Configuración de WPS para Hadoop

Versión: 4.3.3

www.worldprogramming.com

Tabla de contenidos

Introducción

¿Qué es Hadoop?

Hadoop es un marco de trabajo del software de código abierto escalable y tolerante a errores para el almacenamiento distribuido y el procesamiento distribuido de conjuntos de datos muy grandes en clústeres de ordenadores. Se distribuye bajo la licencia Apache.

Para aquellos de ustedes que son novatos en Hadoop, se le recomienda que se refieran inicialmente a *Fundamentos de Hadoop* (pág. 7).

Ventajas ofrecidas por la integración de Hadoop en WPS

- Con la integración de Hadoop, WPS amplía sus capacidades de integración de datos a través de docenas de motores de base de datos.
- El uso compartido de datos entre WPS y HDFS (Hadoop Distributed File System) ofrece interoperabilidad a nivel de datos entre ambos entornos. Aunque no es transparente, es directo: Los datos de Hadoop se pueden importar en WPS para el análisis (estructurado) y, si se desea, posteriormente se envían de vuelta a HDFS.
- Los usuarios de WPS pueden invocar la funcionalidad de Hadoop desde el entorno familiar de la interfaz de usuario de WPS Workbench
- Los usuarios pueden crear y editar nuevas operaciones de Hadoop utilizando un lenguaje similar al SQL, ya que no tienen que conocer Java.

Alcance de este documento

Este documento ofrece una información general de la implementación de WPS y Hadoop, y también abarca la configuración de Kerberos cuando se aplica.

Resumen de integración de WPS/Hadoop

Las siguientes integraciones actualmente implementadas usan las extensiones `filename`, `libname` y `PROC HADOOP` de WPS:

- Conectarse a Hive mediante SQL estándar
- Conectarse a Impala mediante SQL estándar
- Conectarse a Hive mediante SQL de acceso directo
- Conectarse a Impala mediante SQL de acceso directo
- Dicte comandos HDFS y ejecutar scripts de Pig

La integración de Hadoop de WPS se ha certificado contra Cloudera 5 y probada en comparación con otras distribuciones de Hadoop que permanecen cerca del estándar de Apache. Varias muestras de códigos relacionadas con la integración (pág. 18) se dan al final del documento.

Requisitos previos

Hadoop es una pila de tecnología de varias partes compleja. Antes de integrarse con WPS, se necesita instalar y configurar correctamente. Los siguientes pasos preparatorios se deben realizar y revisar:

1. Obtenga el conjunto correcto de archivos `.jar` que corresponden a su instalación de Hadoop.

   Nota:

   Cuando utilice Apache Hive como parte de su instalación de Hadoop con WPS, debe utilizar Apache Hive versión 0.12 o superior.

2. Configure los archivos XML de configuración de acuerdo con su entorno de clúster específico (direcciones IP, puertos y demás).
3. Establezca si su distribución de Hadoop incluye o exige el soporte para Kerberos. Si es así, confirme que la autenticación Kerberos en su servidor funciona, que la entidad se ha configurado correctamente y demás. Sin tener en cuenta si se está utilizando o no Kerberos, complete los pasos restantes de los requisitos previos.
4. Establezca que el clúster funciona correctamente, tal vez consultando con el administrador del clúster que debería tener acceso a los paneles administrativos.
5. Una vez que se haya establecido que el clúster funciona correctamente, establezca que las tareas relacionadas con Hadoop se pueden presentar independientemente de WPS.

Kerberos

Establecer la identidad con una autenticación fuerte es la base para el acceso seguro en Hadoop, con los usuarios que se necesita poder identificar para poder acceder a los recursos, y los recursos del clúster Hadoop que se necesitan autenticar individualmente para evitar que los sistemas maliciosos 'posen' potencialmente como parte del clúster para obtener el acceso a los datos. Para crear esta comunicación segura entre sus diversos componentes, Hadoop puede utilizar Kerberos, que es un mecanismo de autenticación de terceros, mediante el cual los usuarios y servicios que los usuarios desean acceder, confían en el servidor Kerberos para administrar la autenticación.

Nota:

Algunas distribuciones de Hadoop incluyen (o incluso exigir) el soporte para Kerberos. Los detalles de la configuración del servidor Kerberos a menudo varían según el tipo y la versión de distribución, y están fuera del alcance de este documento. Consulte la información de configuración específica de la distribución que se proporciona con su software Hadoop. Consulte *Configuración de Kerberos y Hadoop en el cliente* (pág. 17) para saber cómo configurar Kerberos y Hadoop en el lado del cliente.

Fundamentos de Hadoop

En los entornos analíticos tradicionales, los datos se introducen en un RDBMS mediante un proceso inicial ETL (Extract, Transform, Load). Los datos no estructurados se preparan y cargan en la base de datos, adquiriendo un esquema sobre la marcha. Una vez cargado, se vuelve susceptible a un montón de técnicas de análisis bien establecidas.

Sin embargo, para grandes cantidades de datos, este flujo de trabajo tiene algunos problemas:

1. Si el tiempo necesario para procesar los datos de un día alcanza a un punto en el que no puede completarlo económicamente antes del día siguiente, necesita otro enfoque. El ETL a gran escala ejerce una presión masiva sobre la infraestructura subyacente.
2. A medida que los datos envejecen, a menudo se archivan. Sin embargo, es muy costoso recuperar datos archivados en el volumen (de cinta, blu-ray y demás). Además, una vez que se haya archivado, ya no tiene acceso conveniente y económico a la misma.
3. El proceso ETL es un proceso de abstracción: los datos se agregan y normalizan y su forma original de alta fidelidad se pierde. Si posteriormente la empresa formula un nuevo tipo de pregunta sobre los datos, a menudo no es posible proporcionar una respuesta sin un ejercicio costoso que implica cambiar la lógica ETL, corregir el esquema de la base de datos y recargar.

Hadoop se ha diseñado para proporcionar:

- Escalabilidad en contraposición a informática y datos, eliminando el cuello de botella ETL.
- Economía mejorada para mantener vivos los datos, y en el almacenamiento principal, por más tiempo.
- La flexibilidad de volver y formular nuevas preguntas sobre los datos originales de alta fidelidad.

Comparación de RDBMS y Hadoop

Desde una perspectiva analítica, las principales diferencias entre un RDBMS y Hadoop son como se muestran a continuación.

Tabla 1. Las diferencias clave entre RDBMS y Hadoop

RDBMS	Hadoop
El esquema debe crearse antes de que se puedan cargar los datos.	Los datos se copian simplemente en el almacén de archivos y no se necesita ninguna transformación.
Debe llevarse a cabo una operación ETL explícita, transformando los datos en la estructura interna de la base de datos.	Un serializador/deserializador se aplica en tiempo de lectura para extraer las columnas requeridas.

RDBMS	Hadoop
Las nuevas columnas se deben agregar explícitamente antes de que se puedan cargar nuevos datos para dichas columnas.	Los nuevos datos pueden iniciar a fluir en cualquier momento y aparecerán retrospectivamente una vez que se haya actualizado el serializador/deserializador para analizarlo.

La orientación por esquema de las implementaciones RDBMS convencionales proporciona algunos beneficios clave que han llevado a su adopción generalizada:

- Optimizaciones, índices, particiones y demás, son posibles, permitiendo lecturas muy rápidas para ciertas operaciones, tales como uniones, uniones de varias tablas, etc.
- Un esquema común en toda la organización significa que los diferentes grupos de una empresa pueden comunicar entre sí utilizando un vocabulario común.

Por otro lado, las implementaciones RDBMS pierden cuando se trata de flexibilidad, la capacidad de expandir datos a la velocidad a la que se están evolucionando. Con Hadoop, la estructura sólo se impone a los datos en tiempo de lectura, a través de un serializador/deserializador, y por lo tanto, no hay ninguna fase ETL, los archivos simplemente se copian en el sistema. Fundamentalmente, Hadoop no es una base de datos convencional en el sentido normal, dadas sus propiedades ACID (Atomicidad, Consistencia, Aislamiento, Durabilidad), e incluso si lo fuera, probablemente sería demasiado lento para impulsar la mayoría de las aplicaciones interactivas.

Ambas tecnologías pueden aumentar mutuamente, ambas pueden tener un lugar en la organización informática, es simplemente una cuestión de elegir la herramienta adecuada para el trabajo correcto.

Tabla 2. RDBMS contra Hadoop: Casos de uso clave

Cuando usar RDBMS	Cuando usar Hadoop
OLAP interactivo: tiempo de respuesta de sub-segundos.	Cuando necesita administrar tanto datos estructurados como no estructurados.
Cuando necesita admitir transacciones ACID de varios pasos en datos basados en registros (p. ej. cajeros automáticos, etc.).	Cuando se requiere escalabilidad de almacenamiento y/o computación.
Cuando se requiere el cumplimiento del 100% de SQL.	Cuando tiene necesidades complejas de procesamiento de datos con volúmenes muy grandes de datos.

