IBM SPSS Analytic Server
Version 3.2.1

Guide d'administration

Important

Avant d'utiliser le présent document et le produit associé, prenez connaissance des informations générales figurant à la section «Remarques», à la page 25.

Certaines illustrations de ce manuel ne sont pas disponibles en français à la date d'édition.

Cette édition s'applique à la version 3.2.1 d'IBM SPSS Analytic Server, et à toutes les éditions et modifications ultérieures sauf mention contraire dans les nouvelles éditions.

LE PRESENT DOCUMENT EST LIVRE EN L'ETAT SANS AUCUNE GARANTIE EXPLICITE OU IMPLICITE. IBM DECLINE NOTAMMENT TOUTE RESPONSABILITE RELATIVE A CES INFORMATIONS EN CAS DE CONTREFACON AINSI QU'EN CAS DE DEFAUT D'APTITUDE A L'EXECUTION D'UN TRAVAIL DONNE.

Ce document est mis à jour périodiquement. Chaque nouvelle édition inclut les mises à jour. Les informations qui y sont fournies sont susceptibles d'être modifiées avant que les produits décrits ne deviennent eux-mêmes disponibles. En outre, il peut contenir des informations ou des références concernant certains produits, logiciels ou services non annoncés dans ce pays. Cela ne signifie cependant pas qu'ils y seront annoncés.

Pour plus de détails, pour toute demande d'ordre technique, ou pour obtenir des exemplaires de documents IBM, référez-vous aux documents d'annonce disponibles dans votre pays, ou adressez-vous à votre partenaire commercial.

Vous pouvez également consulter les serveurs Internet suivants :

- http://www.fr.ibm.com (serveur IBM en France)
- http://www.ibm.com/ca/fr (serveur IBM au Canada)
- http://www.ibm.com (serveur IBM aux Etats-Unis)

Compagnie IBM France
Direction Qualité
17, avenue de l'Europe
92275 Bois-Colombes Cedex

Table des matières

Avis aux lecteurs canadiens

Le présent document a été traduit en France. Voici les principales différences et particularités dont vous devez tenir compte.

Illustrations

Les illustrations sont fournies à titre d'exemple. Certaines peuvent contenir des données propres à la France.

Terminologie

La terminologie des titres IBM peut différer d'un pays à l'autre. Reportez-vous au tableau ci-dessous, au besoin.

IBM France	**IBM Canada**
ingénieur commercial	représentant
agence commerciale	succursale
ingénieur technico-commercial	informaticien
inspecteur	technicien du matériel

Claviers

Les lettres sont disposées différemment : le clavier français est de type AZERTY, et le clavier français-canadien de type QWERTY.

OS/2 et Windows - Paramètres canadiens

Au Canada, on utilise :

- les pages de codes 850 (multilingue) et 863 (français-canadien),
- le code pays 002,
- le code clavier CF.

Nomenclature

Les touches présentées dans le tableau d'équivalence suivant sont libellées différemment selon qu'il s'agit du clavier de la France, du clavier du Canada ou du clavier des États-Unis. Reportez-vous à ce tableau pour faire correspondre les touches françaises figurant dans le présent document aux touches de votre clavier.

France	Canada	Etats-Unis
(Pos1)		Home
Fin	Fin	End
(PgAr)		PgUp
(PgAv)		PgDn
Inser	Inser	Ins
Suppr	Suppr	Del
Echap	Echap	Esc
Attn	Intrp	Break
Impr écran	ImpEc	PrtSc
Verr num	Num	Num Lock
Arrêt défil	Défil	Scroll Lock
(Verr maj)	FixMaj	Caps Lock
AltGr	AltCar	Alt (à droite)

Brevets

Il est possible qu'IBM détienne des brevets ou qu'elle ait déposé des demandes de brevets portant sur certains sujets abordés dans ce document. Le fait qu'IBM vous fournisse le présent document ne signifie pas qu'elle vous accorde un permis d'utilisation de ces brevets. Vous pouvez envoyer, par écrit, vos demandes de renseignements relatives aux permis d'utilisation au directeur général des relations commerciales d'IBM, 3600 Steeles Avenue East, Markham, Ontario, L3R 9Z7.

Assistance téléphonique

Si vous avez besoin d'assistance ou si vous voulez commander du matériel, des logiciels et des publications IBM, contactez IBM direct au 1 800 465-1234.

