Configuration de WPS pour Hadoop

Version: 4.4.2

www.worldprogramming.com

Table des matières

Introduction

Qu'est-ce que Hadoop ?

Hadoop est un système open-source offrant une infrastructure logicielle évolutive et résistante aux fautes pour le stockage et le traitement distribués de très grands ensembles de données sur les grappes d'ordinateurs. Il est diffusé sous une licence Apache.

Les lecteurs qui découvrent Hadoop sont invités à consulter d'abord la section *Bases de Hadoop* (p. 7).

Avantages apportés par l'intégration de Hadoop dans WPS

- Avec l'intégration de Hadoop, WPS accroît ses capacités en matière d'acquisition des données à une douzaine de moteurs de base de données.
- Le partage des données entre WPS et HDFS (Hadoop Distributed File System, système de fichiers distribué Hadoop) permet une interopérabilité entre les deux environnements au niveau des données. Bien qu'il ne soit pas transparent, il est simple : il est possible d'importer les données Hadoop dans WPS pour une analyse (structurée) et, si cela est requis, de les renvoyer vers HDFS.
- Les utilisateurs de WPS peuvent invoquer la fonctionnalité depuis l'environnement familier de WPS Workbench.
- Les utilisateurs peuvent créer et modifier de nouvelles opérations Hadoop à l'aide d'un langage similaire à SQL – pas besoin de connaître Java.

Portée du document

Ce document donne un aperçu de l'implémentation de WPS avec Hadoop, et inclut la configuration de Kerberos lorsque cela est applicable.

Résumé de l'intégration WPS/Hadoop

Les intégrations suivantes, actuellement mises en œuvre, utilisent les extensions `filename`, `libname` et `PROC HADOOP` de WPS :

- Connexion à Hive à l'aide de syntaxe SQL standard
- Connexion à Impala à l'aide de syntaxe SQL standard
- Connexion à Hive à l'aide de syntaxe SQL à transfert direct
- Connexion à Impala à l'aide de syntaxe SQL à transfert direct
- Envoi de commandes HDFS et exécution de programmes Pig

L'intégration de WPS avec Hadoop a été certifiée sur Cloudera 5 et testée sur d'autres distributions de Hadoop qui restent proches de la référence Apache. Vous trouverez plusieurs exemples de code concernant l'intégration (p. 17) à la fin de ce document.

Conditions préalables

Hadoop est une pile technologique complexe et comportant de nombreux éléments. Avant de procéder à son intégration avec WPS, il convient de l'installer et de le configurer correctement. Les étapes préparatoires suivantes doivent être réalisées, vérifiées et pointées.

1. Procurez-vous l'ensemble de fichiers `.jar` correspondant à votre installation Hadoop.

   Remarque :

   Si vous utilisez Apache Hive dans le cadre votre installation Hadoop avec WPS, vous devez utiliser Apache Hive version 0.12 ou ultérieure.

2. Paramétrez les fichiers de configuration XML conformément à votre environnement de cluster (adresses IP, ports, etc.).
3. Déterminez si votre distribution de Hadoop inclut ou nécessite la prise en charge de Kerberos. Si c'est le cas, vérifiez que l'authentification Kerberos sur votre serveur fonctionne, que le principal a été configuré correctement, et ainsi de suite. Que Kerberos soit utilisé ou non, effectuez les opérations suivantes pour remplir les conditions préalables.
4. Assurez-vous que le cluster fonctionne correctement, éventuellement en demandant à votre administrateur de cluster, qui devrait avoir accès aux tableaux de bord d'administration.
5. Une fois que vous avez déterminé que le cluster fonctionne correctement, vérifiez qu'il est possible de soumettre des tâches Hadoop indépendamment de WPS.

Kerberos

La sécurisation de l'accès dans Hadoop est basée sur l'établissement de l'identité avec une authentification forte. Les utilisateurs doivent être en mesure de s'identifier pour accéder aux ressources, et les ressources Hadoop elles-mêmes doivent être authentifiées individuellement pour éviter que des systèmes malveillants se « déguisent » et fassent semblant d'appartenir au cluster pour accéder aux données. Pour créer cette communication sécurisée entre ses divers composants, Hadoop peut utiliser Kerberos, un mécanisme d'authentification tiers. Les utilisateurs et les services auxquels ils veulent accéder font alors confiance à Kerberos pour la gestion de l'authentification.

Remarque :

Certaines distributions de Hadoop incluent (voire exigent) la prise en charge de Kerberos. Les détails de la configuration du serveur Kerberos varient souvent en fonction du type de distribution, et ne relèvent pas de ce guide. Consultez les informations de configuration spécifiques fournies avec votre logiciel Hadoop. Voir *Configurer Kerberos et Hadoop côté client* (p. 16) pour savoir comment configurer Kerberos et Hadoop côté client.

Bases de Hadoop

Dans les environnements d'analyse traditionnels, les données sont transmises à un SGBDR via un processus ETL (extraction, transformation et chargement) initial. Les données non structurées sont préparées et chargées dans la bases de données. Un schéma de données leur est appliqué au cours de la préparation. Une fois chargées, elles peuvent être soumises à de nombreuses techniques d'analyse bien établies.

Pour les gros volumes de données, toutefois, ce processus pose quelques problèmes :

1. Si le délai de traitement des données d'une journée est tel qu'il n'est pas possible de terminer l'analyse avant le lendemain de manière économiquement viable, il faut trouver une nouvelle approche. Les opérations d'ETL à grande échelle imposent une pression énorme sur l'infrastructure sous-jacente.
2. À mesure que les données vieillissent, elles sont souvent archivées. Toutefois, il est très coûteux d'extraire de gros volumes de données archivées (que ce soit sur bande, blu-ray ou autre). De plus, une fois archivées, l'accès aux données n'est plus économique ni pratique.
3. L''ETL est un processus d'abstraction : les données sont agrégées et normalisées, et leur format haute fidélité est perdu. Si l'entreprise pose par la suite un nouveau type de question sur les données, il n'est souvent pas possible d'y répondre sans une procédure coûteuse, avec modification de la logique d'ETL, correction du schéma de base de données et rechargement des données.

Hadoop a été conçu pour :

- Offrir de l'évolutivité pour le calcul et les données, en éliminant le goulot d'étranglement de l'ETL.
- Rendre économiquement viable la conservation des données sur le stockage primaire à plus long terme.
- Permettre plus de flexibilité, comme poser de nouvelles questions sur les données haute fidélité d'origine.

Comparaison entre SGBDR et Hadoop

Au niveau de l'analyse, les principales différences entre un SGBDR et Hadoop sont indiquées ci-dessous.

Table 1. Principales différences entre SGBDR et Hadoop

SGBDR	Hadoop
Le schéma doit être créé avant le chargement des données.	Les données sont simplement copiées dans le stockage de fichiers. Aucune transformation n'est nécessaire.

SGBDR	Hadoop
Une opération ETL explicite doit avoir lieu pour appliquer aux données la structure interne de la base de données.	Un sérialiseur/désérialiseur (« Serdé ») est appliqué à la lecture pour extraire les colonnes requises.
L'ajout de nouvelles colonnes doit être demandé de manière explicite pour qu'elles soient chargées.	De nouvelles données peuvent arriver à tout moment, et apparaîtront de manière rétrospective une fois le Serdé mis à jour pour les analyser.

L'orientation schéma des implémentations conventionnelles de SGBDR apporte des avantages importants, qui ont favorisé leur adoption généralisée :

- Optimisations, index, partitionnement et autres deviennent possibles, permettant des lectures très rapides pour certaines opérations telles que les jointures, les jointures multi-tables etc.
- Un schéma d'organisation commun à toute l'entreprise signifie que tous les services peuvent utiliser le même vocabulaire.

Par contre, les implémentations SGBDR sont moins flexibles, et ne peuvent pas faire évoluer le traitement des données aussi vite que leur acquisition. Avec Hadoop, la structure n'est appliquée aux données qu'au moment de la lecture, via un sérialiseur/désérialiseur et, par conséquent, il n'y a pas de phase d'ETL. Les fichiers sont simplement copiés sur le système. Hadoop n'est pas une base de données au sens conventionnel du terme, en raison de ses propriétés ACID (atomicité, cohérence, isolation et durabilité), et même s'il l'était, il serait sans doute trop lent pour la plupart des applications interactives.

Chacune des deux technologies peut bénéficier du soutien de l'autre. Elles ont toutes deux leur place dans l'organisation informatique de l'entreprise. Il suffit de choisir le bon outil pour chaque tâche.

