options

  Nota: Si el problema persiste, aumente el valor `-Djute.maxbuffer` de `2097152` a `4194304` en ambos lugares.

Recomendaciones de la secuencia de IBM SPSS Modeler

Nota: La mayoría de las siguientes recomendaciones también se aplican a los datos pequeños.

Cree un prototipo de datos pequeños

Cuando está experimentando con una secuencia, suele añadir unos cuantos nodos, probar la secuencia hasta ese punto, quizás añadir un nodo para extraer algunas salida tabulada o gráfica, y luego continuar la construcción de la secuencia. Habitualmente no puede permitirse un segundo pase con los datos masivos cada vez que prueba la secuencia.

Cree un ejemplo de datos adecuado de los datos masivos le permite probar la secuencia con datos reales sin entrar en la penalización de tiempo necesaria que implica realizar un pase con los todos los datos completos. El ejemplo de datos debe contener suficientes datos para ejecutar satisfactoriamente la secuencia. Por ejemplo, si tiene previsto analizar transacciones en tiendas que se encuentran en Minnesota, el ejemplo de datos deben contener transacciones de tiendas en Minnesota.

Después del muestreo, puede:

- Crear una memoria caché del ejemplo de datos en el clúster donde residen los datos masivos, o

  Pros -Es simple y no requiere cambiar de nodos de origen

Contras - La memoria caché desaparece cuando finaliza la sesión

- Crear un nuevo origen de datos de Analytic Server que contenga el ejemplo de datos, o

  Pros - Origen de datos permanente

  Contras -Requiere editar/cambiar los nodos de origen

- Descargar los datos de ejemplo en el sistema local y cree un origen de datos local

  Pros -No consume recursos de clúster cuando se crean prototipos; el cliente SPSS Modeler es más eficaz que Analytic Server cuando se trabaja con un volumen de datos pequeño.

  Contras - Requiere cambiar nodos de origen

Cree nodos Tipo y Filtrar independientes de los nodos Origen

Cada nodo de origen de SPSS Modeler también tiene las funciones combinadas de los nodos Filtrar y Tipo. Esto resulta útil para racionalizar el lienzo, pero dificulta el cambio a tipos de nodo Origen diferentes. Además, oculta el hecho de que se están produciendo operaciones de Tipo y Filtrar.

Coloque los nodos Filtrar y Seleccionar lo más cerca posible del nodo Origen

Esto reduce el número de registros en operaciones descendentes.

Evite utilizar el nodo Ordenar siempre que sea posible

Analytic Server no da soporte a la optimización en los nodos que dependen de los datos que se ordenan (como el nodo Fusionar). Así pues, un nodo Ordenar en medio de la secuencia raramente tendrá una función útil. El nodo Ordenar sí tiene un valor cuando va seguido inmediatamente por un nodo Muestrear para obtener los registros Los N más (o Los N más).

Calcule solo los campos que se utilizarán

No debe calcular un campo y filtrarlo inmediatamente después.

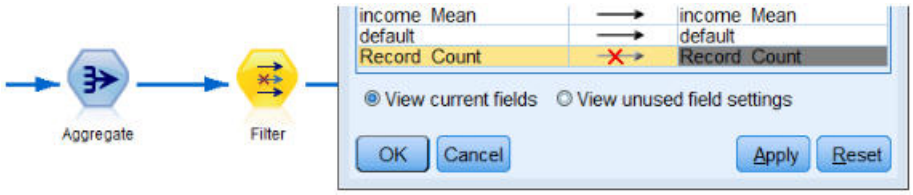

Figura 7. Opciones del campo de Modeler

Siempre que sea posible, sin hacer que las expresiones sean difícil de entender, evite crear muchos campos temporales. Por ejemplo, en lugar de definir el ejemplo siguiente:

```
now = datetime_now()
birthdate = datetime_date(bYear, bMonth, bDay)
age = date_years_difference(birthdate, now)
```

defina en su lugar el siguiente ejemplo:

```
age = date_years_difference(datetime_date(bYear, bMonth, bDay), datetime_now())
```

Esta manera de incluir elementos temporales en expresiones en línea puede aumentar el rendimiento cuando debe transformarse un gran número de campos.