Chapitre 1. Gestion des titulaires

L'utilisation de titulaires (tenants), qui ne peuvent partager les objets, permet d'instituer une séparation de haut niveau entre les utilisateurs, les projets et les sources de données. Chaque utilisateur accède au système dans le contexte d'un titulaire auquel il est affecté.

Vous pouvez gérer des titulaires et leur affecter des utilisateurs via la console Analytic Server. L'affichage de la page des titulaires dépend du rôle de l'utilisateur connecté à la console :

- L'administrateur "superutilisateur" configuré durant l'installation est le gestionnaire du titulaire. Cet utilisateur est le seul à pouvoir créer de nouveaux titulaires et éditer les propriétés d'un titulaire.
- Les utilisateurs possédant le rôle Administrateur peuvent éditer les propriétés du titulaire auxquels ils sont connectés.
- Les utilisateurs possédant le rôle Utilisateur ne peuvent pas éditer les propriétés d'un titulaire. La page des titulaires n'est pas visible pour eux.
- Les utilisateurs dotés du rôle Reader ne peuvent ni éditer des sources de données, ni se connecter à la console Analytic Server.

Les administrateurs peuvent gérer et nettoyer tous types de projets ou de sources de données en accédant aux pages correspondantes. Voir le manuel *IBM® SPSS Analytic Server - Guide d'utilisation* pour plus d'informations.

Liste des titulaires

Les titulaires déjà présents sont affichés dans un tableau. Seul, l'administrateur enregistré comme "superutilisateur" peut effectuer des modifications sur cette page.

- Cliquez sur le nom d'un titulaire pour afficher ses détails et éditer ses propriétés
- Cliquez sur l'URL d'un titulaire pour ouvrir la console dans le contexte de ce titulaire.

  Remarque : Vous serez déconnecté et devrez vous reconnecter avec des données d'identification valides pour le titulaire
- Dans la zone de recherche, saisissez une entrée permettant de filtrer la liste et afficher uniquement les titulaires dont le nom contient la chaîne de recherche.
- Cliquez sur **New** pour créer un titulaire avec le nom spécifié dans la boîte de dialogue **Add new tenant**. Voir «Règles de désignation», à la page 2 pour vérifier les restrictions sur les noms attribués aux titulaires.
- Cliquez sur **Delete** pour retirer les titulaire(s) sélectionné(s).
- Cliquez sur **Refresh** pour mettre à jour la liste.

Détails des titulaires individuels

La zone de contenu est divisée en sections réductibles.

Details

Name Un champ de texte éditable affichant le nom du titulaire.

Description
Un champ de texte éditable vous permettant d'ajouter un texte explicatif concernant le titulaire.

URL Il s'agit de l'URL à donner aux utilisateurs pour se connecter au titulaire via la console

Analytic Server et à utiliser pour configurer le serveur SPSS Modeler. Voir *IBM SPSS Analytic Server - Guide d'installation et de configuration* pour plus de détails sur la configuration de SPSS Modeler.

Status Les titulaires dont le statut est **actif** sont en cours d'utilisation. Le statut **inactif** empêche les utilisateurs de se connecter à ce titulaire mais n'entraîne aucune suppression des informations sous-jacentes.

Principals

Les principaux sont des utilisateurs et groupes dessinés à partir du fournisseur de sécurité configuré pendant l'installation. Vous pouvez ajouter à un titulaire des principaux ayant la fonction d'administrateur, d'utilisateur ou de lecteur.

- Les entrées saisies dans la zone de texte filtrent les utilisateurs et groupes dont le nom contient la chaîne de recherche. Sélectionnez **Administrator**, **User** ou **Reader** dans la liste déroulante pour affecter leur rôle au sein du titulaire. Cliquez sur **Add participant** pour les ajouter à la liste des auteurs.
- Pour supprimer un participant, sélectionnez un utilisateur ou groupe dans la liste des membres et cliquez sur **Remove participant**.

Metrics

Permet de configurer des limites de ressources pour un titulaire. Fournit l'espace disque utilisé par le titulaire actuellement.