Table 2. SGBDR ou Hadoop : Cas d'utilisation

Quand utiliser un SGBDR	Quand utiliser Hadoop
OLAP interactif – temps de réaction inférieur à une seconde.	Lorsque vous devez gérer des données structurées et non structurées.
Lorsque vous devez prendre en charge des transactions ACID à étapes multiples sur des données basées sur des enregistrements (distributeurs de billets, etc.).	Lorsque l'évolutivité du stockage et/ou du calcul est requise.
Lorsqu'une conformité totale à SQL est requise.	Lorsque vous avez des besoins de traitement de données complexes avec de très gros volumes de données.

Architecture Hadoop

Hadoop repose sur deux concepts principaux :

- HDFS (Hadoop Distributed File System, système de fichiers distribué Hadoop) – un système de fichiers basé sur Java qui offre un stockage évolutif et fiable réparti sur de gros clusters de serveurs de grande série (« commodity servers »).
- MapReduce – un modèle de programmation qui simplifie la tâche de rédaction des programmes utilisés dans un environnement informatique parallèle.

Un cluster Hadoop opérationnel comporte de nombreux autres sous-systèmes, mais HDFS et MapReduce sont au centre du modèle de traitement.

HDFS

HDFS est un système de fichiers distribué, évolutif et adaptable écrit en Java. HDFS stocke de gros fichiers (généralement de l'ordre du giga- ou du téra-octet) répartis sur plusieurs machines. Sa fiabilité est due à la réplication des données sur plusieurs hôtes. Par défaut, les blocs de données sont stockés (répliqués) sur trois nœuds – deux sur un même rack, et un sur un autre rack, soit un triplement des coûts par rapport aux stockages non répliqués. Les nœuds de données peuvent communiquer entre eux pour répartir les données, se transmettre des copies et maintenir un niveau élevé de réplication des données.

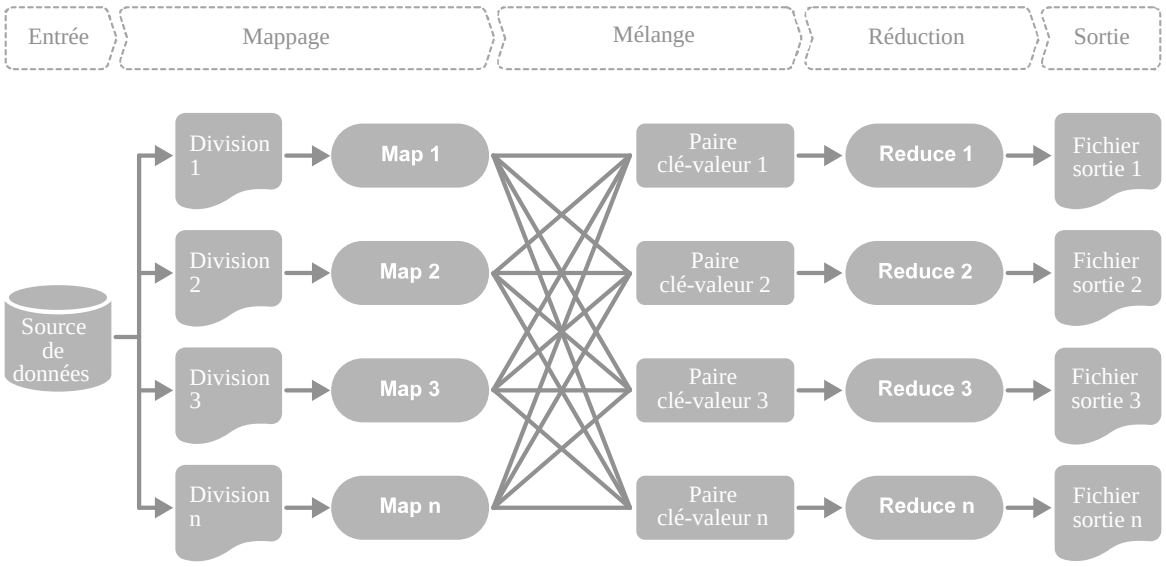
HDFS n'est pas totalement conforme à la norme Posix, et est optimisé pour offrir un débit maximal. Certaines opérations de fichier au niveau atomique sont soit interdites, soit lentes. Ainsi, il n'est pas possible d'insérer des données au milieu d'un fichier, mais vous pouvez en ajouter à la fin.

MapReduce

MapReduce est une infrastructure de programmation qui, si elle est respectée, élimine la complexité de la programmation pour les environnements massivement parallèles.

Un programmeur doit généralement écrire deux fonctions, une fonction Map et une fonction Reduce, et d'autres composants de l'infrastructure Hadoop se chargent de la tolérance aux défaillances, de la distribution, de l'agrégation, du tri et autres. L'exemple souvent cité est la production d'un compte de fréquence des mots sur un grand nombre de documents. Les étapes suivantes sont utilisées :

1. Le système divise les données entrées entre plusieurs nœuds, appelés « mappeurs ». C'est là que le programmeur écrit une fonction qui compte le nombre de mots différents dans un fichier, et le nombre d'occurrences de chaque. Il s'agit de la fonction Map, dont la sortie est un ensemble de paires clé-valeur composées d'un mot et d'un nombre. Chaque mappeur effectue cette opération sur le lot de documents en entrée qui lui est attribué. Ainsi, au total, de nombreux mappeurs produisent de nombreux ensembles de paires clé-valeurs pour l'étape suivante.
2. La phase de mélange a alors lieu – une fonction de hachage est appliquée aux paires clé-valeur, et les valeurs obtenues sont redistribuées aux réducteurs de manière que toutes les paires ayant la même clé aillent au même réducteur.
3. Le programmeur a écrit une fonction Reduce qui, dans ce cas, additionne simplement les occurrences de mots fournies par les flux entrants de paires clé-valeur, et écrit les totaux dans un fichier de sortie :



Ce processus isole le programmeur des problèmes d'évolutivité à mesure que le cluster grandit. Une partie du système Hadoop lui-même se charge de gérer et d'exécuter les ressources. Cette partie est YARN si la version de MapReduce utilisée est 2.0 ou ultérieure.

Il n'est pas garanti que ce processus soit plus rapide qu'un autre système (bien que, en pratique, il soit plus rapide pour certains types de problèmes et pour les gros volumes de données). Le principal avantage de ce modèle de programmation est la possibilité d'exploiter l'opération de mélange, souvent optimisée, et de n'avoir à écrire que les parties mappage et réduction du programme.

L'écosystème Hadoop

Il existe plusieurs manières d'interagir avec un cluster Hadoop.

Java MapReduce

Il s'agit de la méthode d'accès la plus flexible et la mieux adaptée, bien que, comme il s'agit du langage assembleur de Hadoop, le cycle de développement puisse être compliqué.

Streaming MapReduce

Ceci permet le développement pour Hadoop dans le langage de programmation de votre choix, en échange d'une légère perte au niveau des performances et de la flexibilité. Cette méthode s'appuie toujours sur le modèle MapReduce, mais étend la gamme des langages de programmation disponibles.

Crunch

Crunch est une bibliothèque de pipelines MapReduce à plusieurs étapes écrites en Java, sur le modèle de la bibliothèque FlumeJava de Google. Elle offre une API Java pour des tâches telles que la jointure et l'agrégation de données qui sont fastidieuses à réaliser sur un MapReduce ordinaire.

Pig Latin

Pig Latin (souvent simplement appelé « Pig ») est un langage de haut niveau adapté aux charges de travail en flux de données par lot. Avec Pig, il n'est plus nécessaire de penser en termes de MapReduce. Il ouvre le système aux programmeurs ne connaissant pas Java, et fournit des opérations communes telles que les jointures, les groupes, les filtres et les tris.

Hive

Hive est un interpréteur SQL (non conforme aux spécifications SQL) qui inclut un métastore (ou métastockage) qui permet de mapper les fichiers sur leurs schémas, ainsi que des sérialiseurs/désérialiseurs associés. Comme Hive est basé sur SQL, les pilotes ODBC et JDBC permettent d'accéder aux outils d'informatique décisionnelle standard tels qu'Excel.

Oozie

Oozie est un moteur de flux de travail XML en PDL (Prime Data Language) qui permet de créer un flux de travail de tâches constitué des divers éléments présentés plus haut.

HBase

Apache HBase vise l'hébergement de tables très volumineuses – des milliards de lignes et des millions de colonnes – sur des clusters de serveurs de grande série. Sur le modèle du système Bigtable de Google, Hbase offre des fonctionnalités similaires à Bigtable sur Hadoop et HDFS.