Establezca el almacenamiento en el origen de datos

Las operaciones que cambian el tipo de almacenamiento de un campo (por ejemplo, serie a entero) en mitad de la secuencia pueden afectar negativamente el rendimiento global. Puede establecer el almacenamiento de campos, cuando defina orígenes de datos en la consola de Analytic Server, de forma que se evite la repetición de estas conversiones.

Utilice SPSS Modeler cuando trabaje con un volumen de datos pequeño

Manipule los datos masivos con Analytic Server y utilice a continuación SPSS Modeler para finalizar los cálculos en los datos de pequeño volumen.

Seleccione las propiedades de secuencia adecuadas relacionadas con Analytic Server

Configure las propiedades de secuencia pertinentes (**Herramientas** > **Opciones** > **Propiedades de secuencia** > **Analytic Server**) y decida si desea permitir que el proceso de datos abandone Analytic Server y continúe en SPSS Modeler (cuando un nodo no puede ejecutarse en Analytic Server).

De forma predeterminada, SPSS Modeler está configurado para informar de un error y detener su ejecución en esta situación. Puede omitir el error cambiando el valor de `Error` por `Aviso` y ajustando el límite de la cantidad de datos que pueden procesarse en SPSS Modeler. Por ejemplo, puede actualizar la velocidad de transferencia de datos del valor de registro predeterminado de 10000 (si es necesario). Tenga en cuenta que este límite también se aplica cuando se visualizan resultados que utilizan el nodo Tabla de SPSS Modeler. Si se supera el límite, SPSS Modeler informa de que la captación de datos ha superado el límite establecido en las propiedades de la secuencia.

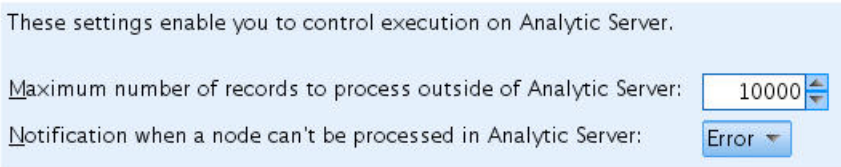

Figura 8. Valores de Analytic Server

Utilice nodos Origen de Analytic Server

Analytic Server puede conectarse a distintos orígenes de datos de base de datos pero SPSS Modeler requiere que todos los nodos Origen sean nodos Origen de Analytic Server (para que la secuencia entera se ejecute como un trabajo de Analytic Server). Para que toda la secuencia se ejecute en Analytic Server, el nodo de origen de base de datos debe cambiarse por un nodo Origen de Analytic Server y debe crearse un origen de base de datos de Analytic Server en la consola de Analytic Server.

Tenga en cuenta cómo se utilizan los nodos no soportados

Analytic Server no da soporte a todos los nodos (el nodo Transponer es un buen ejemplo). Para poder fusionar los resultados de una operación de transposición con el resto de la secuencia y hacer que se ejecute en Analytic Server, debe escribirse una subsecuencia que incluya un nodo Transponer en un origen de datos de Analytic Server que utiliza un nodo Exportar de Analytic Server. A continuación, puede adjuntar un nodo Origen de Analytic Server allí donde se ha interrumpido la secuencia para escribir en Analytic Server.

Nota: La operación de transposición es adecuada para una sola ejecución o para operaciones que se ejecutan en raras ocasiones, pero no debe utilizarse para operaciones de secuencias rutinarias.

Determine si una secuencia funcionará en Analytic Server antes de ejecutarla

Después de preparar una secuencia para ejecutar en Analytic Server, seleccione un nodo de terminal y utilizar la característica de vista previa de SPSS Modeler (el control **Vista previa de ejecución** en la barra de herramientas) para verificar que todos los nodos que participan en el nodo de terminal funcionarán en Analytic Server (sin ejecutar la secuencia). Los problemas se comunican en la ventana de mensajes.

Combine operaciones de fusión en cadena

Combine una serie de nodos Fusionar en un solo nodo cuando tienen las mismas claves y tipo de unión.

Combine subsecuencias idénticas

Intente combinar subsecuencias idénticas siempre que sea posible, especialmente si contienen operaciones costosas (por ejemplo, fusionar y ordenar). SPSS Modeler realiza estas operaciones una vez y utiliza la memoria caché para mejorar el rendimiento. En el ejemplo siguiente, las secuencias son idénticas hasta el nodo **newField**.