Arquitectura Hadoop

Dos conceptos clave se encuentran en el núcleo de Hadoop:

- El HDFS (Hadoop Distributed File System): un sistema de archivos basado en Java que proporciona el almacenamiento de datos escalable y confiable que abarca grandes clústeres de servidores de mercancía.
- MapReduce: un modelo de programación que simplifica la tarea de escribir programas que funcionan en un entorno de computación paralela.

Un clúster operativo de Hadoop tiene muchos otros subsistemas, pero HDFS y MapReduce son fundamentales para el modelo de procesamiento.

HDFS

HDFS es un sistema de archivos distribuido, escalable y portátil escrito en Java. HDFS almacena archivos grandes (normalmente en el rango de gigabytes a terabytes) en varias máquinas. Se logra confiabilidad mediante la replicación de los datos a través de varios hosts. De forma predeterminada, los bloques de datos están almacenados (replicados) en tres nodos, dos en el mismo bastidor y otro en un bastidor diferente (una sobrecarga de 3 veces en contraposición con el almacenamiento no replicado). Los nodos de datos pueden comunicarse entre sí para reequilibrar los datos, mover copias y mantener alta la replicación de datos.

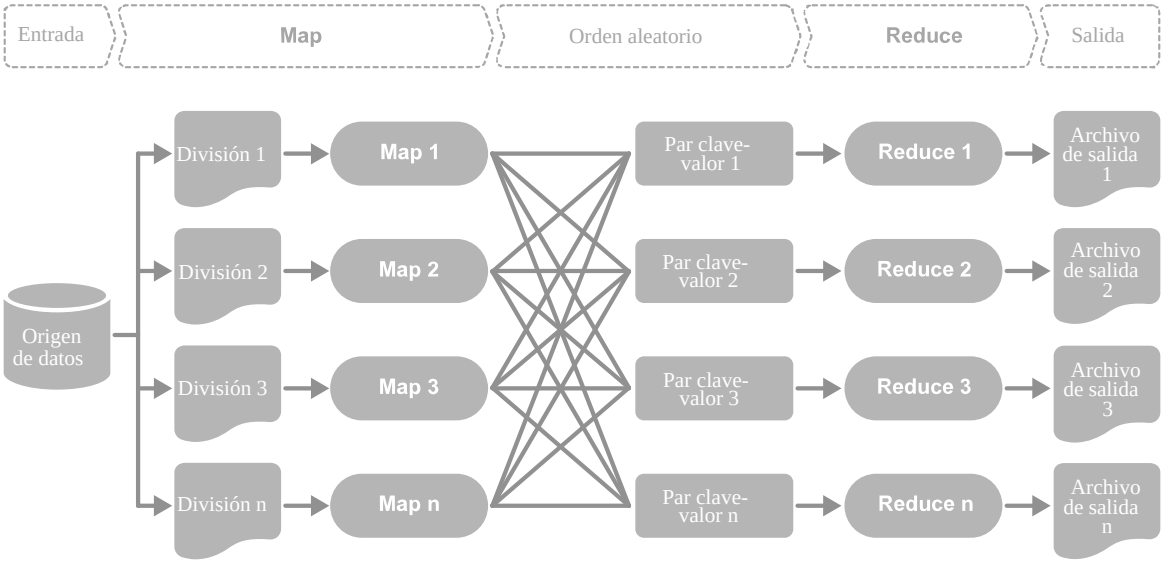
HDFS no es un sistema de archivos completamente compatible con Posix y está optimizado para el rendimiento. Ciertas operaciones de archivos atómicos son prohibidas o lentas. No puede, por ejemplo, insertar datos nuevos en el medio de un archivo, aunque puede anexarlo.

MapReduce

MapReduce es un marco de trabajo de programación que, si se sigue, quita la complejidad de la tarea de programación en entornos masivamente paralelos.

Un programador típicamente tiene que escribir dos funciones, una función Map y una función Reduce, y otros componentes en el marco de trabajo Hadoop se encargará de la tolerancia a errores, distribución, agregación, ordenación y demás. El ejemplo generalmente citado es el problema de producir un recuento de la frecuencia de palabras agregadas a través de un gran número de documentos. Se usan los siguientes pasos:

1. El sistema divide los datos de entrada entre varios nodos denominados asignadores. Aquí, el programador escribe una función que cuenta cada palabra en un archivo y cuántas veces ocurren. Esta es la función Map, cuya salida es un conjunto de pares clave-valor que incluyen una palabra y un recuento de palabras. Cada asignador hace esto a su propio conjunto de documentos de entrada, así que, en total, muchos asignadores producen muchos conjuntos de pares clave-valor para la siguiente etapa.
2. La fase aleatoria sucede, se aplica una función de hash coherente a los pares clave-valor y los datos de salida se redistribuyen a los reductores, de tal manera que todos los pares clave-valor con la misma clave van al mismo reductor.
3. El programador ha escrito una función Reduce que, en este caso, simplemente suma las repeticiones de palabras de los flujos entrantes de pares clave-valor, escribiendo los totales en un archivo de salida:



Este proceso aísla al programador de los problemas de escalabilidad a medida que crece el clúster. Parte del sistema Hadoop se ocupa de la ordenamiento y ejecución de recursos, esta parte es YARN si la versión de MapReduce es 2.0 o superior.

No hay ninguna garantía de que todo este proceso sea más rápido que cualquier otro tipo de sistema alternativo (aunque, en la práctica, es más rápido para ciertos tipos de conjuntos de problemas y grandes volúmenes de datos). El principal beneficio de este modelo de programación es la capacidad de explotar el funcionamiento aleatorio optimizado a menudo, mientras que sólo tiene que escribir las partes de Map y Reduce del programa.

El ecosistema de Hadoop

Hay varias formas de interactuar con un clúster de Hadoop.

Java MapReduce

Este es el método de acceso más flexible y de mejor rendimiento, aunque, dado que este es el lenguaje ensamblador de Hadoop, puede ser un ciclo de desarrollo implicado.

Streaming MapReduce

Esto permite el desarrollo en Hadoop en cualquier lenguaje de programación escogido, a costa de reducciones de rendimiento y flexibilidad de ligeras a modestas. Todavía depende del modelo MapReduce, pero amplía el conjunto de lenguajes de programación disponibles.

Crunch

Esta es una biblioteca para canalizaciones de MapReduce en Java, inspirado en FlumeJava de Google. Ofrece una API de Java para las tareas, tal como la unión y agregación de datos que son tediosas para implementar en MapReduce simple.

Pig Latin

Un lenguaje de alto nivel (a menudo denominado como 'Pig') que es adecuado para cargas de trabajo de flujo de datos por lotes. Con Pig, no hay ninguna necesidad de pensar en cuanto a MapReduce en absoluto. Abre el sistema a los programadores no Java y proporciona operaciones comunes, tales como unir, agrupar, filtrar y ordenar.

Hive

Un intérprete de SQL (no conforme) que incluye un metastore que puede asignar archivos a sus esquemas y serializadores/deserializadores asociados. Como Hive está basado en SQL, los controladores ODBC y JDBC permiten el acceso a herramientas estándar de inteligencia empresarial, tal como Excel.

Oozie

Un motor para el flujo de trabajo XML de PDL que le permite crear un flujo de trabajo de trabajos compuestos de cualquiera de los anteriores.

HBase

Apache HBase está dirigido al host de tablas muy grandes, millardos de filas y millones de columnas, encima de los clústeres de hardware básico. Inspirado en Bigtable de Google, HBase ofrece capacidades similares a las de Bigtable en la parte superior de Hadoop y HDFS.

Zookeeper

Apache Zookeeper es un esfuerzo para desarrollar y mantener un servidor de código abierto que permita una coordinación distribuida altamente confiable. Zookeeper es un servicio centralizado para mantener la información de configuración y la nomenclatura, y para proporcionar la sincronización distribuida y los servicios de grupo.

Implementación de WPS y Hadoop en Windows x64

Instalación de WPS en Windows x64

1. Antes de empezar a instalar WPS, asegúrese de que su copia de Windows disponga de las últimas actualizaciones y service packs.
2. Ambas las instalaciones de las estaciones de trabajo y servidores de Windows utilizan el mismo software WPS; el uso se controla por medio de una clave de licencia aplicada mediante el procedimiento `setinit`.
3. El archivo de instalación de WPS para Windows se puede descargar desde el sitio web de World Programming. Necesitará un nombre de usuario y una contraseña para acceder a la sección de descargas del sitio.
4. Una vez que se haya descargado el archivo de instalación (.msi), sólo haga doble clic en el archivo, lea y acepte el EULA y siga las instrucciones en pantalla.
5. Una vez que se haya instalado el software WPS, será necesario aplicar la clave de licencia. La clave de licencia se le ha enviado por correo electrónico una vez que haya comprado el software WPS. La forma más fácil de aplicar la clave de licencia es mediante la ejecución de WPS Workbench como usuario con acceso de administrador al sistema, y siguiendo las instrucciones.
6. Esto concluye la configuración de WPS.