- Vous avez la possibilité de définir un espace disque maximum pour le titulaire. Une fois que celui-ci a atteint ce quota, il est impossible d'écrire des données supplémentaires sur ce disque. Dans ce cas, le titulaire devra effacer certaines données afin d'augmenter l'espace disque.
- Vous avez la possibilité de définir un niveau d'avertissement concernant un espace disque pour le titulaire. Si la limite est dépassée, les principaux ne peuvent pas envoyer de travaux d'analyse via ce titulaire. Dans ce cas, le titulaire devra effacer certaines données afin d'augmenter l'espace disque.
- Vous avez la possibilité de définir un nombre maximum de travaux parallèles pouvant être exécutés via ce titulaire en une seule fois. Si la limite est dépassée, les principaux ne peuvent pas envoyer de travaux d'analyse via ce titulaire, à moins que le travail en cours d'exécution ne soit terminé.
- Vous pouvez définir le nombre maximum de zones autorisées pour une source de données. Cette limite est vérifiée dès lors qu'une source de données est créée ou mise à jour.
- Vous pouvez définir la taille de fichier maximale en mégaoctets. Cette limite est vérifiée lors du chargement d'un fichier.

Security provider configuration

Permet de spécifier une fournisseur d'authentification d'utilisateur. L'option **Default** vous permet d'utiliser le fournisseur de titulaire par défaut, celui configuré au moment de l'installation et de la configuration. L'option **LDAP** vous permet d'authentifier les utilisateurs avec un serveur LDAP externe tels que Active Directory ou OpenLDAP. Indique les paramètres concernant le fournisseur ainsi que les paramètres de filtre (facultatif) pour contrôler les utilisateurs et groupes disponibles dans la section Principals.

Règles de désignation

Pour attribuer un nom unique aux objets d'Analytic Server, tels que les sources de données et projets, vous devez suivre ces règles :

- Au sein d'un titulaire, les noms des objets de même type doivent être uniques. Par exemple, deux sources de données ne peuvent pas être appelées insuranceClaims. En revanche une source de données et un projet peuvent porter ce même nom.
- Les noms sont sensibles à la casse. Par exemple, insuranceClaims et InsuranceClaims représentent des noms uniques.

- Les noms ne reconnaissent pas les interlignages ni les espaces de fin.
- Les caractères suivants se sont pas admis dans les noms.
  `~, #, %, &, *, {, }, \\, :, <, >, ?, /, |, ", \t, \r, \n`

Chapitre 2. Initiation des utilisateurs

Demandez aux utilisateurs d'accéder à `http://<host>:<port>/<context-root>/admin/<tenant>` et d'entrer leur nom d'utilisateur et leur mot de passe de connexion dans la console Analytic Server.

Remarque : Le nom d'utilisateur saisi à l'invite de connexion à la console Analytic Server ne doit pas inclure le suffixe du nom de domaine. Par conséquent, lorsque plusieurs domaines sont définis, les utilisateurs doivent sélectionner le domaine de leur choix dans la liste déroulante **Domaines**. Lorsqu'un seul domaine est défini, la liste déroulante **Domaines** n'apparaît pas lors de la connexion à Analytic Server.

<host>
Adresse de l'hôte Analytic Server.

<port>
Port sur lequel Analytic Server écoute. Par défaut, il s'agit de 9080.

<context-root>
Racine de contexte d'Analytic Server. Par défaut, la valeur est analyticserver.

<tenant>
Dans un environnement à service partagé, le titulaire auquel vous appartenez. Dans un environnement à service exclusif, le titulaire par défaut est **`ibm`**.

Par exemple, si l'adresse IP de la machine hôte est 9.86.44.232, si vous avez créé le titulaire "masociété" en lui ajoutant des utilisateurs, et si les autres paramètres ont conservé leur valeur par défaut, les utilisateurs doivent utiliser l'adresse `http://9.86.44.232:9080/analyticserver/admin/masociété` pour accéder à la console Analytic Server.

Chapitre 3. Noms des travaux Analytic Server

Analytic Server génère des travaux MapReduce et Spark, que vous pouvez contrôler via l'interface utilisateur du gestionnaire de ressources du cluster Hadoop.

Le nom du travail map-reduce est structuré de la façon suivante.

`AS/{nom_titulaire}/{nom_utilisateur}/{nom_algorithme}`

{nom_titulaire}
: Nom du titulaire utilisé pour exécuter ce travail.

{nom_utilisateur}
: Utilisateur à l'origine de la demande de travail.

{nom_algorithme}
: Algorithme principal dans ce travail. Notez que le flux unique peut générer des travaux map-reduce multiples. Sinon, plusieurs opérations d'un même flux peuvent être incluses dans un travail map-reduce.

Tous les travaux MapReduce s'affichent dans l'interface utilisateur du gestionnaire de ressources. Une application Spark est démarrée pour chaque instance Analytic Server. Ouvrez l'interface utilisateur de l'application Spark pour contrôler les travaux Spark (les noms de travaux s'affichent dans la colonne **Description**).