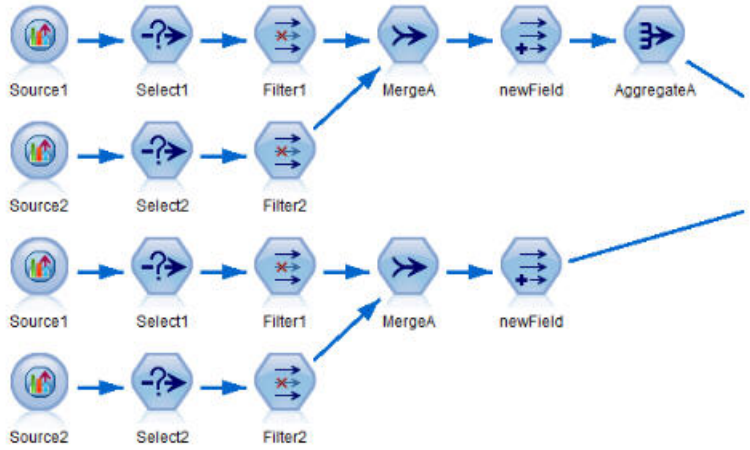

Figura 9. Ejemplo de secuencia

Es más eficaz (y más fácil de mantener) si la subsecuencia se estructura en su lugar de la siguiente manera:

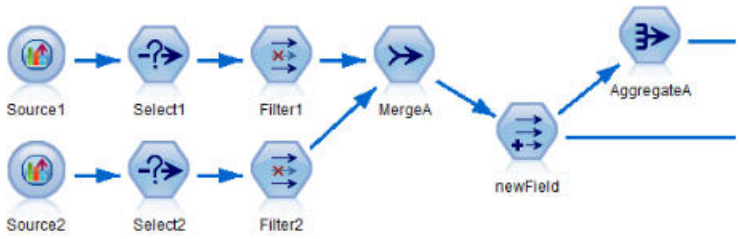

Figura 10. Ejemplo de secuencia

Eliminar nodos Tipo adicionales

Figura 11. Ejemplo de secuencia

Evite nodos Tipo innecesarios cuando se ejecuten para Analytic Server. La operación `Leer valores` inicia un trabajo MapReduce. Normalmente esto se amortiza en una sola vez, a no ser que borre los valores del nodo Tipo.

Documente a fondo cada secuencia

El ejemplo siguiente muestra una secuencia compleja que contiene varias subsecuencias.

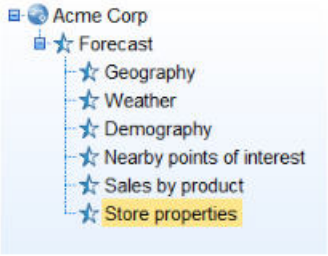

Figura 12. Ejemplo de subsecuencia

En estos casos, es importante denominar adecuadamente los supernodos y documentar la secuencia (de la misma manera que documentaría el código). Un comentario claro puede proporcionar información de gran valor a otros analistas que leen o mantienen la secuencia. Por ejemplo:

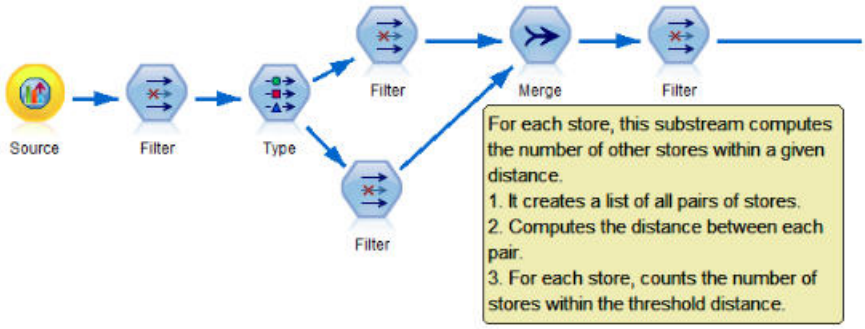

Figura 13. Ejemplo de secuencia con comentarios

Cuando desarrolle secuencias, utilice las memorias caché de SPSS Modeler para almacenar con rapidez los resultados intermedios

En las secuencias que se ejecutan para Analytic Server, el almacenamiento en memoria caché funciona almacenando los datos de una parte específica de la secuencia en archivos temporales en HDFS (en lugar de almacenarlos en el servidor SPSS Modeler). Las memorias caché funcionan bien con datos masivos y pueden utilizarse con seguridad en las secuencias que se ejecutan en Analytic Server.