Configuración de Hadoop en Windows x64

Instalación de Hadoop

Si aún no lo ha hecho, instale Hadoop, consultando, según se necesite, a la documentación proporcionada para su distribución en particular (por ejemplo, Cloudera). Una vez que haya instalado Hadoop, debe continuar con los detalles de configuración descritos a continuación.

Nota:

Siempre y cuando tengas una distribución que funcione de la manera estándar de Apache Hadoop, los detalles de la configuración deben aplicarse, aunque si tu distribución no sea Cloudera. Las distribuciones que desactivan o cambian las funciones estándares de Apache Hadoop no son compatibles.

Archivos de configuración

Todas las llamadas a Cloudera 5 Hadoop se realizan a través de Java y JNI. Los archivos `.jar` del cliente de Cloudera Hadoop se necesitarán obtener y descargar en la máquina local. Los siguientes archivos contienen URLs para varios servicios Hadoop y deben configurarse para que coincidan con la instalación actual de Hadoop:

- core-site.xml
- hdfs-site.xml
- mapred-site.xml

> **Nota:**
> Si está utilizando un cliente Windows contra un clúster de Linux, este último archivo debe establecer el parámetro de configuración `mapreduce.app-submission.cross-platform` como `true`.

Consulte la documentación de Hadoop para más información.

La variable de entorno CLASSPATH

La variable de entorno `CLASSPATH` se necesita configurar para apuntar a los archivos del cliente de Cloudera Java. Esto variará dependiendo de la configuración del cliente y de la máquina específica, pero un ejemplo ficticio podría parecerse a:

```
c:\Cloudera5\conf;c:\Cloudera5\*.jar
```

La variable de entorno HADOOP_HOME

En Windows, la variable de entorno `HADOOP_HOME` necesita configurarse para apuntar a los archivos del cliente de Cloudera Java. Para el ejemplo sobredicho, debe establecerse en: `C:\Cloudera5`.

Configuración de Kerberos en Windows x64

Si su distribución de Hadoop incluye o exige el soporte para Kerberos, continúe con *Configuración de Kerberos y Hadoop en el cliente* (pág. 17).

Implementación de WPS y Hadoop en Linux x64

Instalación de WPS en Linux x64

1. WPS está admitido en cualquier distribución de Linux que está compatible con LSB (Base estándar de Linux) 3.0 o superior. WPS está admitido en Linux que se ejecuta en x86, x86_64 e IBM System z, incluido IFL (Integrated Facility for Linux).
2. Si dispone de una distribución de Linux de 64 bits instalada, tiene la opción de utilizar WPS de 32 o 64 bits. Cabe señalar que algunas distribuciones de Linux de 64 bits sólo instalan las bibliotecas del sistema de 64 bits de manera predeterminada. El uso de WPS de 64 bits en estas distribuciones será listo para usar. Sin embargo, si opta por utilizar WPS de 32 bits, primero deberá instalar las bibliotecas del sistema de 32 bits. Por favor, consulte la documentación de la distribución de Linux para indicaciones sobre cómo lograrlo.
3. WPS para Linux está actualmente sólo disponible como un archivo de almacenamiento tar comprimido. Un instalador nativo, basado en la plataforma RPM, estará disponible en el futuro.
4. El almacenamiento de WPS para Linux se proporciona en formato gzipped tar (`.tar.gz`) y se puede descargar desde el sitio web de World Programming. Necesitará un nombre de usuario y una contraseña para acceder a la sección de descargas del sitio.
5. Para instalar WPS, extraiga los archivos desde el almacenamiento utilizando gunzip y tar de la manera siguiente. Elija una ubicación de instalación adecuada a la que tiene acceso de escritura, y cambie (`cd`) a ese directorio. El almacenamiento es completamente independiente y se puede desempaquetar en cualquier lugar. La ubicación de la instalación se puede encontrar en alguna parte que requiera el acceso a la raíz, tal como `/usr/local` si se instala para todos los usuarios, o se puede encontrar en su directorio principal.
6. Descomprima el archivo de instalación escribiendo: `tar -xzof <archivo-de-instalación-de-wps>.tar.gz` o: `gzip -cd <archivo-de-instalación-de-wps>.tar.gz | tar xvf -`
7. Necesitará una clave de licencia para ejecutar WPS. Se puede aplicar desde la interfaz gráfica de usuario o desde la línea de comandos, iniciando cualquier aplicación de la manera siguiente.
   a. Para iniciar la interfaz gráfica de usuario WPS Workbench, dicte el siguiente comando: `<wps-@versión-completa-producto-corta@-directorio-instalación>/eclipse/workbench`. El sistema abrirá un cuadro de diálogo donde puede importar la clave de licencia.
   b. Para iniciar WPS desde la línea de comandos, dicte el siguiente comando: `<wps-@versión-completa-producto-corta@-directorio-instalación>/bin/wps -stdio -setinit < <archivo-de-la-clave-de-wps>` Un mensaje confirmará que la licencia se ha aplicado correctamente.

8. Esto concluye la configuración de WPS.

Configuración de Hadoop en Linux x64

Instalación de Hadoop

Si aún no lo ha hecho, instale Hadoop, consultando, según se necesite, a la documentación proporcionada para su distribución en particular (por ejemplo, Cloudera). Una vez que haya instalado Hadoop, debe continuar con los detalles de configuración descritos a continuación.

Nota:

Siempre y cuando tengas una distribución que funcione de la manera estándar de Apache Hadoop, los detalles de la configuración deben aplicarse, aunque si tu distribución no sea Cloudera. Las distribuciones que desactivan o cambian las funciones estándares de Apache Hadoop no son compatibles.

Archivos de configuración

Todas las llamadas a Cloudera 5 Hadoop se realizan a través de Java y JNI. Los archivos `.jar` del cliente de Cloudera Hadoop se necesitarán obtener y descargar en la máquina local. Los siguientes archivos contienen URLs para varios servicios Hadoop y deben configurarse para que coincidan con la instalación actual de Hadoop:

- core-site.xml
- hdfs-site.xml
- mapred-site.xml

Consulte la documentación de Hadoop para más información.

La variable de entorno CLASSPATH

La variable de entorno `CLASSPATH` se necesita configurar para apuntar a los archivos del cliente de Cloudera Java. Para un ejemplo ficticio, las siguientes líneas podrían agregarse al perfil del usuario (tal como `.bash_profile`):

```
CLASSPATH=/opt/cloudera5/conf:/opt/cloudera5/*.jar
```

```
EXPORT CLASSPATH
```

Configuración de Kerberos en Linux x64

Si su distribución de Hadoop incluye o exige el soporte para Kerberos, continúe con *Configuración de Kerberos y Hadoop en el cliente* (pág. 17).

Configuración de Kerberos y Hadoop en el cliente

Tanto en Windows como en Linux, es posible que necesite ejecutar primero el comando `kinit` y escribir su contraseña en el mensaje. Esto puede ser la implementación de SO de `kinit` (en Linux) o el binario `kinit` en el directorio JRE dentro de WPS.

En Windows:

- Debe iniciar sesión como un usuario del directorio activo, no como un usuario de la máquina local
- Su usuario no puede ser un administrador local en la máquina
- Debe establecer una clave de registro para dejar que Windows permita el acceso de Java a la clave de sesión TGT:

```
HKEY_LOCAL_MACHINE\System\CurrentControlSet\Control\Lsa\Kerberos\Parameters
Value Name: allowtgtsessionkey
Value Type: REG_DWORD
Value: 0x01
```

- Los archivos de la Política de Jurisdicción de Fuerza Ilimitada de JCE (Java Cryptography Extension) se deben instalar en su JRE (es decir, el JRE dentro del directorio de instalación de WPS).

A continuación, tendrá que configurar varias entidades Kerberos en los archivos de configuración XML de Hadoop. Con Cloudera, están disponibles a través de Cloudera Manager. La lista de archivos de configuración incluye:

- `dfs.namenode.kerberos.principal`
- `dfs.namenode.kerberos.internal.spnego.principal`
- `dfs.datanode.kerberos.principal`
- `yarn.resourcemanager.principal`
- `yarn.resourcemanager.principal`

Nota:

La lista de arriba no es exhaustiva y a menudo puede ser específica de un sitio: las instrucciones `libname` requieren además que el parámetro `hive_principal` se establezca en `hive_principal` del clúster Kerberos.