Chapitre 4. Meilleures pratiques et recommandations pour IBM SPSS Analytic Server

Les sections ci-après contiennent les meilleures pratiques et les recommandations à suivre pour Analytic Server en matière de sources de données, de configuration de cluster et de flux IBM SPSS Modeler.

Sources de données

Analytic Server prend en charge les types de source de données suivants :

- Les sources de données basées sur des fichiers, comme des fichiers délimités, à texte fixe et Microsoft Excel.
- Les bases de données relationnelles comme Db2, Oracle, Microsoft SQL Server, Teradata, Postgres, Netezza, MySQL et Amazon Redshift.
- Les sources de données Hive/HCatalog qui incluent tous les types de données intégrés (par exemple ORC et Parquet), ainsi que tout type de données personnalisé pour lequel la mise en oeuvre du sérialiseur-désérialiseur dans Hive est disponible. En outre, vous pouvez configurer Analytic Server pour accéder à des bases de données NoSQL, comme HBase, MongoDB, Accumulo, Cassandra, Oracle NoSQL et d'autres bases de données pour lesquelles une mise en oeuvre appropriée du gestionnaire de stockage Hive est disponible.
- Les sources de données de type géospatial (basées sur des fichiers shapefile et des services cartographiques).

Limitations d'Analytic Server sur des sources de données Hive/HCatalog

- Si la procédure "pushback" dans Hive est nécessaire pour le noeud Sélectionner de SPSS Modeler, l'expression de filtrage ne peut référencer que des colonnes partitionnées de type STRING. A partir d'Analytic Server 3.0, une prise en charge des types de données a été ajoutée pour les colonnes partitionnées suivantes : TINYINT, SMALLINT, INT et BIGINT. L'expression de filtrage statique indiquée pour la source de données Hive peut inclure des expressions de filtrage pour des colonnes partitionnées de n'importe quel type de données.
- Analytic Server ne prend pas en charge les sources de données basées sur des vues Hive.

Configuration du cluster - Sécurité

Emprunt d'identité Kerberos

Avant la version 3.0.1, les instances Analytic Server utilisaient un nom de principal d'utilisateurs dans le fichier de clés Analytic Server pour authentifier les opérations HDFS lorsque la sécurité Kerberos était activée. A partir de la version 3.0.1, Analytic Server utilise un nom de principal de service dans le fichier Analytic Server avec le nom d'utilisateur demandeur (qui fait la demande REST) pour authentifier les opérations HDFS qui utilisent l'emprunt d'identité Kerberos. Analytic Server version 3.0.1 ou suivante est nécessaire pour ajouter les attributs de configuration d'emprunt d'identité à HDFS (ou les configurations de service Hive) lors d'une exécution activée pour utiliser Kerberos. Dans le cas du système de fichiers HDFS, vous devez ajouter les propriétés suivantes au fichier `core-site.xml` HDFS :

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

où `<analytic_server_service_principal_name>` est la valeur par défaut `as_user` spécifiée dans la zone `Analytic_Server_User` de la configuration d'Analytic Server.

Vous devez ajouter les propriétés suivantes dans le fichier `core-site.xml` HDFS dans les cas où les données sont accessibles à partir du système de fichiers HDFS via Hive/HCatalog :

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

Authentification interdomaine Kerberos

Analytic Server prend en charge l'authentification interdomaine Kerberos. Pour activer cette fonction, vous devez d'abord vérifier que l'authentification interdomaine KDC est activée, puis ajouter le paramètre suivant à la section **Custom analytics.cfg** de la configuration Ambari d'Analytic Server :

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

Configuration du cluster - Paramètres d'optimisation des performances et résultats

Configuration Spark

Analytic Server utilise le mode `yarn-client` pour interagir avec YARN et exécuter des travaux Spark sur le cluster Hadoop.

Configuration **Analytic Server personnalisée :**

- Les paramètres Ambari sont définis dans la section **Custom analytics.cfg** de la configuration Analytic Server Ambari.
- Les paramètres Cloudera sont situés dans la section **Analytic Server Advanced Configuration Snippet (Safety Valve) for analyticserver-conf/config.properties** de Cloudera Manager.