Capítulo 6. Resolución de problemas

Analytic Server proporciona varias herramientas útiles para la determinación de problemas.

Registro

Analytic Server crea archivos de rastreo y archivos de registro de cliente que son de utilidad a la hora de diagnosticar un problema. Con la instalación predeterminada de Liberty, puede encontrar los archivos de registro en el directorio `{RAÍZ_AS}/ae_wlpserver/usr/servers/aeserver/logs`.

La configuración de registro predeterminada produce dos archivos de registro que se renuevan a diario.

`as.log`
Este archivo contiene el resumen de alto nivel de los mensajes informativos de error y aviso. Cuando se produzca un error de servidor que no pueda resolverse a partir del mensaje de error mostrado en la interfaz de usuario, empiece por consultar este archivo.

`as_trace.log`
Este archivo contiene todas las entradas de `ae.log`, e información adicional destinada principalmente al soporte y desarrollo de IBM a efectos de depuración.

Analytic Server utiliza por debajo Apache LOG4J como recurso de registro. Con log4j, el registro puede ajustarse de forma dinámica editando el archivo de configuración `{AS_SERVER_ROOT}/configuration/log4j.xml`. Es posible que el soporte de IBM le pida esto para ayudarle a diagnosticar un problema, o puede que le interese modificarlo para limitar el número de archivos de registro que se mantienen. Los cambios en el archivo se detectan automáticamente en unos segundos, de modo que no es necesario reiniciar Analytic Server.

Para obtener información adicional relativa a log4j y al archivo de configuración, consulte la documentación en el sitio oficial de Apache http://logging.apache.org/log4j/.

Información sobre la versión

Puede determinarse la versión instalada de Analytic Server consultando la carpeta `{RAÍZ_AS}/properties/version`. Los siguientes archivos contienen información de versión.

`IBM_SPSS_Analytic_Server-*.swtag`
Contiene información de producto detallada.

`version.txt`
Versión y número de compilación del producto instalado.

Recopilador de registros

Cuando no se pueden resolver los problemas consultando directamente los archivos de registro, puede empaquetar todos los registros y enviarlos al soporte de IBM. Se proporciona un programa de utilidad para simplificar la recopilación de todos los datos necesarios.

Utilizando un shell de mandatos, ejecute los siguientes mandatos:

```
cd  {RAÍZ_AS}/bin
   run  >sh ./logcollector.sh
```

Estos mandatos crean un archivo comprimido bajo `{RAÍZ_AS}/bin`. El archivo comprimido contiene todos los archivos de registro y la información de versión del producto.

Problemas comunes

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

Ejecución de secuencias

Los trabajos de R convierten palabras no inglesas a Unicode

En los clústeres de Cloudera, si la codificación del sistema de los servidores Hadoop no es UTF-8, R convierte las palabras que no sean inglesas a Unicode.

1. Vaya a la pestaña de configuración de YARN en la consola de Cloudera Manager.
2. Añada los valores siguientes en el campo "NodeManager Environment Advanced Configuration Snippet (Safety Valve)".

```
LC_ALL=""
LANG=en_US.utf8
```

La ejecución de los trabajos PySpark ha fallado

Asegúrese de que el servicio Spark se haya desplegado en todos los nodos de Analytic Server y en todos los gestores de nodos.

La ejecución de los trabajos PySpark ha fallado en entornos habilitados para Kerberos

Debe ejecutar el mandato `kinit` y, a continuación, reiniciar Analytic Server antes de que las pruebas de PySpark se ejecuten satisfactoriamente. Por ejemplo:

HDP Kerberos

```
cd /etc/security/keytabs/
sudo -u as_user kinit -k -t as_user.headless.keytab as_user/lyrh1.fyre.ibm.com@IBM.COM
```

CDH Kerberos

```
cd /run/cloudera-scm-agent/process/387-analyticserver-ANALYTIC_SERVER
sudo -u as_user kinit -k -t analyticserver.keytab as_user/cdh12-1.fyre.ibm.com@IBM.COM
```