Ejemplos de código relacionados con la integración

Conexión a Hive mediante SQL estándar

```
libname lib hadoop schema=default server="clouderademo" user=demo
password=demo;

proc sql;
 drop table lib.people;
run;

data people1;
 infile 'd:\testdata.csv' dlm=',' dsd;
 input id $ hair $ eyes $ sex $ age dob :date9. tob :time8.;
run;

proc print data=people1;
 format dob mmddyy8. tob time8.;
run;

data lib.people;
  set people1;
run;

data people2;
  set lib.people;
run;

proc contents data=people2;
run;

proc print data=people2;
 format dob mmddyy8. tob time8.;
run;

proc means data=lib.people;
 by hair;
  where hair = 'Black';
run;
```

Conexión a Impala mediante SQL estándar

```
libname lib hadoop schema=default server="clouderademo" user=demo
password=demo port=21050 hive_principal=nosasl;
```

```
proc sql;
 drop table lib.peopleimpala;
run;

data people1;
 infile 'd:\testdata.csv' dlm=',' dsd;
 input id $ hair $ eyes $ sex $ age dob :date9. tob :time8.;
run;

proc print data=people1;
 format dob mmddyy8. tob time8.;
run;

data lib.peopleimpala;
  set people1;
run;

data people2;
  set lib.peopleimpala;
run;

proc contents data=people2;
run;

proc print data=people2;
 format dob mmddyy8. tob time8.;
run;

proc means data=lib.peopleimpala;
 by hair;
  where hair = 'Black';
run;
```

Conexión a Hive mediante SQL de acceso directo

```
proc sql;
connect to hadoop as lib (schema=default server="clouderademo" user=demo
 password=demo);
    execute (create database if not exists mydb) by lib;
    execute (drop table if exists mydb.peopledata) by lib;
    execute (CREATE EXTERNAL TABLE mydb.peopledata(id STRING, hair STRING, eye
 STRING, sex STRING, age INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES
 TERMINATED BY '\n' STORED AS TEXTFILE LOCATION '/user/demo/test') by lib;
 select * from connection to lib (select * from mydb.peopledata);
  disconnect from lib;
quit;

/* options sastrace=,,,d; */
libname lib2 hadoop schema=mydb server="clouderademo" user=demo password=demo;
data mypeopledata;
  set lib2.peopledata;
run;

proc print data=mypeopledata;
run;
```

Conexión a Impala mediante SQL de acceso directo

```
proc sql;
  connect to hadoop as lib (schema=default server="clouderademo" user=demo
 password=demo port=21050 hive_principal=nosasl);
    execute (create database if not exists mydb) by lib;
    execute (drop table if exists mydb.peopledataimpala) by lib;
    execute (CREATE EXTERNAL TABLE mydb.peopledataimpala(id STRING, hair STRING, eye
 STRING, sex STRING, age INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES
 TERMINATED BY '\n' STORED AS TEXTFILE LOCATION '/user/demo/test') by lib;
 select * from connection to lib (select * from mydb.peopledataimpala);
  disconnect from lib;
quit;

libname lib2 hadoop schema=mydb server="clouderademo" user=demo password=demo
 port=21050 hive_principal=nosasl;
data mypeopledata;
  set lib2.peopledataimpala;
run;

proc print data=mypeopledata;
run;
```

Ejecución de comandos HDFS y scripts de Pig a través de WPS

Ejemplo de código de WPS

```
filename script 'd:\pig.txt';
proc hadoop options='d:\hadoop.xml' username = 'hdfs' verbose;
       hdfs delete='/user/demo/testdataout' recursive;
run;

proc hadoop options='d:\hadoop.xml' username = 'demo' verbose;
       pig code = script;
run;

proc hadoop options='d:\hadoop.xml' username = 'demo';
       hdfs copytolocal='/user/demo/testdataout/part-r-00000' out='d:\output.txt'
 overwrite;
run;

data output;
 infile "d:\output.txt" delimiter='09'x;
 input field1 field2 $;
run;

proc print data=output;
run;
```

Ejemplo de código de Pig

```
input_lines = LOAD '/user/demo/test/testdata.csv' AS (line:chararray);
-- Extract words from each line and put them into a pig bag
-- datatype, then flatten the bag to get one word on each row
words = FOREACH input_lines GENERATE FLATTEN(TOKENIZE(line)) AS word;

-- filter out any words that are just white spaces
filtered_words = FILTER words BY word MATCHES '\\w+';
```

```
-- create a group for each word
word_groups = GROUP filtered_words BY word;

-- count the entries in each group
word_count = FOREACH word_groups GENERATE COUNT(filtered_words) AS count, group AS
 word;

-- order the records by count
ordered_word_count = ORDER word_count BY count DESC;
STORE ordered_word_count INTO '/user/demo/testdataout';
```

Uso de WPS con Hadoop Streaming

El streaming de Hadoop es una utilidad que viene con la distribución de Hadoop. La utilidad le permite crear y ejecutar trabajos de MapReduce con cualquier ejecutable o script como un asignador y/o un reductor.

La sintaxis de esquema es de la manera siguiente:

```
$HADOOP_HOME/bin/hadoop jar $HADOOP_HOME/hadoop-streaming.jar \
    -input myInputDirs \
    -output myOutputDir \
    -mapper /bin/cat \
    -reducer /bin/wc
```

Los asignadores y reductores reciben su entrada y salida en `stdin` y `stdout`. La vista de datos está orientada a líneas y cada línea se procesa como un par clave-valor separado por el carácter de 'tabulación'.

Puede utilizar el streaming de Hadoop para emplear la potencia de WPS con el fin de distribuir programas escritos en el lenguaje de SAS a través de muchos ordenadores en un clúster de Hadoop, como en el ejemplo del trabajo de MapReduce puesto a continuación.

Nota:

Debido a la amplia distribución de los programas, cualquier ejemplo necesariamente usa un enfoque no tradicional para el lenguaje de SAS, en que cada asignador y reductor sólo detecta un subconjunto limitado de los datos.

Antes de continuar, asegúrese de estar familiarizado con los conceptos de HDFS y MapReduce en *Arquitectura Hadoop* (pág. 8).

El siguiente ejemplo muestra la creación y ejecución de un trabajo de MapReduce para producir recuentos de palabras que aparecen en los archivos de texto en el directorio que se proporciona como entrada al trabajo MapReduce. Cada uno de los recuentos de una palabra se procesan como el par clave-valor `<palabra><tabulación>1`.

1. Asegúrese de que el directorio `input` se ha configurado en el HDFS, por ejemplo utilizando `hadoop fs -mkdir /user/rw/input`, y que los archivos de texto que contienen las palabras que se van a contar se han agregado al directorio. Cada clúster puede detectar este directorio.
2. Asegúrese de que WPS se haya instalado en la misma ubicación en cada nodo del clúster, así que cualquiera de los asignadores y reductores lo pueda llamar.

3. Cree un programa asignador se llama **map.sas**:

```
options nonotes;

data map;
 infile stdin firstobs=1 lrecl=32767 encoding='utf8' missover dsd;
 informat line $32767.;
 input line;
 do i=1 by 1 while(scan(line, i, ' ') ^= '');
  key = scan(line, i,' ');
  value = 1;
  drop i line;
  output;
 end;
run;

proc export data=map outfile=stdout dbms=tab replace;
 putnames=no;
run;
```

4. Cree uno script denominado **map.sh** para llamar **map.sas**:

```
#!/bin/bash
/opt/wps/bin/wps /home/rw/map.sas
```

5. Cree un programa reductor denominado **reduce.sas**:

```
options nonotes;

data reduce;
 infile stdin delimiter='09'x firstobs=1 lrecl=32767 missover dsd;
 informat key $45.;
 informat value best12.;
 input key value;
run;

proc sql;
 create table result as select key as word, sum(value) as total from reduce
 group by key order by total desc;
quit;

proc export data=result outfile=stdout dbms=tab replace;
 putnames=no;
run;
```

6. Cree uno script denominado **reduce.sh** para llamar **reduce.sas**:

```
#!/bin/bash
/opt/wps/bin/wps /home/rw/reduce.sas
```