Zookeeper

Apache Zookeeper est un projet de développement et de maintenance d'un serveur open-source qui offre une coordination extrêmement fiable des systèmes distribués. Zookeeper est un service centralisé pour la maintenance des informations de configuration et du nommage, et offre également une synchronisation distribuée et des services de groupe.

Implémentation de WPS avec Hadoop sur Windows x64

Installer WPS sur Windows x64

1. Avant d'installer WPS, vérifiez que les mises à jour et service packs les plus récents ont été appliqués à Windows.
2. Le même logiciel WPS est utilisé pour les installations Windows sur poste de travail et sur serveur. L'utilisation est contrôlée par la clé de licence appliquée lors de la procédure `setinit`.
3. Le fichier d'installation de WPS pour Windows est disponible sur le site Web de World Programming. Vous devez disposer d'un nom d'utilisateur et d'un mot de passe pour accéder à la section Téléchargement du site.
4. Une fois le fichier d'installation (.msi) téléchargé, il suffit de double-cliquer dessus, de lire et d'accepter le CLUF et de suivre les intructions à l'écran.
5. Une fois le logiciel WPS installé, vous devez appliquer la clé de licence. Cette clé vous a été envoyée par courriel lorsque vous avez acheté le logiciel WPS. La méthode la plus simple pour appliquer la licence consiste à lancer WPS Workbench en tant qu'utilisateur bénéficiant de l'accès administrateur système et à suivre les instructions.
6. La configuration de WPS est terminée.

Configurer Hadoop sur Windows x64

Installer Hadoop

Si ce n'est pas encore fait, installez Hadoop en vous référent à la documentation fournie pour votre distribution (Cloudera, par exemple). Une fois Hadoop installé, vous pouvez configurer les détails décrits ci-dessous.

Remarque :

Si vous avez une distribution qui fonctionne conformément au standard Apache Hadoop, les détails de configuration devraient être applicables, même si vous n'utilisez pas une distribution Cloudera. Les distribution qui désactivent ou modifient les fonctionnalités du standard Apache Hadoop ne sont pas prises en charge.

Fichiers de configuration

Tous les appels à Hadoop Cloudera 5 sont réalisés via Java et JNI. Vous devez vous procurer les fichiers `.jar` du client Hadoop Cloudera et les télécharger sur la machine locale. Les fichiers suivants contiennent des URL pour les divers services Hadoop, et doivent être configurés de sorte à correspondre à l'installation actuelle de Hadoop :

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

> **Remarque :**
> Si vous utilisez un client Windows avec un cluster Linux, il faut définir le paramètre `mapreduce.app-submission.cross-platform` de ce fichier sur `true`.

Veuillez consulter la documentation de Hadoop pour pour plus d'informations.

Variable d'environnement CLASSPATH

La variable d'environnement `CLASSPATH` doit être configurée de sorte à pointer vers les fichiers du client Java Cloudera. Elle peut varier selon la configuration du client utilisé et la machine, mais elle peut ressembler à :

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

Variables d'environnement HADOOP_HOME

Sur Windows, la variable d'environnement `HADOOP_HOME` doit être configurée de sorte à pointer vers les fichiers du client Java Cloudera. Pour l'exemple ci-dessus, elle devrait avoir la valeur : `C:\Cloudera5`.

Configurer Kerberos sur Windows x64

Si votre distribution de Hadoop inclut ou nécessite la prise en charge de Kerberos, allez à la section *Configurer Kerberos et Hadoop côté client* (p. 16).

Implémentation de WPS avec Hadoop sur Linux x64

Installer WPS sur Windows x64

1. WPS est compatible avec toute distribution de Linux conforme à Linux Standard Base (LSB) 3.0 ou ultérieur. WPS est compatible avec Linux sur les plates-formes x86, x86_64 et IBM System z, y compris Integrated Facility for Linux (IFL).
2. Si vous avez installé une distribution 64 bits de Linux, vous pouvez utiliser une version 32 ou 64 bits de WPS. Il faut cependant noter que certaines distributions 64 bits de Linux n'installent par défaut que les bibliothèques 64 bits. WPS 64 bits est directement utilisable sur ces distributions. En revanche, si vous installez WPS 32 bits, vous devez d'abord installer les bibliothèques système 32 bits. Consultez la documentation de votre distribution de Linux pour savoir comment réaliser cette opération.
3. WPS pour Linux n'est actuellement disponible qu'en tant que fichier d'archive compressé tar. Un programme d'installation natif basé sur RPM est prévu.
4. L'archive de WPS pour Linux est fournie au format tar gzip (`.tar.gz`) et est disponible sur le site Web de World Programming. Vous devez disposer d'un nom d'utilisateur et d'un mot de passe pour accéder à la section Téléchargement du site.
5. Pour installer WPS, extrayez les fichiers de l'archive à l'aide de gunzip et tar comme suit. Choisissez un emplacement d'installation approprié et auquel vous avez accès en écriture et accédez à ce répertoire à l'aide de la commande `cd`. L'archive est complète, et peut être ouverte à l'emplacement de votre choix. L'emplacement d'installation peut nécessiter un accès en tant qu'utilisateur racine (root), comme par exemple `/usr/local`, si vous installez le logiciel pour tous les utilisateurs, ou être installé dans votre répertoire d'accueil.
6. Dézippez et extrayez le fichier d'installation à l'aide de la commande : `tar -xzof <fichier-installation-wps>.tar.gz` ou : `gzip -cd <fichier-installation-wps>.tar.gz | tar xvf -`
7. Une clé de licence est nécessaire pour utiliser WPS. Elle peut être appliquée soit depuis l'interface graphique, soit en ligne de commande en lançant l'application d'une des manières suivantes.
   a. Pour lancer l'interface graphique WPS Workbench, passez la commande suivante : `<wps-@product-version-full-short@-installation-dir>/eclipse/workbench`. Le système ouvre une boîte de dialogue permettant d'importer votre clé de licence.
   b. Pour lancer WPS en ligne de commande, passez la commande suivante : `<wps-@product-version-full-short@-installation-dir>/bin/wps -stdio -setinit < <fichier-clé-wps>`. Un message confirme que la clé de licence a bien été appliquée.
8. La configuration de WPS est terminée.

Configurer Hadoop sur Linux x64

Installer Hadoop

Si ce n'est pas encore fait, installez Hadoop en vous référent à la documentation fournie pour votre distribution (Cloudera, par exemple). Une fois Hadoop installé, vous pouvez configurer les détails décrits ci-dessous.

Remarque :

Si vous avez une distribution qui fonctionne conformément au standard Apache Hadoop, les détails de configuration devraient être applicables, même si vous n'utilisez pas une distribution Cloudera. Les distribution qui désactivent ou modifient les fonctionnalités du standard Apache Hadoop ne sont pas prises en charge.

Fichiers de configuration

Tous les appels à Hadoop Cloudera 5 sont réalisés via Java et JNI. Vous devez vous procurer les fichiers `.jar` du client Hadoop Cloudera et les télécharger sur la machine locale. Les fichiers suivants contiennent des URL pour les divers services Hadoop, et doivent être configurés de sorte à correspondre à l'installation actuelle de Hadoop :

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

Veuillez consulter la documentation de Hadoop pour pour plus d'informations.

Variable d'environnement CLASSPATH

La variable d'environnement `CLASSPATH` doit être configurée de sorte à pointer vers les fichiers du client Java Cloudera. Par exemple, il serait possible d'ajouter les lignes suivantes au profil utilisateur (tel que le fichier `.bash_profile`) :

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

```
EXPORT CLASSPATH
```

Configurer Kerberos sur Linux x64

Si votre distribution de Hadoop inclut ou nécessite la prise en charge de Kerberos, allez à la section *Configurer Kerberos et Hadoop côté client* (p. 16).

Configurer Kerberos et Hadoop côté client

Sur Windows comme sur Linux, il peut être nécessaire d'exécuter d'abord la commande `kinit` et de saisir votre mot de passe à l'invite. Il peut s'agir de l'implémentation OS de `kinit` (sur Linux) ou du binaire `kinit` présent dans le répertoire JRE de WPS.

Sur Windows :

- Vous devez être connecté en tant qu'utilisateur Active Directory, pas en tant qu'utilisateur de la machine locale.
- Votre utilisateur ne peut pas être un administrateur local de la machine.
- Vous devez définir une clé de registre afin que Windows permette à Java l'accès à la clé de session TGT :

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

- Les fichiers de JCE (Java Cryptography Extension) Unlimited Strength Jurisdicton Policy doivent être installés dans votre JRE (le JRE inclus dans le répertoire d'installation de WPS).