El nodo XGBoost-AS no se ejecuta

Puede que encuentre el siguiente error al intentar ejecutar el nodo XGBoost-AS:

```
Error al ejecutar ASL: La ejecución ha fallado. Motivo: ml.dmlc.xgboost4j.java.XGBoostError:
ha fallado el entrenamiento de XGBoostModel
```

Configuración personalizada de Analytic Server

1. Aplique los siguientes valores de configuración personalizada de Analytic Server para resolver el problema.

```
spark.executor.memory=12g (o superior)
```

   - Los valores de Ambari están definidos en la sección **Custom analytics.cfg** de la configuración de Ambari.
   - Los valores de Cloudera se encuentran en la sección **Analytic Server Advanced Configuration Snippet (Safety Valve) for analyticserver-conf/config.properties** de Cloudera Manager.

2. Reinicie el servicio Analytic Server tras actualizar la configuración personalizada de Analytic Server.

Configuración de IBM SPSS Modeler

Habilite el valor IBM SPSS Modeler Nodo XGBoost-AS **Opciones de compilación** > **General** > **Utilizar memoria externa**.

Errores de memoria

Configuración de YARN tras errores de memoria de ejecución

Puede producirse el siguiente error cuando la memoria de ejecución necesaria supera el umbral máximo:

```
Caused by: com.spss.mapreduce.exceptions.JobException:
    java.lang.IllegalArgumentException: Required executor memory (1024+384  MB) is above the max
    threshold (1024 MB) of this cluster! Please  increase the value of
    'yarn.scheduler.maximum-allocation-mb'.
```

Los pasos siguientes proporcionan los valores de la configuración de YARN necesarios para resolver el problema.

Para Ambari

1. En la interfaz de usuario de Ambari, vaya a **YARN** > **Configs** > **Settings**.
2. Aumente el **nodo de memoria (la memoria asignada para todos los contenedores de YARN)** a 8192MB.
3. Aumente los valores de contenedor:
   - **Minimum Container Size (Memory)** a 682MB
   - **Maximum Container Size (Memory)** a 8192MB
4. Aumente el valor de **Maximum Container Size (VCores)** a 3.
5. Reinicie YARN, Spark y el servicio Analytic Server.

Para Cloudera

1. Aumente `yarn.nodemanager.resource.memory-mb` a 8GB
   - En la interfaz de usuario de Cloudera Manager, vaya a **YARN service** > **Configurations** > **Search Container Memory** y aumente el valor a 8GB.
2. En la interfaz de usuario de Cloudera Manager, vaya a **YARN service** > **Quick Links** y seleccione **Dynamic Resource Pools**.
3. Bajo **Configuration**, pulse **edit** para cada una de las agrupaciones disponibles y bajo **YARN** establezca el valor **Max Running Apps** en 4.
4. Reinicie YARN, Spark y el servicio Analytic Server.

Memoria de executor necesaria

Puede producirse el siguiente error cuando la memoria de ejecución necesaria supera el valor actual.

```
Ejecutor de AS eliminado:SparkListenerExecutorRemoved(1557300399468,716,Contenedor terminado
por YARN por superar los límites de memoria.
2.0 GB de 2 GB de memoria física utilizados. Considere la posibilidad de aumentar
spark.yarn.executor.memoryOverhead.)
```

1. Aplique los siguientes valores de configuración personalizada de Analytic Server para resolver el problema.

   ```
   spark.executor.memory=4g (o superior)
   ```

   - Los valores de Ambari están definidos en la sección **Custom analytics.cfg** de la configuración de Ambari.
   - Los valores de Cloudera se encuentran en la sección **Analytic Server Advanced Configuration Snippet (Safety Valve) for analyticserver-conf/config.properties** de Cloudera Manager.
2. Reinicie el servicio Analytic Server tras actualizar la configuración personalizada de Analytic Server.

Hadoop con Apache Spark 2.2

- La mayoría de los trabajos `forcespark` y `forcehadoop` fallan cuando Hadoop y Apache Spark 2.2 existen en el mismo entorno. El error aparece en el registro de aplicaciones de YARN como: `java.lang.NoClassDefFoundError: org/apache/hadoop/fs/FSDataInputStream`.