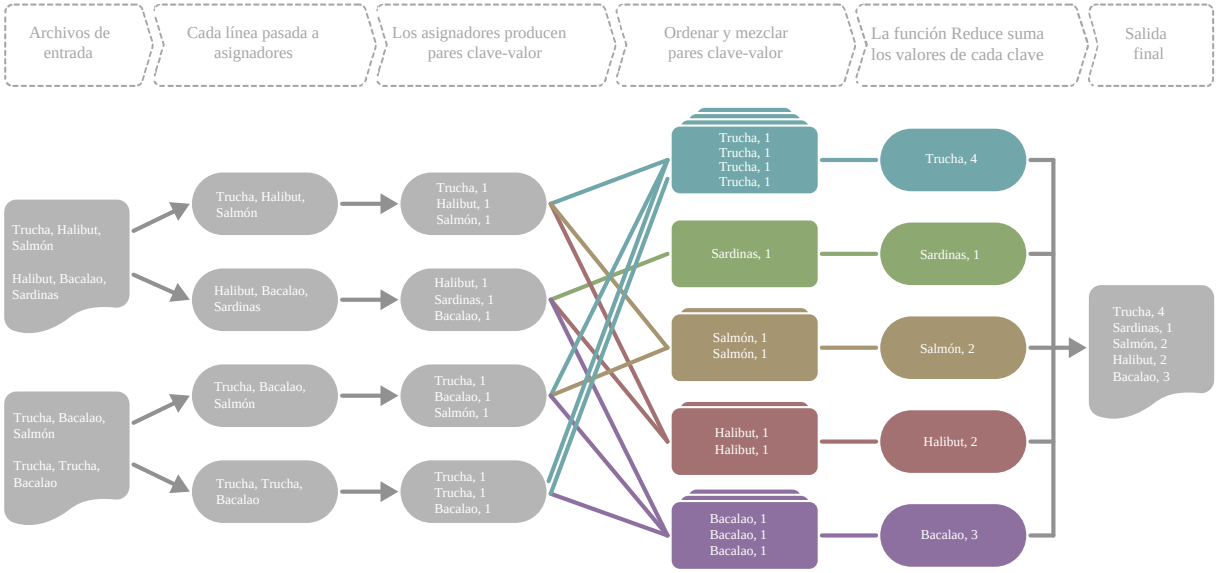
7. Asegúrese de que **map.sh**, **map.sas**, **reduce.sh** y **reduce.sas** se copian en la misma ubicación de cada nodo del clúster, así que los asignadores y reductores puedan ejecutarse cuando sea necesario.

8. Asegúrese de que la variable de entorno `CLASSPATH` se ha configurado en el equipo cliente para su sistema operativo, de acuerdo con *Configuración de Hadoop en Windows x64* (pág. 12) o *Configuración de Hadoop en Linux x64* (pág. 15).

9. Ejecute la siguiente línea de comandos desde un equipo cliente con un cliente Hadoop instalado, ajustando los números de versión según corresponda:

```
hadoop jar hadoop-streaming-2.5.0-cdh5.3.2.jar -input input -output output -
mapper "/home/rw/map.sh" -reducer "/home/rw/reduce.sh"
```

La ejecución del comando tiene el efecto de iniciar el trabajo MapReduce en el clúster específico. Cada instancia de un asignador (donde se trata del script **map.sh** en un nodo específico que invoca **map.sas**) produce un conjunto de pares clave-valor que cada uno consta de una palabra y un recuento de 1. La fase aleatoria entonces sucede con los pares clave-valor con la misma clave que va al mismo reductor. Cada instancia de un reductor (donde este es el script **reduce.sh** en un nodo específico que invoca **reduce.sas**) suma las repeticiones de las palabras para su clave particular en un archivo de salida. La salida resultante es una serie de palabras y recuentos asociados. Se puede ilustrar de la manera siguiente:



Nota:

La salida final puede generar la división en más de un archivo dentro del directorio de salida, dependiendo de la configuración del clúster.

Referencia

Las definiciones de los diagramas de sintaxis de ferrocarril son anotaciones que ayudan a explicar la sintaxis de los lenguajes de programación, y se usan en esta guía para describir la sintaxis del lenguaje.

Cómo leer los diagramas de sintaxis del ferrocarril

Los diagramas del ferrocarril son una notación de la sintaxis gráfica que acompaña estructuras lingüísticas significativas, tales como procedimientos, instrucciones y demás.

La descripción de cada concepto lingüístico comienza con su diagrama de sintaxis.

Introducción de texto

El texto que se debe introducir exactamente como está visualizado, se muestra en una fuente de máquina de escribir:

Este ejemplo describe un fragmento de sintaxis en el que la palabra clave `OUTPUT` termina con un carácter de punto y coma: `;`. La forma de diagrama de sintaxis es: 
.

Generalmente, las mayúsculas y minúsculas del texto no son significativas, pero en este aspecto, la convención es usar mayúsculas para palabras clave.

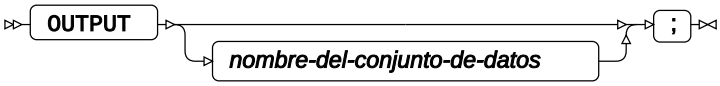
Elementos de los marcadores de posición

Los marcadores de posición que se deben sustituir con el texto pertinente y dependiente del contexto, se reproducen en una fuente minúscula y cursiva:

Aquí, la palabra clave `OUTPUT` se debe introducir literalmente, pero *nombre-del-conjunto-de-datos* se debe sustituir con algo apropiado al programa, en este caso, el nombre de un conjunto de datos al que agregar una observación.

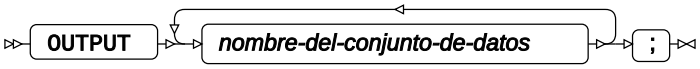
Opcionalidad

Cuando los elementos son opcionales, aparecen en una rama por debajo de la línea principal en diagramas del ferrocarril. La opcionalidad es representada por una ruta de acceso alternativa sin obstáculos a través del diagrama:



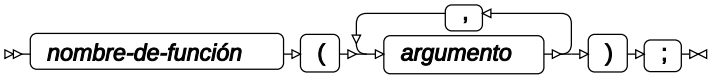
Repetición

En los diagramas de sintaxis, la repetición se representa con un bucle de retorno que opcionalmente especifica el separador que se debe colocar entre varias instancias.



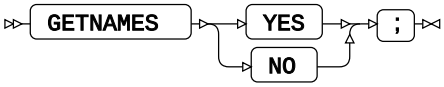
La palabra clave `OUTPUT` arriba se debe introducir literalmente y terminar con una o más repeticiones de *nombre-de-conjunto-de-datos*, en este caso, no se requiere ningún separador a excepción de un espacio.

El siguiente ejemplo muestra el uso de un separador.



Elecciones

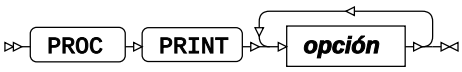
En los diagramas de sintaxis, la elección se muestra con varias ramas paralelas.



En el ejemplo de arriba, la palabra clave `GETNAMES` se debe introducir literalmente y, a continuación, la palabra clave `YES` o la palabra clave `NO`.

Fragmentos

Cuando la sintaxis es demasiado complicada para encajarse en una definición, podría dividirse en fragmentos:



opción

Arriba, la sintaxis completa se divide en fragmentos separados de diagramas de sintaxis. El primero indica que `PROC PRINT` debe terminarse con una o más instancias de una *opción*, cada una de las cuales debe adherir a la sintaxis proporcionada en el segundo diagrama.

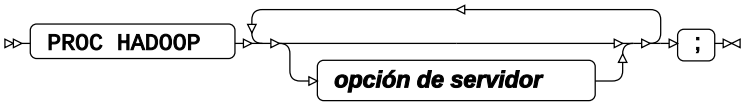
Procedimiento HADOOP

Instrucciones admitidas

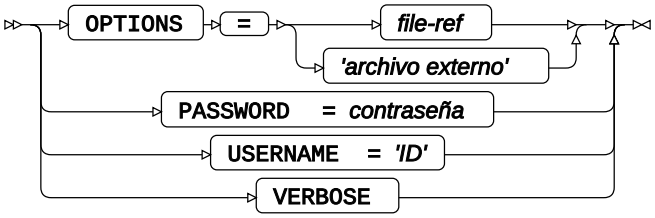
- *PROC HADOOP* (pág. 28)
- *HDFS* (pág. 29)
- *MAPREDUCE* (pág. 29)
- *PIG* (pág. 30)

PROC HADOOP

Accede a Hadoop a través de WPS.