1. Envisagez d'augmenter la valeur du paramètre de configuration **spark.driver.memory** en ajoutant un élément de configuration à la configuration Analytic Server personnalisée (s'il n'est pas défini explicitement, cet élément reçoit la valeur par défaut `1g`). Exemple :

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

2. Sélectionnez l'une des options suivantes d'utilisation d'Analytic Server avec Spark.
   - **Option A : configuration d'allocation de ressources statique**

     Trois paramètres doivent être définis dans la configuration Analytic Server personnalisée :

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

     La procédure suivante explique comment déterminer les valeurs des paramètres.

     a. Etablissez le pourcentage, en termes d'unité centrale et de mémoire, qu'Analytic Server peut en permanence allouer à Spark. Cette opération permet d'obtenir un nombre de coeurs (C) spécifique et une quantité de mémoire fixe qui peut être utilisée sur chaque machine (M).

     b. Etablissez le nombre de programmes d'exécution (E) que chaque machine peut exécuter. Ces programmes d'exécution s'exécutent sous la forme de conteneurs Hadoop (processus) distincts sur chaque noeud de cluster. En général, une valeur supérieure à 2 est appropriée mais la valeur doit être inférieure au nombre total de coeurs. La mémoire allouée à Spark est répartie entre ces programmes d'exécution. La sélection d'une valeur élevée pour ce paramètre réduit la quantité de mémoire allouée à chaque conteneur.

     c. Etablissez le nombre de coeurs utilisés pour chaque programme d'exécution (CE). Généralement, cette valeur correspond à C/E (nombre de coeurs de chaque machine alloués à l'application Spark, divisé par le nombre total de programmes d'exécution).

     d. Etablissez la quantité de mémoire utilisée pour chaque programme d'exécution (ME). Cette valeur correspond généralement à M/E.

     Remarque : Le nombre de programmes d'exécution et de coeurs utilisés doit être équilibré afin que la quantité de mémoire de chaque programme d'exécution soit supérieure à 3G * CE.

Chaque coeur de chaque programme d'exécution doit disposer d'au moins 3 Go de mémoire qui seront utilisés comme mémoire de stockage ou de calcul.

```
spark.executor.instances = <E>*N /<E> // valeur déterminée à l'étape b où N correspond au nombre de noeuds de traitement
spark.executor.cores = <CE> // valeur déterminée à l'étape c
spark.executor.memory = <ME> // valeur déterminée à l'étape d
```

Figure 1. Paramètres de Spark dans la section Custom analytics.cfg

- **Option B : Configuration d'allocation de ressources dynamique**

  Si vous utilisez cette option, tous les programmes d'exécution alloués par YARN sont augmentés/diminués dynamiquement en fonction des ressources disponibles effectives dans le cluster complet.

  La configuration minimale est la suivante :

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

  Configuration standard :

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

  Remarques :

  – `spark.executor.instances = <E>` ne doit pas être utilisé, faute de quoi l'allocation de ressources statique est employée.

  – Les considérations concernant les valeurs à affecter au nombre de coeurs de programme d'exécution et à la mémoire sont identiques à celles de l'option A.

3. Vous pouvez désactiver le cache dans la configuration Analytic Server personnalisée en spécifiant les paramètres suivants :

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

Figure 2. Paramètres du cache de Spark dans la section Custom analytics.cfg

Le cache de Spark ne doit pas être désactivé lorsque des flux IBM SPSS Modeler volumineux sont utilisés. La désactivation du cache Spark dans cette instance entraîne un ralentissement des flux en exécution mais permet d'éviter des problèmes de saturation de la mémoire pouvant survenir lorsque la quantité de mémoire spécifiée pour chaque programme d'exécution est faible.

Configuration de la machine virtuelle Java (JVM)

Paramètres Ambari :

1. Dans la configuration Ambari d'Analytic Server, spécifiez la quantité de mémoire que le serveur peut utiliser pour le traitement local. La valeur par défaut (2 Go) peut être utilisée pour des flux petits et moyens mais une valeur de segment de mémoire supérieure (par exemple, 10 Go) doit être utilisée pour les flux plus importants.

   Analytic Server > **Configuration** > **Advanced analytic-jvm-options**
2. Remplacez `-Xmx2048M` par `-Xmx10G`, enregistrez la configuration et redémarrez Analytic Server.

Figure 3. Paramètres Advanced analytic-jvm-options

Paramètres Cloudera :

1. Dans Cloudera Manager, accédez à l'onglet **Configuration** du service Analytic Server et mettez à jour le contrôle `jvm-options` en spécifiant la quantité de mémoire que peut utiliser le serveur pour traitement local. La valeur par défaut (2 Go) peut être utilisée pour des flux petits et moyens mais une valeur de segment de mémoire supérieure (par exemple, 10 Go) doit être utilisée pour les flux plus importants.