  El problema se puede resolver editando manualmente el archivo `/etc/spark2/conf/spark-defaults.conf` como se indica a continuación:

  ```
  #spark.hadoop.mapreduce.application.classpath=
  #spark.hadoop.yarn.application.classpath=
  ```

- Cuando hay instaladas dos versiones de JDK en el mismo sistema, Cloudera utiliza JDK 1.7 mientras que Spark 2.2 utiliza JDK 1.8. La ejecución de los trabajos `forcespark` o `forcehadoop` con Apache Spark 2.x puede dar lugar a que fallen todos los trabajos con el siguiente mensaje de error:

```
La ejecución ha fallado. Motivo: org/apache/spark/api/java/function/PairFunction :
major.minor version 52.0 no soportado
```

Para Cloudera, añada la línea siguiente en la sección **Analytic Server Advanced Configuration Snippet (Safety Valve) for server.env** de Cloudera Manager:

```
JAVA_HOME=/usr/java/jdk1.8.0_152
```

Otorgar la autorización admin a los usuarios de UDF de Apache Hive

Puede que aparezca el error `Función no válida` una vez registrada la UDF de Apache Hive de Analytic Server. De forma predeterminada, existen dos roles de Hive (`admin` y `public`). Los usuarios de Hive pertenecen al rol `public`. La UDF de Hive requiere que los usuarios registrados tengan el privilegio `admin` (la seguridad de Hive está habilitada).

Para otorgar la autorización `admin` a los usuarios de UDF de Hive:

1. Inicie la sesión en Beeline como Hive:

   ```
   !connect jdbc:hive2://localhost:10000/default;principal=hive/cdh51501.fyre.ibm.com@IBM.COM
   ```

2. Ejecute el mandato siguiente en Beeline:

   ```
   grant admin to user hive WITH ADMIN OPTION;
   ```

   Nota: Otros mandatos SQL útiles son:

   Mostrar qué roles ya están asignados al usuario `hive`

   ```
   show role grant user hive;
   ```

   Mostrar qué usuarios están asignados al rol `public`

   ```
   show principals public;
   ```

3. Reinicie Hive y vuelva a registrar la UDF de Hive de Analytic Server.

   ```
   sudo -u hive kinit -k -t hive.keytab hive/cdh51501.fyre.ibm.com@IBM.COM
   sudo -u hive hive -f /opt/cloudera/parcels/AnalyticServer/bin/udfUnregister.sql
   sudo -u hive hive -f /opt/cloudera/parcels/AnalyticServer/bin/udfRegister.sql
   ```

Error de la base de datos de Hive

Es posible que reciba el siguiente error al escribir en la base de datos de Hive:

```
(AEQAE4805E) La ejecución ha fallado. Razón: com.google.common.io.Closeables.closeQuietly(Ljava/io/Closeable;)
```

La causa del error es la existencia de varias versiones del archivo `guava-*.jar` en el clúster de Hadoop. El error puede resolverse realizando los pasos siguientes (el ejemplo utiliza HDP 3.1):

1. Abra la consola de Ambari y detenga el servicio Analytic Server.
2. Copie `/usr/hdp/3.1.0.0-78/spark2/jars/guava-14.0.1.jar` en `{RAÍZ_AS}/ae_wlpserver/usr/servers/aeserver/apps/AE_BOOT.war/WEB-INF/lib`.
3. En la consola de Ambari, renueve el servicio Analytic Server e inicie de nuevo el servicio Analytic Server.

Mezclar orígenes de datos de Hive y de HCatalog

Analytic Server no permite mezclar orígenes de datos de Hive y de HCatalog en la misma secuencia de IBM SPSS Modeler.

Ajuste del rendimiento

En esta sección se describen maneras de optimizar el rendimiento del sistema.

Analytic Server es un componente de la infraestructura de Ambari que utiliza otros componentes como, por ejemplo, HDFS, YARN y Spark. Las técnicas comunes de ajuste de rendimiento para Hadoop, HDFS y Spark se aplican a las cargas de trabajo de Analytic Server. Cada carga de trabajo de Analytic Server es diferente, por lo que deberá hacer pruebas de ajustes en función de su carga de trabajo de despliegues específica. Las propiedades y los consejos de ajustes siguientes son cambios clave que han impactado en los resultados de las pruebas de escalado y de benchmarking de Analytic Server.