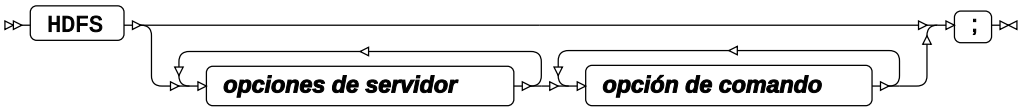


opción de servidor

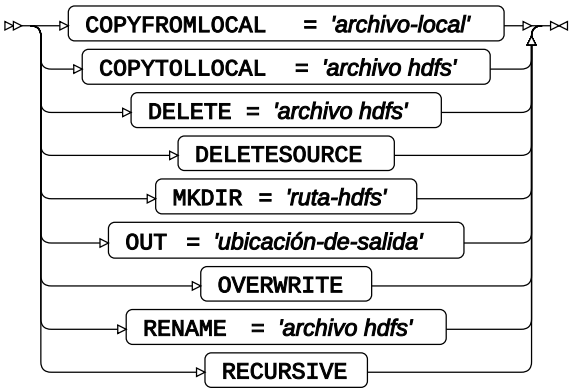


HDFS

Especifica el sistema de archivos distribuido por Hadoop para el uso.

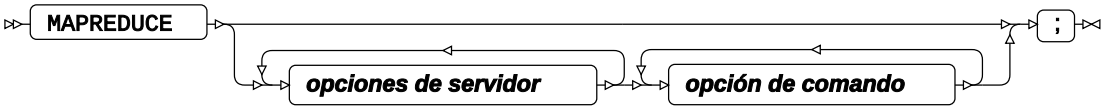


opción de comando

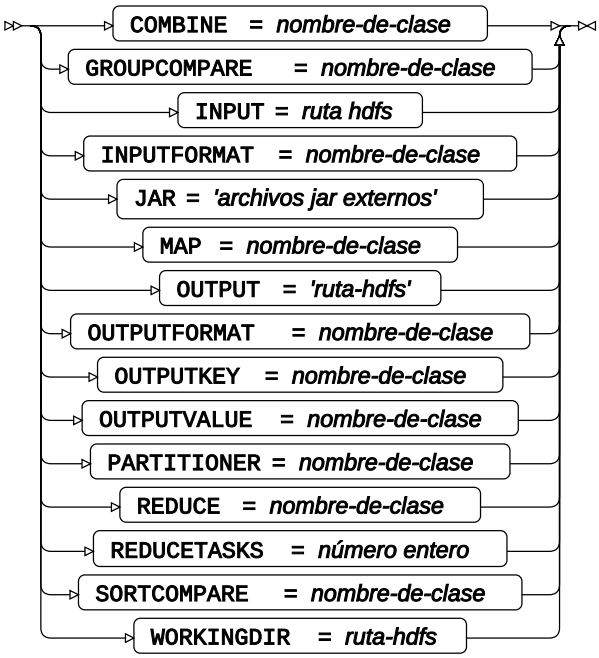


MAPREDUCE

Inicia trabajos de MapReduce.

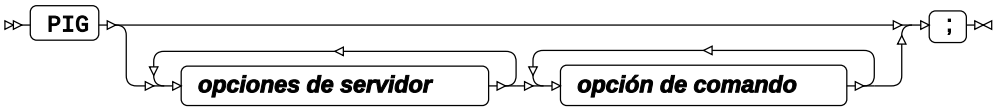


opción de comando

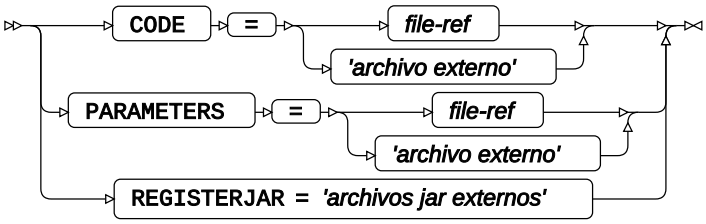


PIG

Permite el envío de archivos externos a un clúster.

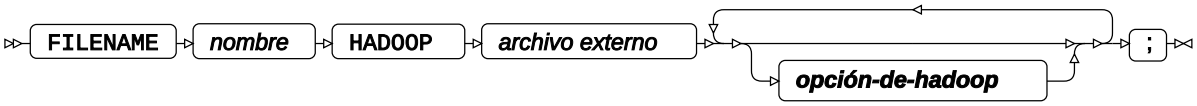


opción de comando

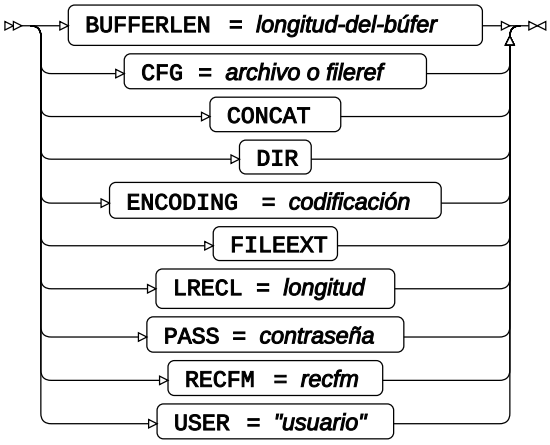


Instrucciones globales

Método de acceso de FILENAME, HADOOP



opción-de-hadoop

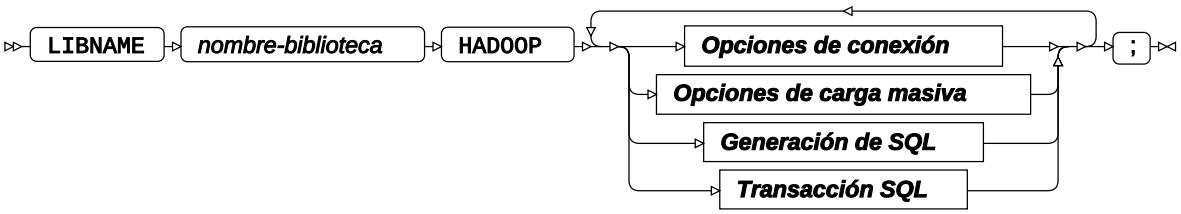


Motor de WPS para Hadoop

El motor WPS para Hadoop proporciona conectividad a una base de datos de Hadoop.

HADOOP

Conéctese a una base de datos especificando el nombre del motor de la base de datos en la instrucción de conexión de la biblioteca `LIBNAME`.



La instrucción `LIBNAME` permite que un programa de lenguaje SAS acceda a una base de datos utilizando el nombre definido en *nombre-biblioteca*. Una referencia de biblioteca sólo está activa durante la duración de una sesión de WPS Analytics. Puede utilizar la referencia de biblioteca especificada en programas para acceder a datos almacenados en una base de datos, siempre que los programas se ejecuten en la sesión de WPS Analytics en la que se especifica la referencia de biblioteca. La instrucción `LIBNAME` contiene opciones que, cuando se especifican, determinan cómo interactúan los programas de lenguaje SAS con la base de datos, agrupados de la siguiente manera:

- *Opciones de conexión* (pág. 32): crean la conexión al servidor de la base de datos.
- *Opciones de carga masiva* (pág. 36): insertan rápidamente grandes cantidades de datos en una base de datos mediante el mecanismo de carga masiva (inserción masiva).
- *Opciones de generación de SQL* (pág. 37): determinan cómo se formatea y utiliza la información de descripción de la tabla o las instrucciones de consulta.
- *Opciones de transacción de SQL* (pág. 41): afectan el comportamiento transaccional de las instrucciones procesadas por WPS Analytics y el servidor de la base de datos.

nombre-biblioteca

Especifica el nombre utilizado en otras instrucciones de lenguaje SAS para hacer referencia a esta conexión a la base de datos.

Por ejemplo, la siguiente instrucción:

```
LIBNAME ExLib HADOOP DATASOURCE=testdb USER=BJames PASSWORD=xxxxxxxx;
```

crea una conexión a una base de datos con el nombre `ExLib`. Este nombre se puede utilizar, por ejemplo, en el procedimiento de SQL:

```
PROC SQL;
  INSERT INTO ExLib.person VALUES (32, 'Smith', 'John', 479216691);
QUIT;
```

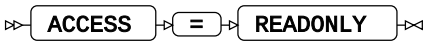
En este programa, el procedimiento de SQL se utiliza para insertar datos en la tabla de la base de datos `person` en la base de datos referenciada por `ExLib`.

Opciones de conexión

crean la conexión al servidor de la base de datos.

ACCESS

Especifica el modo de acceso para la conexión de la biblioteca.



READONLY

La conexión de la biblioteca sólo se puede utilizar para leer datos.

Si especifica esta opción, anula la configuración de inserción o actualización en otras opciones y puede provocar que los datos no se modifiquen como se esperaba.

Si no se especifica esta opción, la conexión a la biblioteca utiliza un modo de acceso de *lectura-escritura* que permite operaciones de lectura, inserción y actualización.

AUTHDOMAIN

Especifica el dominio de autorización del Concentrador de WPS.

Tipo: Cadena

El dominio de autorización proporciona permisos para acceder a un servidor de base de datos. WPS Analytics usa el Concentrador como un dominio de autorización y un servidor del Concentrador debe estar disponible para su sistema.