Vous devez ensuite définir les divers principaux Kerberos dans les fichiers de configuration XML Hadoop. Avec Cloudera, ces composants sont disponibles via Cloudera Manager. La liste des fichiers de configuration inclut :

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

Remarque :

La liste ci-dessus n'est pas exhaustive, et peut être très spécifique au guide : les déclarations `libname` nécessitent en outre que le paramètre `hive_principal` soit défini sur le `hive_principal` du cluster Kerberos.

Exemples de code concernant l'intégration

Connexion à Hive à l'aide de syntaxe SQL standard

```
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;
```

Connexion à Impala à l'aide de syntaxe SQL standard

```
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;
```

Connexion à Hive à l'aide de syntaxe SQL à transfert direct

```
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;
```

Connexion à Impala à l'aide de syntaxe SQL à transfert direct

```
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;
```

Exécution de commandes HDFS et de programmes Pig via WPS

Exemple de code 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;
```

Exemple de code 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';
```

Utiliser WPS avec Hadoop Streaming

Hadoop Streaming est un utilitaire fourni avec la distribution de Hadoop. Il permet de créer et d'exécuter des jobs MapReduce avec tout exécutable ou programme en tant que mappeur ou réducteur.

La syntaxe générale est la suivante :

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

Les mappeurs et réducteurs reçoivent leurs entrées et sorties sur `stdin` et `stdout`. La vue des données est orientée ligne, et chaque ligne est traitée comme une paire clé-valeur séparée par une tabulation.

Vous pouvez utiliser Hadoop Streaming pour tirer parti de la puissance de WPS afin de distribuer des programmes écrits en langage SAS sur de nombreux ordinateurs sur un cluster Hadoop, comme dans l'exemple de job MapReduce donné ci-dessous.

Remarque :

En raison de la large diffusion des programmes, tout exemple utilise nécessairement une approche non traditionnelle du langage SAS, car chaque mappeur et chaque réducteur ne voit qu'une partie limitée des données.

Avant de continuer, familiarisez-vous avec les concepts de HDFS et de MapReduce en lisant la section *Architecture Hadoop* (p. 8).

L'exemple suivant montre la création et l'exécution d'un job MapReduce pour produire le décompte des mots figurant dans les fichiers texte du répertoire fourni comme entrée du job. Chaque instance de mot est traitée comme la paire clé-valeur `<mot><tabulation>1`.

1. Vérifiez que le répertoire `input` a bien été défini sur le HDFS, en utilisant par exemple `hadoop fs -mkdir /user/rw/input`, et que les fichiers texte contenant les mots à compter ont été ajoutés au répertoire. Chaque cluster peut voir ce répertoire.
2. Vérifiez que WPS a été installé au même emplacement sur chaque nœud du cluster, afin que chacun des mappeurs et réducteurs puisse l'invoquer.

3. Créez un programme mappeur appelé **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. Créez un script appelé **map.sh** pour appeler **map.sas** :

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

5. Créez un programme réducteur appelé **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. Créez un script appelé **reduce.sh** pour appeler **reduce.sas** :

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

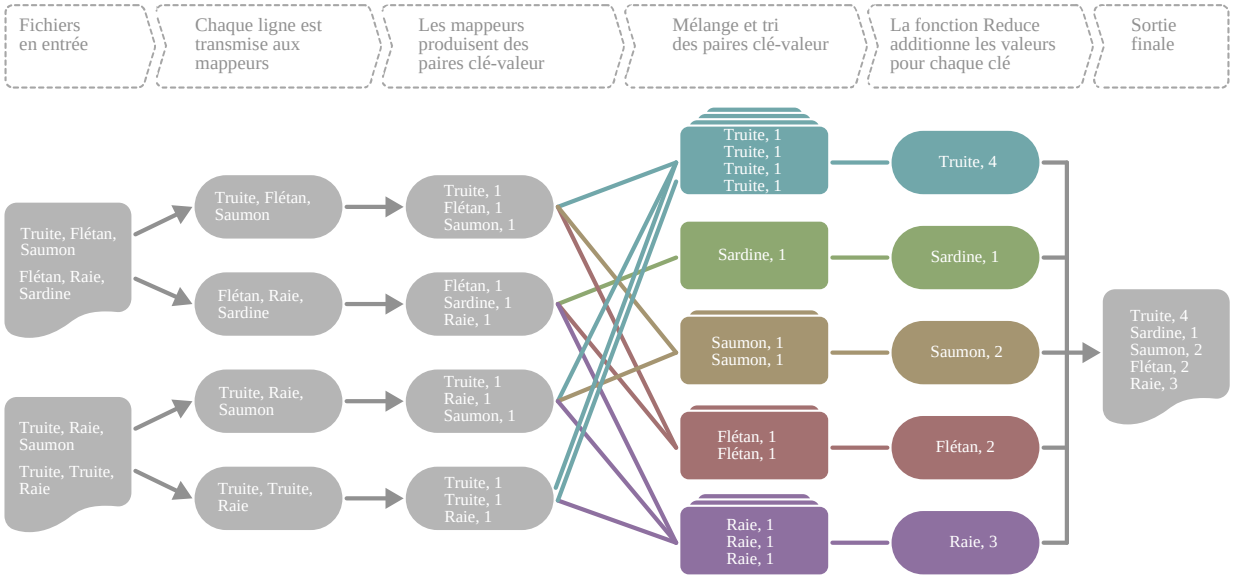
7. Vérifiez que **map.sh**, **map.sas**, **reduce.sh** et **reduce.sas** ont été placés au même emplacement sur chaque nœud du cluster, afin que chacun des mappeurs et réducteurs puisse fonctionner lorsqu'il est invoqué.

8. Vérifiez que la variable d'environnement `CLASSPATH` a été configurée sur la machine client pour votre système d'exploitation, comme indiqué à la section *Configurer Hadoop sur Windows x64* (p. 12) ou *Configurer Hadoop sur Linux x64* (p. 15).

9. Exécutez la ligne de commande suivante depuis une machine client où un client Hadoop est installé, en adaptant au besoin les numéros de version :

```
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"
```

L'exécution de la commande a pour effet de lancer le job MapReduce sur le cluster en question. Chaque instance de mappeur (script **map.sh** sur un nœud, invoquant **map.sas**) produit un ensemble de paires clé-valeur consistant chacune d'un mot et du nombre 1. La phase de mélange a ensuite lieu : les paires clé-valeur ayant la même clé sont envoyées au même réducteur. Chaque instance de réducteur (script **reduce.sh** sur un nœud, invoquant **reduce.sas**) fait la somme du nombre d'occurrences pour sa clé et la place dans un fichier de sortie. La sortie résultante est une série de mots avec les nombres associés. Le processus peut être représenté comme suit :



Remarque :

La sortie finale peut être divisée en plusieurs fichiers dans le répertoire de sortie, selon la configuration du cluster.

Référence

Les diagrammes syntaxiques les descriptions sont des notations permettant d'expliquer la syntaxe de langages de programmation. Ils sont tous deux utilisés dans ce manuel.

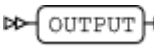
Comment lire les diagrammes syntaxiques

Les diagrammes syntaxiques sont une notation syntaxique graphique représentant des structures importantes du langage telles que les procédures, les instructions et autres.

La description de chaque concept du langage commence par un diagramme syntaxique.

Saisie de texte

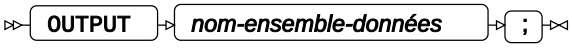
Le texte qui doit être saisi dans un programme tel qu'il est indiqué est représenté dans une police à chasse fixe :

Cet exemple décrit un fragment de syntaxe dans lequel le mot-clé `OUTPUT` est suivi d'un point-virgule : `;`. Le diagramme syntaxique a l'aspect suivant : 
.

En règle générale, la casse du texte n'est pas prise en compte. Dans le présent document, par convention, nous utilisons les majuscules pour les mots-clés.

Espaces réservés

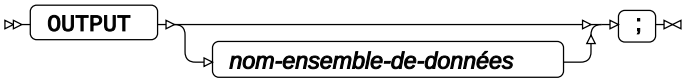
Les espaces réservés sont du texte qui doit être remplacé par une chaîne appropriée. Ils sont en minuscules et en italique :



Ici, le mot-clé `OUTPUT` doit être saisi tel quel, mais *nom-ensemble-de-données* doit être remplacé par une chaîne appropriée – dans ce cas, le nom d'un ensemble de données où le programme va ajouter une observation.