Cuando se ejecute el primer trabajo en Analytic Server, el servidor iniciará una aplicación Spark persistente que estará activa hasta que se concluya Analytic Server. La aplicación Spark persistente asignará y mantendrá todos los recursos de clúster que se le hayan asignado durante la ejecución de Analytic Server, aunque haya algún trabajo de Analytic Server que no se esté ejecutando activamente. Deberá prestarse atención a la cantidad de recursos asignados a la aplicación de Spark de Analytic Server. Si todos los recursos de clúster se asignar a la aplicación Spark de Analytic Server, es posible que los demás trabajo se retrasen, o no se ejecuten. Estos trabajos pueden colocarse en la cola, a la espera de que haya suficientes recursos libres, y dichos recursos los consumirá la aplicación Spark de Analytic Server.

Si se han configurado y desplegado varios servicios de Analytic Server, cada instancia de servicio podría asignar, potencialmente, su propia aplicación Spark persistente. Por ejemplo, si se han desplegado dos servicios de Analytic Server para dar soporte a la migración tras error de alta disponibilidad, es posible que vea dos aplicaciones Spark persistentes activas, cada una de ellas asignando recursos de clúster.

Una complejidad adicional es que, en ciertas situaciones, Analytic Server puede comenzar un trabajo de MapReduce que necesitará recursos de clúster. Estos trabajos de MapReduce necesitarán recursos que no estén asignados a la aplicación Spark. Los componentes específicos que requieren trabajos de MapReduce son creaciones del modelo PSM.

Las propiedades siguientes pueden configurarse para que asignen recursos a la aplicación Spark. Si se establecen en el archivo spark-defaults.conf de la instalación de Spark, se asignarán a todos los trabajos de Spark que se ejecuten en el entorno. Si se establecen en la configuración de Analytic Server como propiedades personalizadas, en la sección “Custom analytic.cfg”, se asignarán solamente a la aplicación Spark de Analytic Server.

spark.executor.memory

Cantidad de memoria que debe utilizarse por cada proceso executor.

spark.executor.instances

El número de procesos executor que deben iniciarse.

spark.executor.cores

El número de hebras de trabajo de executor que deben utilizarse por cada proceso executor. Este valor debe oscilar entre 1 y 5.

Un ejemplo de configuración de las tres propiedades clave de Spark. Hay 10 nodos de datos en un clúster de HDFS y cada nodo de datos tiene 24 núcleos lógicos y 48 GB de memoria, y sólo ejecuta procesos HDFS. He aquí un modo de configurar las propiedades de este entorno, presuponiendo que sólo se ejecutan trabajos de Analytic Server en este entorno y se desea la asignación máxima a una única aplicación Spark de Analytic Server.

- Set spark.executor.instances=20. Este valor tratará de ejecutar 2 procesos executor de Spark por cada nodo de datos.
- Set spark.executor.memory=22G. Este valor establecerá el tamaño de almacenamiento dinámico máximo para cada proceso executor de Spark en 22 GB, asignando 44 GB en cada nodo de datos. Las demás JVM y el sistema operativo necesitan memoria adicional.
- Set spark.executor.cores=5. Este valor proporcionará 5 Para los sistemas para cada executor de Spark, para un total de 10 hebras de trabajo por nodo de datos.

Supervisar los trabajos en ejecución a través de la interfaz de usuario de Spark

Si ve la opción Spill to disk, esto podría afectar al rendimiento. He aquí algunas posibles soluciones:

- Aumente al memoria y asígnela a los procesos executor de Spark a través de **spark.executor.memory**.
- Reduzca el número de **spark.executor.cores**. Esto reducirá el número de hebras de trabajo concurrentes que asignan memoria, pero también reducirá la cantidad de paralelismo para los trabajos.
- Cambie las propiedades de la memoria de Spark. Porcentaje de asignación de **spark.shuffle.memoryFraction** y **spark.storage.memoryFraction** del almacenamiento dinámico de executor de Spark para Spark.

Asegúrese de que el nodo de nombres tenga memoria suficiente

Si el número de bloques en HDFS es grande y sigue creciendo, asegúrese de que el almacenamiento dinámico del nodo de nombres aumente para ajustarse a este crecimiento. Ésta es una recomendación de ajuste de HDFS común.

Altere la cantidad de memoria utilizada para la colocación en memoria caché

De forma predeterminada, **spark.storage.memoryFraction** tiene un valor de 0,6. Este valor puede aumentarse hasta 0,8 en el caso de que el tamaño de bloque de HDFS de los datos sea de 64 MB. Si el tamaño de bloque de HDFS de los datos de entrada es mayor de 64 MB, este valor sólo puede aumentarse si la memoria asignada por tarea es mayor que 2 GB.