En este ejemplo, los permisos para acceder al Concentrador se proporcionan como opciones del sistema y el nombre del dominio de autorización que contiene los detalles de autorización en el Concentrador, se especifica como `AUTHDOMAIN`.

```
OPTIONS HUB_SERVER='blue_streak' HUB_PORT=309 HUB_PROTOCOL='HTTP'
HUB_USER='ARichards' HUB_PWD='xxxxxxxx';
LIBNAME ExLib HADOOP DATASRC=ExDB AUTHDOMAIN='OracleAuth';
```

CONFIG

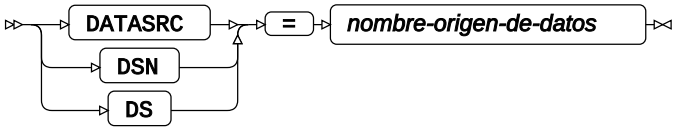
Especifica las opciones de configuración.

Tipo: Cadena

Las opciones están separadas por espacios. La lista debe estar encerrada entre comillas.

DATASRC

Especifica un nombre de origen de datos (DSN).

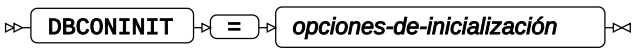


Tipo: Cadena

Se debe crear un DSN antes de que se pueda especificar en *nombre-origen-de-datos*.

DBCONINIT

Especifica instrucciones de SQL o comandos de base de datos que se procesan cada vez que se abre la conexión a la biblioteca.



Tipo: Cadena

La conexión se abre cuando esta instrucción `LIBNAME` se ejecuta por primera vez, y luego cada vez que se realiza una conexión a la base de datos; por ejemplo, ejecutando una instrucción de SQL en el procedimiento de SQL.

DBCONTERM

Especifica instrucciones de SQL o comandos de base de datos que se procesan cada vez que se cierra la conexión a la biblioteca.

Tipo: Cadena

La conexión de la biblioteca se cierra al final de un procedimiento, un paso DATA y al final de una sesión.

DBLIBINIT

Especifica instrucciones de SQL o comandos de base de datos que se procesan inmediatamente después de que se crea la conexión a la biblioteca usando la instrucción `LIBNAME` .

Tipo: Cadena

DBLIBTERM

Especifica instrucciones de SQL o comandos de base de datos que se procesan inmediatamente antes de que se desconecte una conexión de biblioteca cuando finaliza la sesión de WPS Analytics.

Tipo: Cadena

HIVE_PRINCIPAL

Especifica el principal de un Hive.

Tipo: Cadena

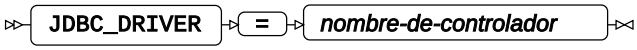
JDBC_CONNECTION_STRING

Especifica la cadena de conexión para JDBC.

Tipo: Cadena

JDBC_DRIVER

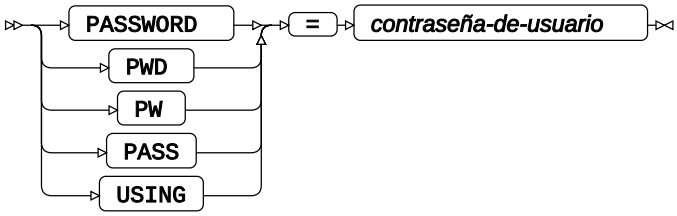
Especifica el controlador JDBC que se utilizará para la conexión de Hive.



Tipo: Cadena

PASSWORD

Especifica la contraseña para el nombre de usuario.



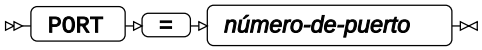
Tipo: Cadena

Si se utilizan caracteres especiales como parte de la cadena, debe escribir *contraseña-de-usuario* entre comillas. El nombre de usuario se especifica con la opción `USER`.

Si ha especificado `AUTHDOMAIN` para habilitar la autorización a través del Concentrador, no es necesario que especifique el nombre de usuario y la contraseña.

PORT

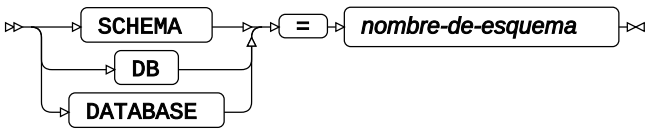
Especifica el número de puerto del servidor de base de datos especificado por la opción `SERVER`.



Tipo: Numérico

SCHEMA

Especifica el nombre del esquema de la base de datos con el que interactúa la conexión.

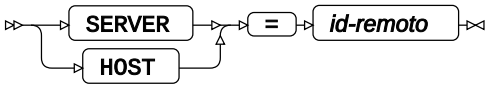


Tipo: Cadena

El esquema es una agrupación de datos y objetos de la base de datos a la que puede acceder el usuario y que puede manipularse mediante instrucciones de SQL.

SERVER

Especifica el nombre o la dirección TCP/IP del equipo que aloja el servidor de la base de datos.

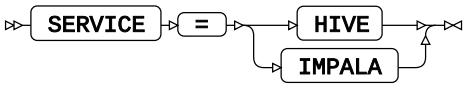


Tipo: Cadena

Si no se especifica un servidor de base de datos externo, se asume que el servidor de base de datos y la base de datos están en el dispositivo en el que WPS Analytics está instalado y se ejecuta.

SERVICE

Especifica el tipo de servicio Hadoop.



HIVE

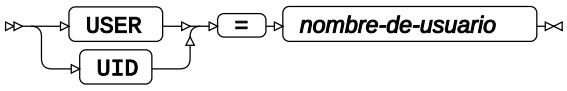
Especifica que el servicio es Hive.

IMPALA

Especifica que el servicio es Impala.

USER

Especifica el nombre de usuario necesario para acceder a la base de datos.



Tipo: Cadena

Si se utilizan caracteres especiales como parte de la cadena, debe escribir *nombre-de-usuario* entre comillas.

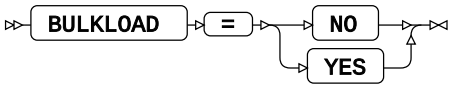
Si ha especificado `AUTHDOMAIN` para habilitar la autorización a través del Concentrador, no tiene que especificar el nombre de usuario y la contraseña.

Opciones de carga masiva

insertan rápidamente grandes cantidades de datos en una base de datos mediante el mecanismo de carga masiva (inserción masiva).

BULKLOAD

Especifica si los datos se insertan en una tabla mediante la funcionalidad de carga masiva.



NO

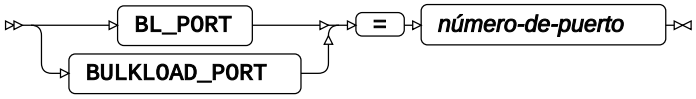
Las inserciones de datos no utilizan la funcionalidad de carga masiva. Todas las demás opciones de carga masiva se ignoran.

YES

Los insertos de datos utilizan la funcionalidad de carga masiva.

BL_PORT

Especifica un puerto de servidor a utilizar para la carga masiva de datos.



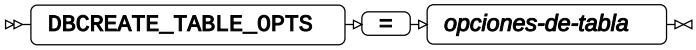
Tipo: Numérico

Generación de SQL

Afecta cómo se crean las instrucciones de SQL y si WPS Analytics o el servidor de la base de datos crea las instrucciones.

DBCREATE_TABLE_OPTS

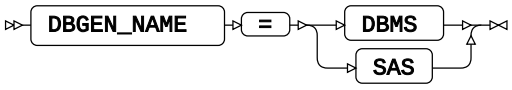
Especifica opciones de tabla adicionales que se agregarán a una instrucción SQL `CREATE TABLE` después de que se hayan definido las columnas de la tabla.



Tipo: Cadena

DBGEN_NAME

Especifica cómo modificar los nombres de columna no válidos para que sean nombres de variable de lenguaje SAS válidos.



Valor predeterminado: `SAS`

DBMS

Si el nombre de la columna contiene caracteres no válidos, esos caracteres se reemplazan por un guión bajo. Si este cambio da como resultado un nombre que entra en conflicto con el de otra columna, el recuento de la columna se agrega al nombre de la columna.

Por ejemplo, si la tabla de su base de datos contiene respectivamente las columnas `id$x`, `id#x` y `id_x`, las variables en el conjunto de datos de lenguaje SAS sería `ID_X`, `ID_X0` y `ID_X1`.

Nota:

La numeración comienza en la primera repetición del nombre de la variable y comienza en `0`.

SAS

Si el nombre de la columna contiene caracteres no válidos o tiene el mismo nombre que otra columna de la tabla, el nombre se reemplaza con la cadena `_COL`. Si este cambio da como resultado un conflicto de nombres con otras columnas, el recuento de columnas se agrega al nombre: `_COL`*x* donde *x* es el recuento, comenzando en 0 (cero).