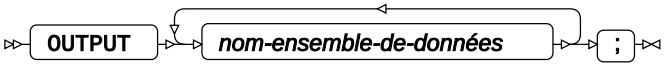
Éléments facultatifs

Lorsque des éléments sont facultatifs, ils apparaissent sur une branche située sous la ligne principale dans les diagrammes syntaxiques. Les éléments facultatifs sont signalés par le fait qu'ils ont un trajet alternatif sans obstacle traversant le diagramme :



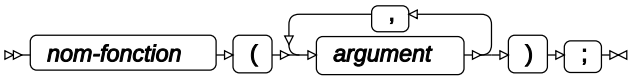
Répétition

Dans les diagrammes syntaxiques, la répétition est décrite par une boucle de retour qui spécifie de manière facultative le séparateur qui doit être placé entre les instances.



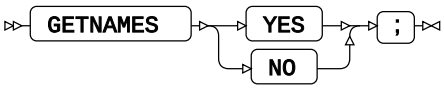
Ci-dessus, le mot-clé `OUTPUT` doit être saisi tel quel, suivi d'une ou plusieurs instances de *nom-ensemble-de-données* – dans ce cas, aucun séparateur autre qu'un espace n'est requis.

L'exemple ci-dessous montre l'utilisation d'un séparateur.



Choix

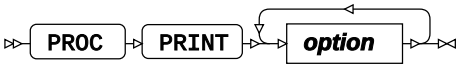
Dans les diagrammes syntaxiques, le choix est représenté par des branches parallèles.



Dans l'exemple ci-dessus, le mot-clé `GETNAMES` doit être saisi tel quel, suivi soit par le mot-clé `YES`, soit par le mot-clé `NO`.

Fragments

Lorsque la syntaxe est trop complexe pour être décrite dans une seule définition, elle peut être divisée en fragments :



option

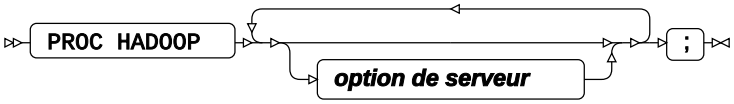
Ci-dessus, la syntaxe est divisée en fragments de diagramme syntaxique distincts. Le premier indique que `PROC PRINT` doit être suivi d'une ou plusieurs instances d'une *option*, chacune devant respecter la syntaxe indiquée dans le deuxième diagramme.

Procédure HADOOP

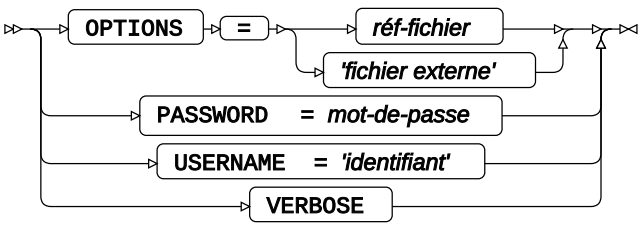
Instructions prises en charge

PROC HADOOP

Accède à Hadoop via WPS.

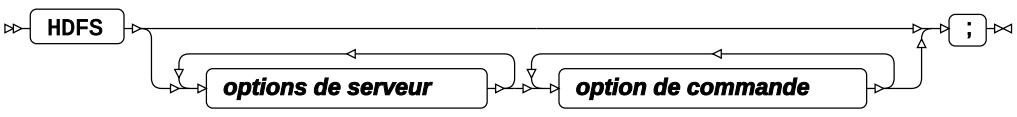


option de serveur

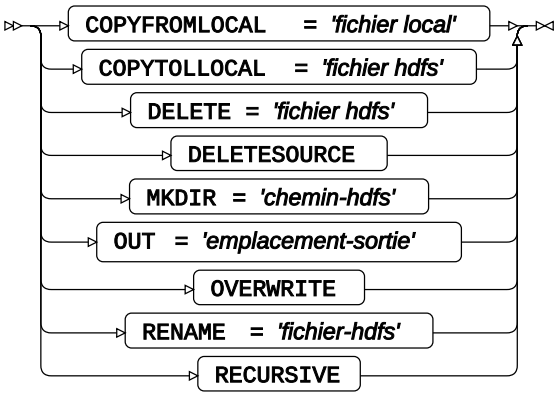


HDFS

Spécifie le système de fichiers distribué Hadoop à utiliser.

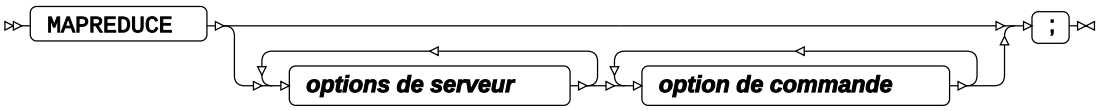


option de commande

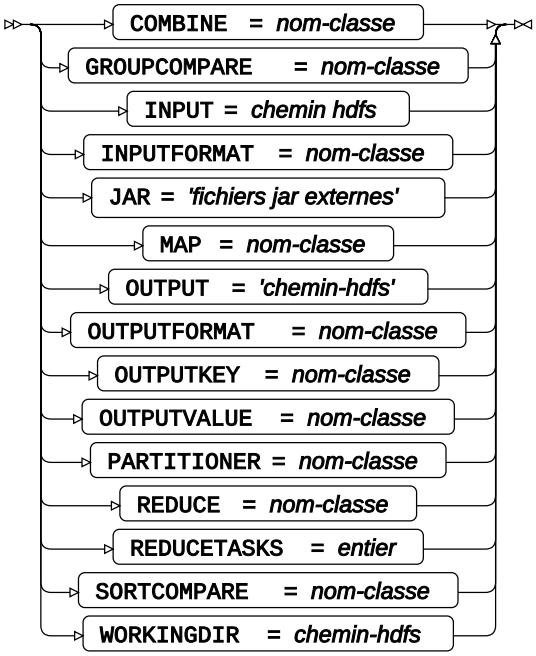


MAPREDUCE

Lance les tâches MapReduce.

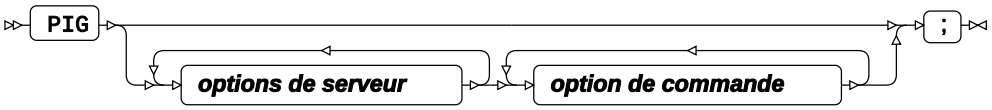


option de commande

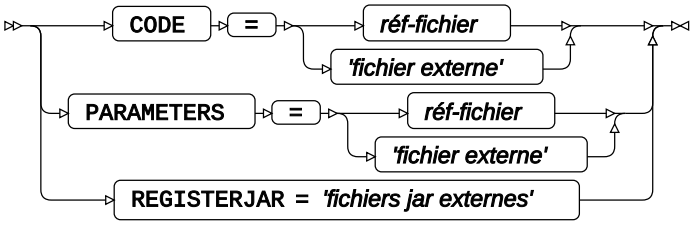


PIG

Active les fichiers externes à soumettre à une grappe (cluster).



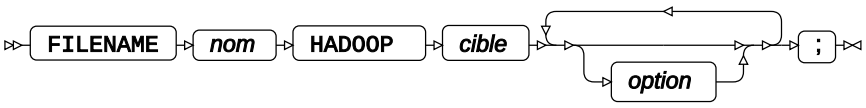
option de commande



Instructions globales

FILENAME, méthode d'accès HADOOP

Lit et écrit des données dans un système de fichiers Hadoop.



Cette méthode d'accès permet à un programme en langage SAS d'accéder aux fichiers d'un système de fichiers Hadoop en utilisant le nom défini dans *nom*. Le nom défini est présenté comme une *référence de fichier* (*fileref*). Vous pouvez utiliser la référence de fichier spécifiée dans un programme afin de lire et d'écrire des données, à condition que le programme soit exécuté au cours de la session WPS Analytics où la référence de nom de fichier a été spécifiée. Une référence de nom de fichier reste active pour la durée d'une session WPS Analytics, sauf si la référence de nom de fichier est effacée par l'instruction `FILENAME CLEAR`.

nom

Nom à utiliser comme référence de nom de fichier.

HADOOP

Spécifie que cette référence de nom de fichier est utilisée pour accéder à un fichier dans un système de fichiers Hadoop.

cible

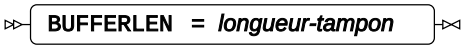
Nom de fichier ou répertoire du fichier auquel accéder.

option

Option qui modifie le comportement de l'instruction.

BUFFERLEN

Spécifie la longueur du tampon utilisé pour les données.



longueur-tampon

Longueur du tampon en octets.

La valeur peut être spécifiée comme :

- Nombre d'octets ; par exemple, vous pouvez saisir `21000`.
- Nombre de kibioctets ou mébioctets, en ajoutant `K` ou `M` après la valeur. Par exemple, si la valeur est `0.3K`, la taille du tampon est 0.3 KiO.