   Analytic Server service > **Configuration** > **jvm-options**
2. Remplacez `-Xmx2048M` par `-Xmx10G`, enregistrez la configuration et redémarrez Analytic Server.

Configuration Yarn MapReduce2 :

- Si vous devez exécuter des travaux MapReduce en parallèle de travaux Spark pour l'exécution d'Analytic Server, vous devez configurer le cluster Yarn pour allouer 4 Go de mémoire par conteneur Yarn.

Configuration Zookeeper :

- Cloudera requiert une mise à jour manuelle de la configuration Zookeeper. Pour plus d'informations, voir https://www.cloudera.com/documentation/enterprise/5-4-x/topics/cdh_ig_zookeeper_server_maintain.html.
- Si vous utilisez des flux SPSS Modeler complexes ou des données étendues (nombre de zones élevé), il est possible que des travaux échouent en raison de l'interruption d'une connexion Analytic Server–Zookeeper. Le problème est dû à la taille élevée du programme que SPSS Modeler Server envoie à Analytic Server. Il est possible que le problème se produise moins souvent dans Analytic Server 3.0 (ou suivante). Pour résoudre le problème, procédez comme suit :
  1. Dans la console Ambari, accédez au service Zookeeper dans l'onglet **Configs**, ajoutez la ligne suivante au modèle `zookeeper-env` sous **Advanced zookeeper-env**, puis redémarrez le service Zookeeper.

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

Figure 4. Paramètres du modèle zookeeper-env

2. Dans la console Ambari, accédez à l'onglet **Configs** du service Analytic Server, ajoutez les éléments suivants à **Advanced analytics-jvm-options**, puis redémarrez le service Analytic Server.

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

Figure 5. Paramètres Advanced analytics-jvm-options

Remarque : Si le problème persiste, augmentez la valeur de `-Djute.maxbuffer` en la faisant passer de `2097152` à `4194304` aux deux endroits.

Recommandations pour les flux IBM SPSS Modeler

Remarque : La plupart des recommandations ci-après s'appliquent également aux petits volumes de données.

Prototype pour les petits volumes de données

Lorsque vous concevez un flux, vous ajoutez souvent quelques noeuds, testez le flux à ce stade, ajoutez éventuellement un noeud pour vérifier la sortie tabulaire et graphique, puis continuez à élaborer le flux. En général, vous ne pouvez pas vous permettre de transmettre des données volumineuses chaque fois que vous testez le flux.

La création d'un échantillon de données pertinent à partir de vos mégadonnées permet de tester le flux par rapport aux données réelles sans perdre du temps à exécuter des tests de données complets. L'échantillon de données doit inclure suffisamment de données pour permettre une exécution réussie de votre flux. Par exemple, si vous prévoyez d'analyser des transactions dans des magasins situés dans le Minnesota, l'échantillon de données doit inclure des transactions provenant des magasins du Minnesota.

Une fois l'échantillonnage effectué, vous pouvez :

- Créer un cache de l'échantillon de données sur le cluster où les mégadonnées se trouvent

  Avantages : Simplicité. Ne requiert pas le changement de noeuds source

  Inconvénients : Le cache est vidé à la fin de la session
- Créer une source de données Analytic Server qui contient l'échantillon de données

  Avantages : Source de données permanentes

  Inconvénients : Requiert l'édition ou le changement de noeuds source
- Télécharger l'échantillon de données sur le système local et créer une source de données locale

  Avantages : N'utilise pas des ressources de cluster lors de la définition du prototype ; le client SPSS Modeler est plus efficace qu'Analytic Server lorsque vous travaillez avec un petit volume de données.

Avantages : Requiert le changement de noeuds source

Créez des noeuds Type et Filtrer distincts des noeuds source

Chaque noeud source SPSS Modeler inclut également la fonction combinée des noeuds Filtrer et Type. Cette fonction est utile pour conserver un modèle cohérent mais il rend votre travail plus difficile lorsque vous basculez vers des types de noeud source différents. En outre, il masque les opérations Type et Filtrer qui sont effectuées.

Placez les noeuds Filtrer et Sélectionner aussi près que possible du noeud source

Cette action permet de réduire le nombre d'enregistrements dans des opérations en aval.