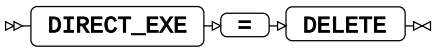
Por ejemplo, si su tabla contiene las columnas `id$x`, `id#x` y `id_x`, el conjunto de datos tendría las variables correspondientes `_COL0`, `_COL1` y `ID_X`.

Nota:

La numeración comienza en la instancia del nombre creado y comienza en `0`.

DIRECT_EXE

Especifica si las instrucciones de eliminación las procesa el motor de la base de datos o WPS Analytics.



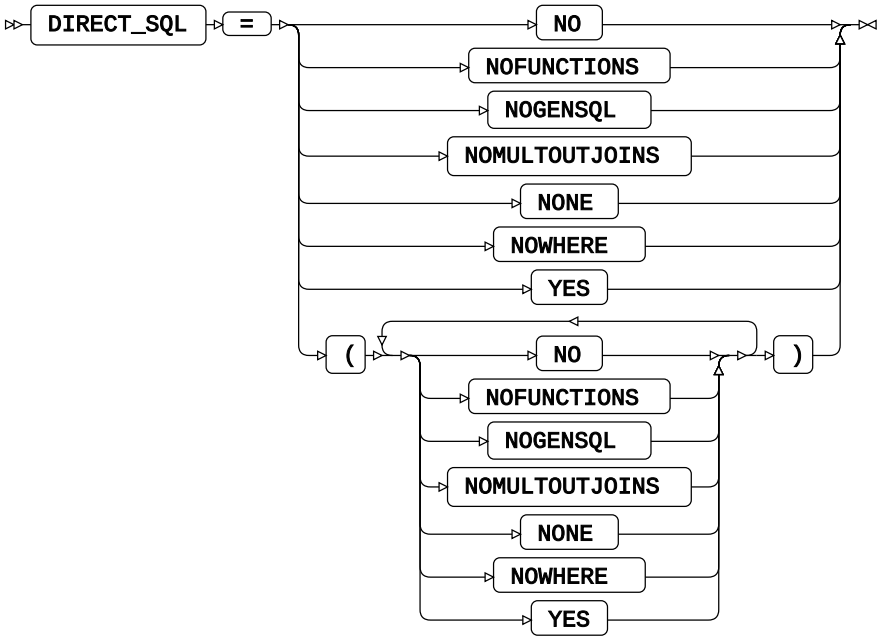
Cuando no se especifica, la tabla de la base de datos se analiza y las filas que coinciden con los criterios de eliminación se devuelven a WPS Analytics. WPS Analytics crea una instrucción de SQL `DELETE FROM` para cada fila afectada y la pasa a la base de datos para su procesamiento.

DELETE

Especifica que un comando de SQL `DELETE FROM` se pasa a la base de datos y el motor de la base de datos lo procesa en su totalidad.

DIRECT_SQL

Especifica si las instrucciones de SQL se pasan para que las procese el motor de la base de datos o WPS Analytics.



Valor predeterminado: `YES`

NO

WPS Analytics procesa todas las instrucciones de SQL.

NOFUNCTIONS

WPS Analytics procesa cualquier instrucción de SQL que contenga llamadas a funciones.. Todas las demás instrucciones que el motor de la base de datos puede procesar se pasan al servidor de la base de datos para su procesamiento. Especificar `DIRECT_SQL = NOFUNCTIONS` anula la opción `SQL_FUNCTIONS`.

NOGENSQL

WPS Analytics procesa todas las instrucciones de SQL..

NOMULTOUTJOINS

WPS Analytics procesa cualquier instrucción de SQL que contenga múltiples uniones externas.. Todas las demás instrucciones que el motor de la base de datos puede procesar se pasan al servidor de la base de datos para su procesamiento.

NONE

WPS Analytics procesa todas las instrucciones de SQL.

NOWHERE

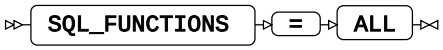
WPS Analytics procesa cualquier instrucción de SQL que contenga clausulas `WHERE`.. Todas las demás instrucciones que el motor de la base de datos puede procesar se pasan al servidor de la base de datos para su procesamiento.

YES

Se le pasan todas las instrucciones de SQL que puede procesar el servidor de la base de datos.

SQL_FUNCTIONS

Especifica qué funciones del lenguaje SAS procesa el motor de la base de datos.



Si no se especifica esta opción, las siguientes funciones se pasan al motor de la base de datos para su procesamiento cuando se utilizan en instrucciones de lenguaje SAS, tal como en el procedimiento de SQL o la opción del conjunto de datos `WHERE`.

ABS()	ARCOS()	ARSIN()	ATAN()
CEIL()	COALESCE()	COS()	COUNT()
DAY()	EXP()	FLOOR()	HOUR()
INDEX()	LOG()	LOG10()	LOWCASE()
MAX()	MIN()	MINUTE()	MONTH()
SECOND()	SIN()	SQRT()	STD()
STRIP()	SUBSTR()	SUM()	TAN()
TRANSTRN()	UPCASE()	VAR()	YEAR()

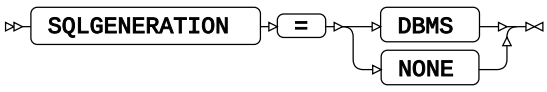
ALL

Especifica que, además del conjunto predeterminado de funciones del lenguaje SAS, lo siguiente también se pasa al motor de la base de datos para su procesamiento:

COMPRESS()	DATE()	DATEPART()
DATETIME()	LENGTH()	REPEAT()
TODAY()	TRIMN()	

SQLGENERATION

Especifica si el servidor de la base de datos o WPS Analytics procesa los datos y las instrucciones de SQL.



DBMS

Las instrucciones se generan y los datos se procesan en el servidor de la base de datos.

NONE

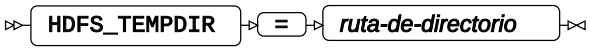
Las instrucciones se generan y los datos se procesan en WPS Analytics. A continuación, los resultados se envían al servidor de la base de datos.

Transacción SQL

afectan el comportamiento transaccional de las instrucciones procesadas por WPS Analytics y el servidor de la base de datos.

HDFS_TEMPDIR

Especifica la ubicación utilizada para los directorios HDFS temporales.

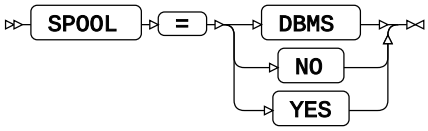


Tipo: Cadena

De forma predeterminada, el parámetro de configuración de Hadoop `hadoop.tmp.dir` especifica la ubicación de los archivos temporales locales y HDFS. Puede utilizar este parámetro para especificar un directorio diferente para los archivos HDFS temporales.

SPOOL

Especifica si se crea un archivo de cola de impresión.



Se utiliza un archivo de cola de impresión para garantizar que todas las operaciones de lectura realizadas durante la ejecución de una solicitud de SQL sean las mismas. La cola de impresión siempre ocurre si es necesario. Lo determina el planificador de consultas.

DBMS

El motor de la base de datos administra la cola de impresión.

NO

Proporcionado por compatibilidad. Sin embargo, como la cola de impresión siempre ocurre cuando es necesario, si esta opción está configurada, el motor de la base de datos la administra.

YES

WPS Analytics administra la cola de impresión. Este es el valor predeterminado.

Avisos legales

Marcas comerciales

WPS e World Programming son marcas registradas o comerciales de World Programming Limited en la Unión Europea y en otros países. (r) o ® indican una marca comunitaria.

SAS y todos los otros nombres de productos o servicios de SAS Institute Inc. son marcas registradas o comerciales de SAS Institute Inc. en los EE.UU. y en otros países. ® indica la registración en los EE.UU.

Todas las otras marcas comerciales mencionadas pertenecen a sus respectivos propietarios.

Avisos generales

World Programming Limited no está asociada de ninguna manera con SAS Institute Inc.

WPS no es SAS System.

Las expresiones "SAS", "lenguaje SAS" y "lenguaje de SAS" utilizadas en este documento, se usan en referencia al lenguaje de programación, llamado a menudo en una de dichas maneras.

Las expresiones "programa", "programa SAS" y "programa en el lenguaje SAS" utilizadas en este documento, se usan en referencia a los programas escritos en el lenguaje SAS. También se conocen como "scripts", "scripts SAS" o "scripts en el lenguaje SAS".

Las expresiones "IML", "lenguaje IML", "sintaxis IML", "Interactive Matrix Language" y "lenguaje de IML" utilizadas en este documento, se usan en referencia al lenguaje de programación, llamado a menudo en una de dichas maneras.

WPS incluye software desarrollado por terceros. Se puede encontrar más información en el archivo THANKS o acknowledgments.txt, incluidos en la instalación de WPS.