CFG

Spécifie le fichier de configuration Hadoop à utiliser

fichier-ou-fileref

Nom du fichier ou référence de nom de fichier.

CONCAT

Spécifie que si le programme lit plusieurs fichiers, ils sont concaténés.

Si la cible est un répertoire, ou un nom de fichier caractères génériques, tous les fichiers du répertoire ou tous les fichiers sélectionnés sont lus.

DIR

Spécifie que la cible spécifiée dans *cible* est un répertoire.

En spécifiant que la *cible* est un répertoire, il est possible d'utiliser la syntaxe de nom de membre avec `FILE` et `INFILE`.

Par exemple, si vous spécifiez :

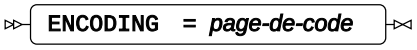
```
FILENAME inhd HADOOP "/temp" USER=exusenm PASSWORD=xxx11xxx DIR;
```

alors il est présumé que `temp` est un répertoire, et vous pouvez spécifier un membre (c'est-à-dire un fichier) de ce répertoire ; par exemple :

```
INFILE inhd (exf.txt);
```

ENCODING

Spécifie la page de code à utiliser pour le contenu d'un fichier.



page-de-code

Une des valeurs de page de code décrites à la section *Encoding values* du document *WPS Reference for Language Elements*.

FILEEXT

Spécifie que, si un nom de fichier spécifié en tant que membre ne comporte pas d'extension, une extension est ajoutée.

Cette option n'est efficace que si `DIR` a été spécifié. Si c'est le cas, et que le nom du membre spécifié pour l'instruction `FILENAME` n'a pas d'extension, l'extension `.data` est ajoutée. Par conséquent, si votre programme crée un fichier dont le nom n'a pas d'extension, l'extension `.data` est ajoutée au nom enregistré. Si votre programme lit un fichier spécifié qui n'a pas d'extension, il est présumé qu'un fichier du même nom avec l'extension `.data` existe ; si ce n'est pas le cas, le programme signale une erreur.

LRECL

Spécifie la longueur des enregistrements.

n

Spécifie la longueur des enregistrements en octets.

La valeur peut être spécifiée comme :

- Nombre d'octets ; par exemple, vous pouvez saisir `21000`.
- Nombre de kibioctets ou mébioctets, en ajoutant `K` ou `M` après la valeur. Par exemple, si la valeur est `0.3K`, la taille du tampon est 0.3 KiO.

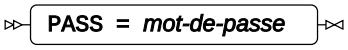
La valeur par défaut est 256.

La longueur d'enregistrement par défaut est la valeur spécifiée par l'option système `LRECL`.

Si un enregistrement lu ou écrit dans un fichier dépasse la longueur spécifiée par cette option, il est tronqué.

PASS

Spécifie le mot de passe d'accès au système de fichiers Hadoop, si nécessaire.

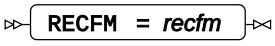


mot-de-passe

Mot de passe correspondant au nom d'utilisateur spécifié par `USER`.

RECFM

Spécifie le format de l'enregistrement.



recfm

Format de l'enregistrement.

F

Longueur fixe sans bloc. Un bloc contient un enregistrement.

N

Format de flux. L'intégralité du fichier est un ensemble continu d'octets, sans remplissage ou interruption des enregistrements. Lit comme `RECFM=F` et écrit comme `RECFM=S`.

S

Format de flux. Similaire à `N`, mais lit et écrit les données telles quelles.

V

Longueur variable sans bloc. Un bloc contient un enregistrement.

Il s'agit de la valeur par défaut.

USER

Spécifie le nom d'utilisateur pour l'accès au système de fichiers Hadoop, si nécessaire.

utilisateur

Nom d'utilisateur requis.

Vous pouvez spécifier le mot de passe correspondant à l'aide de `PASS`.

Les informations concernant la connexion sont écrites dans le journal. Cela inclut le nom du serveur, l'adresse IP, le nom de chemin, etc. Vous pouvez générer plus d'informations sur la connexion en utilisant l'option `DEBUG`.

Exemple simple

Dans cet exemple, des données sont écrites dans le fichier `exfile` sur le système de fichiers Hadoop. Le nom d'utilisateur et le mot de passe sont spécifiés, et le format d'enregistrement et la longueur des enregistrements logiques sont définies.

```
FILENAME outhd HADOOP "/temp/books/exfile" USER=exusenm PASSWORD=xxx11xxx;
DATA _NULL_;

  SET books.books;
  FILE outhd;

  WHERE author EQ "Asimov, Isaac";

  PUT author title;

RUN;
```

Exemple – Lire un fichier et spécifier que les noms de fichier ont l'extension .dat

Dans cet exemple, des données sont écrites dans le fichier `exfile` sur le système de fichiers Hadoop. Le nom d'utilisateur et le mot de passe sont spécifiés, et le format d'enregistrement et la longueur des enregistrements logiques sont définies. Le nom de fichier `exfile2` n'a pas d'extension, donc l'option `FILEEXT` a été spécifiée pour que seul le fichier `exfile2.dat` soit lu.

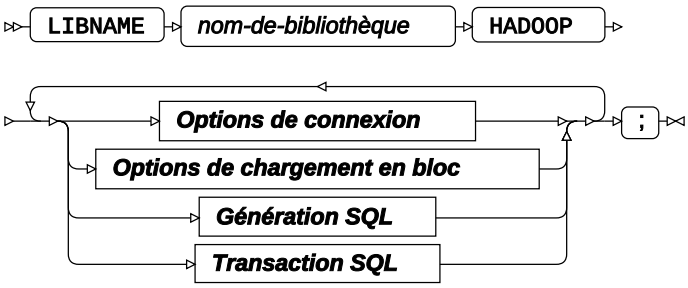
```
FILENAME inhd HADOOP '/temp/books/exfile2' USER=exusenm PASSWORD=xxx111xxx FILEEXT
 RECFM=F LRECL=8192;
DATA out;

  FILE inhd dlm='#';

  INPUT a $ b $ ;

RUN;
```

Moteur WPS pour Hadoop

Le moteur WPS pour Hadoop permet la connectivité entre WPS et une base de données Hadoop.

HADOOP

Cett option permet de connecter WPS à une base de données en spécifiant le moteur de base de données à l'aide de l'instruction de connexion à une bibliothèque `LIBNAME`.



L'instruction `LIBNAME` permet à un programme en langage SAS d'accéder à une base de données en utilisant le nom défini dans *nom-de-bibliothèque*. Une référence de bibliothèque ne reste active que pour la durée d'une session WPS Analytics. Vous pouvez utiliser la référence de bibliothèque spécifiée dans des programmes afin d'accéder aux données stockées dans une base de données, à condition que les programmes soient exécutés au cours de la session où la référence de bibliothèque a été spécifiée. L'instruction `LIBNAME` contient des options qui, lorsqu'elles sont spécifiées, déterminent comment les programmes en langage SAS interagissent avec la base de données. Ces options sont regroupées comme suit :

- *Options de connexion* (p. 34) : Établissent la connexion avec le serveur de base de données.
- *Options de chargement en bloc* (p. 38) : Insèrent rapidement de grandes quantités de données dans une base de données à l'aide du mécanisme de chargement en bloc (bulk insert).
- *Options de génération SQL* (p. 39) : Déterminent comment les informations de description de table ou les instructions de requête sont formatées et utilisées.
- *Options de transaction SQL* (p. 42) : Affectent le comportement transactionnel des instructions traitées par WPS Analytics et le serveur de base de données.

nom-de-bibliothèque

Spécifie le nom utilisé dans d'autres instructions en langage SAS pour faire référence à cette connexion à la base de données.

Par exemple, l'instruction suivante :

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

crée une connexion avec une base de données en utilisant le nom `ExLib`. Vous pouvez alors utiliser ce nom, par exemple, dans la procédure SQL :

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

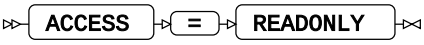
Dans ce programme, la procédure SQL permet d'insérer des données dans la table de base de données `person` dans la base de données référencée par `ExLib`.

Options de connexion

Établissent la connexion avec le serveur de base de données.

ACCESS

Spécifie le mode d'accès de la connexion à la bibliothèque.



READONLY

La connexion à la bibliothèque ne peut être utilisée que pour lire des données.

Si vous spécifiez cette option, elle remplace les paramètres d'insertion ou de mise à jour d'autres options, et les données ne sont alors pas modifiées.

Si cette option n'est pas spécifiée, la connexion à la bibliothèque utilise un mode d'accès *lecture-écriture* qui permet les opérations de lecture, d'insertion et de mise à jour.