Evitez le noeud Trier chaque fois que cela est possible

Analytic Server ne prend pas en charge les optimisations sur les noeuds qui dépendent des données triées (comme le noeud Fusionner). A ce titre, un noeud Trier au milieu du flux est rarement utile. Le noeud Trier présente un intérêt lorsqu'il est immédiatement suivi d'un noeud Echantillon pour obtenir les N premiers (ou les N derniers) enregistrements.

Effectuez des calculs uniquement pour les zones qui doivent être utilisées

Ne calculez pas une zone pour la filtrer immédiatement après.

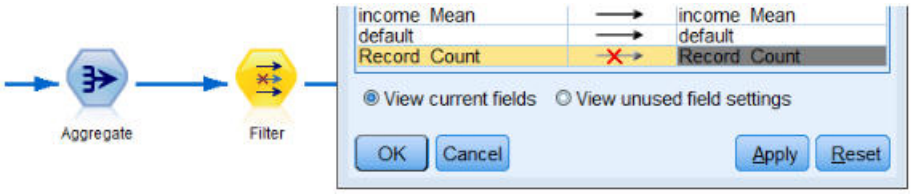

Figure 6. Options de zone de Modeler

Dans la mesure du possible, évitez de créer de nombreuses zones temporaires sans pour autant rendre les expressions difficiles à comprendre. Par exemple, au lieu de définir l'exemple suivant :

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

définissez l'exemple suivant :

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

L'intégration de données temporaires dans des expressions en ligne peut accroître les performances lorsque de nombreuses zones sont transformées.

Définissez le stockage dans la source de données

Les opérations qui modifient le type de stockage d'une zone (par exemple, string remplacé par integer) au milieu du flux risquent d'altérer les performances globales. Vous pouvez définir le stockage pour les zones lors de la définition des sources de données dans la console Analytic Server afin d'éviter la répétition de ces conversions.

Utilisez SPSS Modeler lorsque vous utilisez des volumes de données limités

Manipulez les mégadonnées avec Analytic Server, puis utilisez SPSS Modeler pour terminer les calculs appliqués aux petits volumes de données.

Sélectionnez les propriétés de flux Analytic Server appropriées

Configurez les propriétés de flux appropriées (**Tools** > **Options** > **Stream Properties** > **Analytic Server**) et décidez si vous souhaitez autoriser le traitement de données à quitter Analytic Server et à continuer dans SPSS Modeler (lorsqu'un noeud ne peut pas être exécuté dans Analytic Server).

Par défaut, SPSS Modeler est configuré pour signaler une erreur et arrêter de s'exécuter dans cette situation. Vous pouvez ignorer l'erreur en remplaçant le paramètre `Error` par `Warn` et en adaptant la quantité de données maximale qui peuvent être traitées dans SPSS Modeler. Par exemple, vous pouvez mettre à jour la vitesse de transfert des données en modifiant la valeur par défaut de 10000 enregistrements, si nécessaire. Notez que cette limite s'applique également lors de l'affichage de résultats qui utilisent le noeud des tables SPSS Modeler. Si la limite est dépassée, SPSS Modeler indique que l'extraction des données a dépassé la limite définie dans les propriétés de flux.

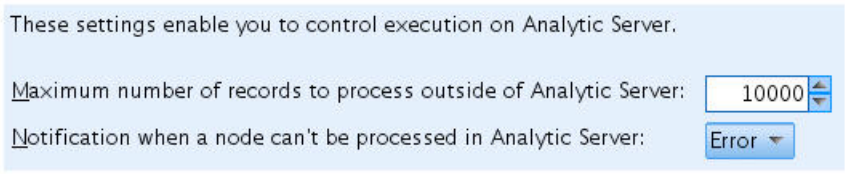

Figure 7. Paramètres Analytic Server

Utilisez des noeuds source Analytic Server

Analytic Server ne peut pas se connecter à des sources de données Base de données différentes mais SPSS Modeler requiert que tous les noeuds source soient des noeuds source Analytic Server (pour que l'ensemble du flux s'exécute en tant que travail Analytic Server). Pour permettre l'exécution de l'ensemble du flux dans Analytic Server, vous devez remplacer le noeud source de base de données par un noeud source Analytic Server et vous devez créer une source de données Base de données Analytic Server dans la console Analytic Server.

Examinez l'utilisation des noeuds qui ne sont pas pris en charge

Analytic Server ne prend pas en charge tous les noeuds (le noeud Transposer est un bon exemple). Pour fusionner les résultats d'une opération Transposer avec le reste du flux et l'exécuter dans Analytic Server, un sous-flux qui inclut un noeud Transposer doit être placé dans une source de données Analytic Server qui utilise un noeud Exporter Analytic Server. Vous pouvez ensuite joindre un noeud source Analytic Server où le flux a été interrompu pour le copier dans Analytic Server.