Ajuste del rendimiento de la puntuación de modelos

Puede mejorar el rendimiento de los trabajos de puntuación de modelos en conjuntos de datos masivos con el motor de Apache Spark, utilizando los pasos siguientes. Tenga en cuenta que dichos pasos no impactarán en el funcionamiento de los servicios del clúster que no sean de Analytic Server.

1. Compruebe si ya se ha instalado libtcmalloc_minimal.so{/version} en cada nodo del clúster.

   ```
   whereis libtcmalloc_minimal.so.*
   ```

2. Si no se ha instalado libtcmalloc_minimal.so, instale el paquete específico del sistema operativo que contenga la biblioteca libtcmalloc_minimal en cada nodo del clúster, o compile e instale libtcmalloc_minimal manualmente. Por ejemplo:

   Ubuntu:

   ```
   sudo apt-get install libgoogle-perftools-dev
   ```

   Red Hat Enterprise Linux 6.x (x64):

   a. Instale el repositorio EPEL para Red Hat (si aún no se ha instalado)

      ```
      wget http://dl.fedoraproject.org/pub/epel/6/x86_64/epel-release-6-8.noarch.rpm
      sudo rpm -Uvh epel-release-6*.rpm
      ```

   b. `sudo yum install gperftools-libs.x86_64`

   Compilación manual:

   a. Descargue el archivo gperftools-2.4.tar.gz desde el enlace https://github.com/gperftools/gperftools/releases
   b. `tar zxvf gperftools-2.4.tar.gz`
   c. `cd gperftools-2.4`
   d. `./configure --disable-cpu-profiler --disable-heap-profiler --disable-heap-checker --disable-debugalloc --enable-minimal`
   e. `make`
   f. `sudo make install`

3. Observe que una de las ubicaciones del archivo de biblioteca instalado libtcmalloc_minimal.so{.version}, se devuelve al ejecutar el mandato siguiente en uno o más de los nodos.

```
whereis libtcmalloc_minimal.so.*
```

   Si el clúster tiene nodos en los que se ejecute una combinación de sistemas operativos, puede haber varias ubicaciones para este archivo.

4. En la consola de Ambari, vaya a la configuración de Analytic Server y, en la sección Custom analytics.cfg, configure la clave spark.executorEnv.LD_PRELOAD utilizando la ubicación de la biblioteca como valor. Tras efectuar este cambio, reinicie el servicio de Analytic Server. Por ejemplo, si la biblioteca está instalada en /usr/lib64/libtcmalloc_minimal.so.4, le configuración sería:

```
spark.executorEnv.LD_PRELOAD=/usr/lib64/libtcmalloc_minimal.so.4
```

   Si se necesitan varias ubicaciones, utilice un espacio para separarlas, tal como se muestra en el ejemplo siguiente.

```
spark.executorEnv.LD_PRELOAD=/usr/lib64/libtcmalloc_minimal.so.4 /usr/lib/
libtcmalloc_minimal.so
```

   Si algún nodo no tiene instalada la biblioteca `libtcmalloc_minimal.so` en una de las ubicaciones configuradas, esto no provocará ningún error, pero el rendimiento de la puntuación de modelos puede ser más lenta es dicho nodo.

Unión Spark de lado de correlación

La implementación de unión Spark de Analytic Server no da soporte a la funcionalidad de unión de lado de correlación (una unión Spark es principalmente reducir un lado). La implementación no se beneficia de las uniones de lado de correlación de uniones para optimizar las uniones cuando una entrada es pequeña. El resultado de no aprovechar la unión de lado de correlación es un trabajo Spark que utiliza recursos de manera extremadamente intensiva y que finalmente resulta fallido.

Para optimizar las uniones cuando se ejecutan uniones Spark de lado de correlación de Analytic Server (o trabajos Spark nativos basados en el tamaño de RDD más pequeño), puede añadir la propiedad `spark.msj.maxBroadcast` al archivo `analytics.cfg` (`SPSS Analytic Server/Configs/Custom analytics.cfg`) o a `analytics-meta`.

Avisos

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

LICENCIA DE DERECHOS DE AUTOR:

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

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

Marcas registradas

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

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

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

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

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

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

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

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

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

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