AUTHDOMAIN

Spécifie le domaine d'authentification WPS Hub.

Type : Chaîne de caractères

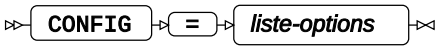
Le domaine d'authentification accorde les autorisations d'accès à un serveur de base de données. WPS Analytics utilise WPS Hub comme domaine d'authentification, et votre système doit avoir accès à un serveur WPS Hub.

Dans cet exemple, le programme fournit les autorisations d'accès à WPS Hub sous forme d'options système, et le nom du domaine d'authentification contenant les détails d'authentification dans WPS Hub est spécifié par `AUTHDOMAIN`.

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

CONFIG

Spécifie les options de configuration.

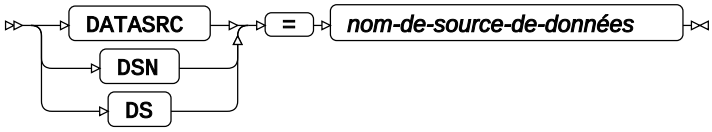


Type : Chaîne de caractères

Les options sont séparées par des espaces. La liste doit être entourée de guillemets.

DATASRC

Spécifie un nom de source de données (DSN).



Type : Chaîne de caractères

Vous ne pouvez spécifier un DSN dans *nom-de-source-de-données* que s'il a déjà été créé.

DBCONINIT

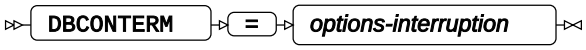
Spécifie les instructions SQL ou les commandes de base de données qui sont traitées chaque fois que la connexion à la bibliothèque est établie.

Type : Chaîne de caractères

La connexion est établie lors de l'exécution de l'instruction `LIBNAME`, puis à chaque fois qu'une connexion est établie avec la base de données ; par exemple, lors de l'exécution d'une instruction SQL dans la procédure SQL.

DBCONTERM

Spécifie les instructions SQL ou les commandes de base de données qui sont traitées chaque fois que la connexion à la bibliothèque est interrompue.



Type : Chaîne de caractères

La connexion à la bibliothèque est interrompue à la fin d'une procédure, d'une étape DATA, et d'une session.

DBLIBINIT

Spécifie les instructions SQL ou les commandes de base de données qui sont traitées immédiatement lors de la création de la connexion à la bibliothèque à l'aide de l'instruction `LIBNAME`.

Type : Chaîne de caractères

DBLIBTERM

Spécifie les instructions SQL ou les commandes de base de données qui sont traitées juste avant l'interruption de la connexion à la bibliothèque à la fin de la session WPS Analytics.

Type : Chaîne de caractères

HIVE_PRINCIPAL

Spécifie le serveur principal d'une ruche Hive.

Type : Chaîne de caractères

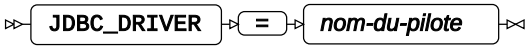
JDBC_CONNECTION_STRING

Spécifie la chaîne de connexion pour JDBC.

Type : Chaîne de caractères

JDBC_DRIVER

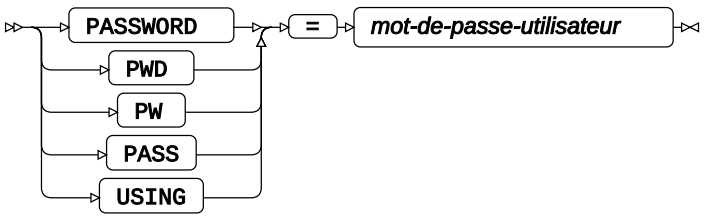
Spécifie le pilote JDBC à utiliser pour la connexion à la ruche Hive.



Type : Chaîne de caractères

PASSWORD

Spécifie le mot de passe correspondant au nom d'utilisateur.



Type : Chaîne de caractères

Si vous utilisez des caractères spéciaux dans la chaîne, vous devez saisir *mot-de-passe-utilisateur* entre guillemets. Le nom d'utilisateur est spécifié par l'option `USER`.

Si vous avez spécifié `AUTHDOMAIN` pour activer l'authentification via WPS Hub, il n'est pas nécessaire de spécifier le nom d'utilisateur et le mot de passe.

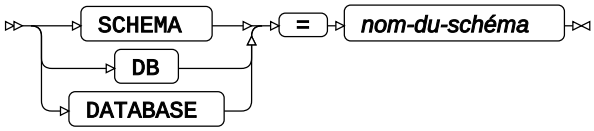
PORT

Spécifie numéro de port du serveur de base de données spécifié par l'option `SERVER`.

Type : Numérique

SCHEMA

Spécifie le nom du schéma de base de données avec lequel la connexion interagit.

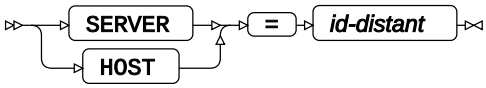


Type : Chaîne de caractères

Un schéma est un groupe d'objets de base de données et de données auquel l'utilisateur a accès et qu'il peut manipuler à l'aide d'instructions SQL.

SERVER

Spécifie le nom ou l'adresse TCP/IP de l'ordinateur hébergeant le serveur de base de données.

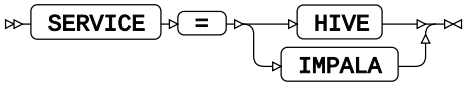


Type : Chaîne de caractères

Si aucun serveur de base de données externe n'est spécifié, il est présumé que le serveur de base de données et les données sont sur le dispositif où WPS Analytics est installé et s'exécute.

SERVICE

Spécifie le type de service Hadoop.



HIVE

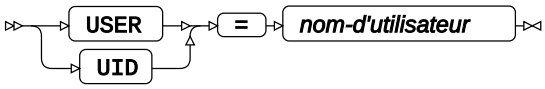
Spécifie que le service est Hive.

IMPALA

Spécifie que le service est Impala.

USER

Spécifie le nom d'utilisateur requis pour accéder à la base de données.



Type : Chaîne de caractères

Si vous utilisez des caractères spéciaux dans la chaîne, vous devez saisir *nom-d'utilisateur* entre guillemets.

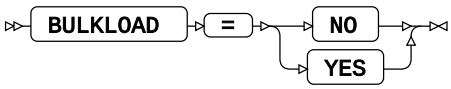
Si vous avez spécifié `AUTHDOMAIN` pour activer l'authentification via WPS Hub, il n'est pas nécessaire de spécifier le nom d'utilisateur et le mot de passe.

Options de chargement en bloc

Insèrent rapidement de grandes quantités de données dans une base de données à l'aide du mécanisme de chargement en bloc (bulk insert).

BULKLOAD

Spécifie si les données sont insérées dans une table en utilisant la fonctionnalité de chargement en bloc.



NO

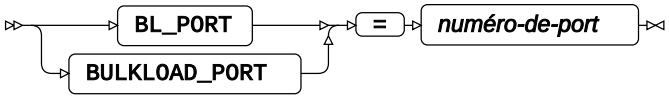
L'insertion de données n'utilise pas la fonctionnalité de chargement en bloc. Toute autre option de chargement en bloc spécifiées pour l'instruction n'est pas prise en compte.

YES

L'insertion de données utilise la fonctionnalité de chargement en bloc.

BL_PORT

Spécifie le port de serveur à utiliser pour le chargement en bloc des données.



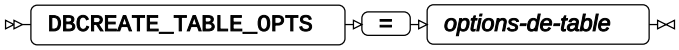
Type : Numérique

Génération SQL

Affecte la manière dont les instructions SQL sont créées, et si elles sont créées par WPS Analytics ou par le serveur de base de données.

DBCREATE_TABLE_OPTS

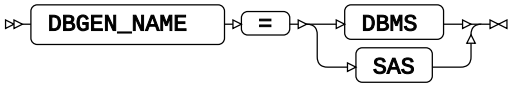
Spécifie des options de table supplémentaires à ajouter à une instruction SQL `CREATE TABLE` après la définition des colonnes de la table.



Type : Chaîne de caractères

DBGEN_NAME

Spécifie comment modifier les noms de colonne non valides pour qu'ils soient des noms de variable valides en langage SAS.



Valeur par défaut : `SAS`

DBMS

Si le nom de colonne contient des caractères non valides, ceux-ci sont remplacés par un tiret de soulignement. Si cela crée un nom qui entre en conflit avec le nom d'une autre colonne, le numéro de la colonne est ajouté au nom de la colonne.

Par exemple, si votre base de données contient les colonnes `id$x`, `id#x` et `id_x`, les variables de l'ensemble de données en langage SAS seraient respectivement `ID_X`, `ID_X0` et `ID_X1`.

Remarque :

La numérotation commence dès la première répétition du nom de variable et débute par `0` (zéro).

SAS

Si le nom de la colonne comporte des caractères non valides, ou est identique au nom d'une autre colonne de la table, le nom est remplacé par la chaîne `_COL`. Si cela crée un nom qui entre en conflit avec une autre colonne, le numéro de la colonne est ajouté au nom : `_COL`*x*, où *x* est le numéro, en débutant par 0 (zéro).

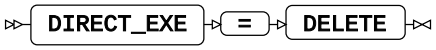
Par exemple, si votre table contient les colonnes `id$x`, `id#x` et `id_x`, l'ensemble de données aurait respectivement les variables `COL0`, `COL1` et `ID_X0`.

Remarque :

La numérotation commence dès la première instance du nom créé et débute par `0` (zéro).

DIRECT_EXE

Spécifie si les instructions de suppression sont traitées par le moteur de base de données ou par WPS Analytics.



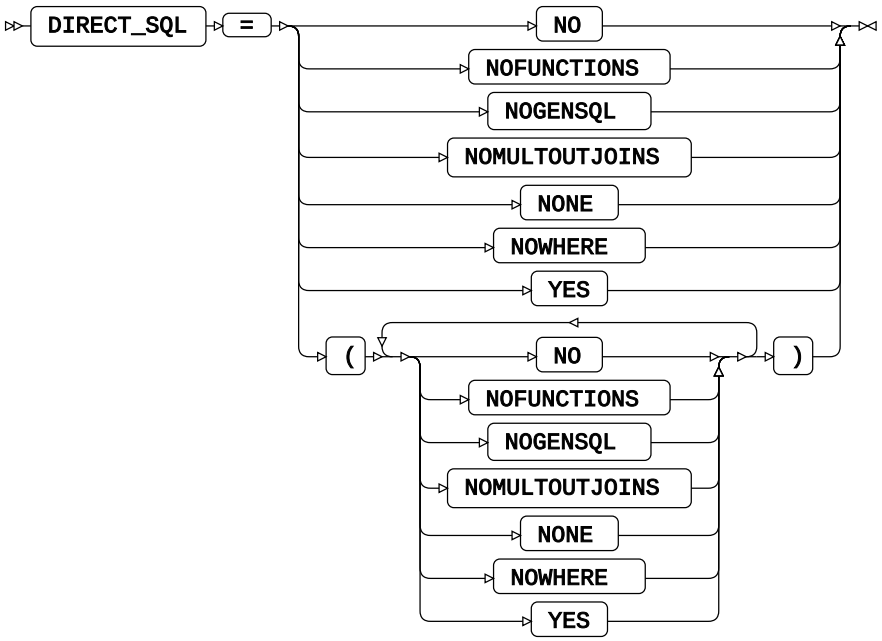
Si ce n'est pas spécifié, la table de base de données est analysée, et les lignes correspondant aux critères de suppression sont renvoyés à WPS Analytics. Une instruction SQL `DELETE FROM` est créée par WPS Analytics pour chaque ligne affectée et transmise à la base de données pour traitement.

DELETE

Spécifie qu'une commande SQL `DELETE FROM` est transmise à la base de données et traitée entièrement par le moteur de base de données.

DIRECT_SQL

Spécifie si les instructions SQL sont transmises pour être traitées par le moteur de base de données ou traitées par WPS Analytics.



Valeur par défaut : `YES`

NO

Toutes les instructions SQL sont traitées par WPS Analytics.

NOFUNCTIONS

Toutes les instructions SQL contenant des appels de fonction sont traitées par WPS Analytics.. Toute autre instruction pouvant être traitée par le moteur de base de données est envoyée au serveur de base de données pour traitement. Si vous spécifiez `DIRECT_SQL = NOFUNCTIONS`, vous remplacez l'option `SQL_FUNCTIONS`.

NOGENSQL

Toutes les instructions SQL sont traitées par WPS Analytics..

NOMULTOUTJOINS

Toutes les instructions SQL contenant plusieurs jointures externes sont traitées par WPS Analytics.. Toute autre instruction pouvant être traitée par le moteur de base de données est envoyée au serveur de base de données pour traitement.

NONE

Toutes les instructions SQL sont traitées par WPS Analytics.

NOWHERE

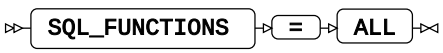
Toutes les instructions SQL contenant des clauses `WHERE` sont traitées par WPS Analytics.. Toute autre instruction pouvant être traitée par le moteur de base de données est envoyée au serveur de base de données pour traitement.

YES

Toute instruction SQL pouvant être traitée par le serveur de base de données est envoyée à ce dernier.

SQL_FUNCTIONS

Spécifie quelles fonctions du langage SAS sont traitées par le moteur de base de données.



Si cette option n'est pas spécifiée, les fonctions suivantes sont transmises au moteur de base de données pour traitement lorsqu'elles sont utilisées dans des instructions du langage SAS telles que la procédure SQL, ou l'option d'ensemble de données `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()

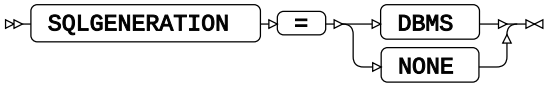
ALL

Spécifie que, en plus de l'ensemble par défaut de fonctions du langage SAS, les options suivantes sont également transmises au moteur de base de données pour traitement :

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

SQLGENERATION

Specifies whether the SQL statements and data are processed by the database server, or by WPS Analytics.



DBMS

Les instructions sont générées et les données sont traitées sur le serveur de base de données.

NONE

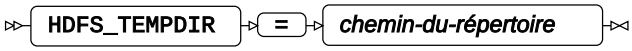
Les instructions sont générées et les données sont traitées par WPS Analytics. Les résultats sont alors envoyés au serveur de base de données.

Transaction SQL

Affectent le comportement transactionnel des instructions traitées par WPS Analytics et le serveur de base de données.

HDFS_TEMPDIR

Spécifie l'emplacement utilisé pour les répertoires HDFS temporaires.

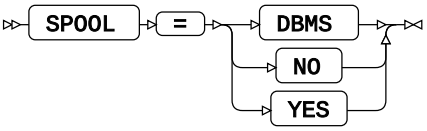


Type : Chaîne de caractères

Par défaut, le paramètre de configuration de Hadoop `hadoop.tmp.dir` spécifie l'emplacement des fichiers temporaires locaux et HDFS. Vous pouvez utiliser ce paramètre pour spécifier un autre répertoire pour les fichiers temporaires HDFS.

SPOOL

Spécifie si un fichier de spoule est créé.



Un fichier de spoule permet de faire en sorte que toutes les opérations effectuées lors de l'exécution d'une requête SQL soient identiques. L'utilisation du spoule a toujours lieu si nécessaire. Elle est déterminée par le planificateur de requêtes.

DBMS

Le spouling est géré par le moteur de base de données.

NO

Fourni pour améliorer la compatibilité. Toutefois, comme le spouling a toujours lieu si nécessaire, si cette option est définie, le spouling est géré par le moteur de base de données.

YES

Le spouling est géré par WPS Analytics. Il s'agit de la valeur par défaut.

Notices légales

Marques

WPS et World Programming sont des marques commerciales ou des marques déposées de World Programming Limited dans l'Union européenne et dans d'autres pays. Le sigle (r) ou ® indique l'enregistrement au niveau de l'Union européenne (« marque communautaire »).

SAS et tous les autres noms de produits et de services de SAS Institute Inc. sont des marques déposées ou des marques commerciales de SAS Institute Inc. aux États-Unis et dans d'autres pays. ® indique que la marque est déposée aux États-Unis.

Toutes les autres marques commerciales sont la propriété de leurs détenteurs respectifs.

Notices générales

World Programming Limited n'est associé d'aucune manière à SAS Institute Inc.

WPS n'est pas le système SAS.

Les expressions « SAS » et « langage SAS » utilisées dans ce document font référence au langage de programmation SAS qui est souvent désigné par ces termes.

Les expressions « programme », « programme SAS » et « programme en langage SAS » utilisées dans ce document font référence aux programmes écrits en langage SAS. Ils peuvent également être appelés « scripts », « scripts SAS » ou « scripts en langage SAS ».

Les expressions « IML » et « langage IML », « syntaxe IML » et « Interactive Matrix Language » utilisées dans ce document font référence au langage de programmation informatique qui est souvent désigné par ces termes.

WPS inclut du logiciel développé par des tiers. Vous trouverez plus d'informations dans le fichier THANKS ou acknowledgements-fr.txt inclus dans l'installation de WPS.