Remarque : L'opération Transposer est adaptée à des opérations exceptionnelles ou exécutées rarement mais elle ne doit pas être utilisée pour des opérations de flux courantes.

Déterminez si un flux fonctionne dans Analytic Server avant de l'exécuter

Après avoir préparé un flux à exécuter dans Analytic Server, sélectionnez un noeud de terminal et utilisez la fonction d'aperçu SPSS Modeler (option **d'aperçu de l'exécution** dans la barre d'outils) pour vérifier que les noeuds impliqués dans l'exécution du noeud de terminal fonctionnent dans Analytic Server (sans exécuter le flux). Des problèmes sont signalés dans la fenêtre de messages.

Combinez des opérations Fusionner consécutives

Lorsqu'ils sont de type Joindre et qu'ils possèdent les mêmes clés, combinez une série de noeuds Fusionner pour obtenir un noeud unique.

Combinez des sous-flux identiques

Dans la mesure du possible, essayez de combiner des sous-flux identiques, en particulier s'ils incluent des opérations complexes (par exemple, fusion et tri). SPSS Modeler effectue ces opérations une fois et utilise le cache pour améliorer les performances. Dans l'exemple suivant, les flux sont identiques jusqu'au noeud **newField**.

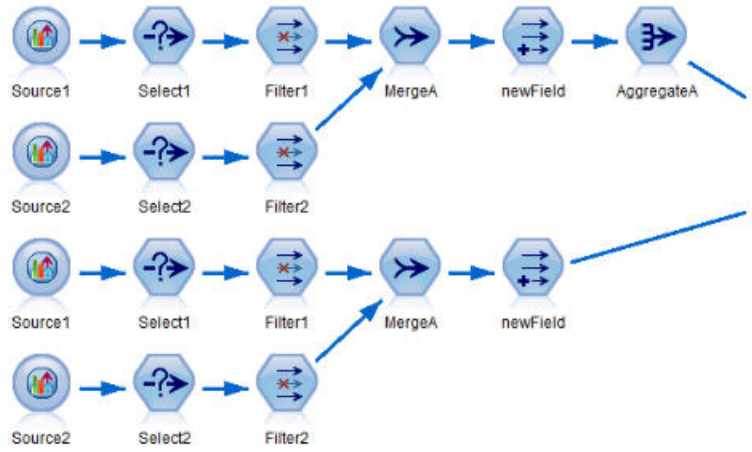

Figure 8. Exemple de flux

Un sous-flux structuré de la manière suivante est plus efficace (et plus simple à gérer) :

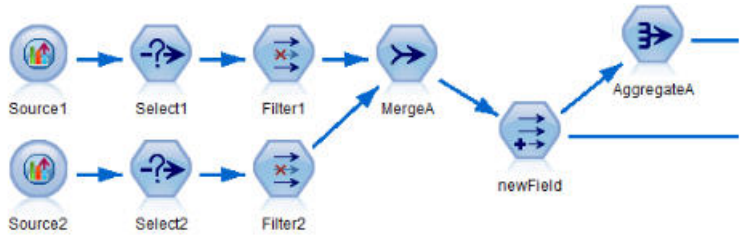

Figure 9. Exemple de flux

Supprimez les noeuds Type superflus

Figure 10. Exemple de flux

Evitez les noeuds Type inutiles lors d'une exécution sur Analytic Server. L'opération de `lecture de valeurs` du noeud Type démarre un travail MapReduce. En général, elle permet de faire des gains ponctuellement sauf si vous effacez les valeurs des noeuds Type.

Documentez chaque flux de manière exhaustive

L'exemple ci-après décrit un flux complexe qui contient un certain nombre de sous-flux.

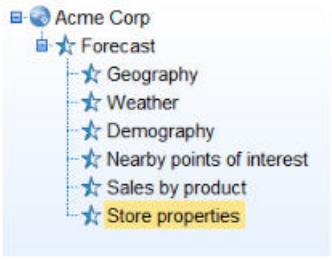

Figure 11. Exemple de sous-flux

Dans ces cas de figure, vous devez nommer correctement les supernoeuds et documenter le flux (comme vous le feriez pour le code). Un commentaire clair peut fournir des informations importantes aux autres analystes qui lisent ou gèrent le flux. Exemple :

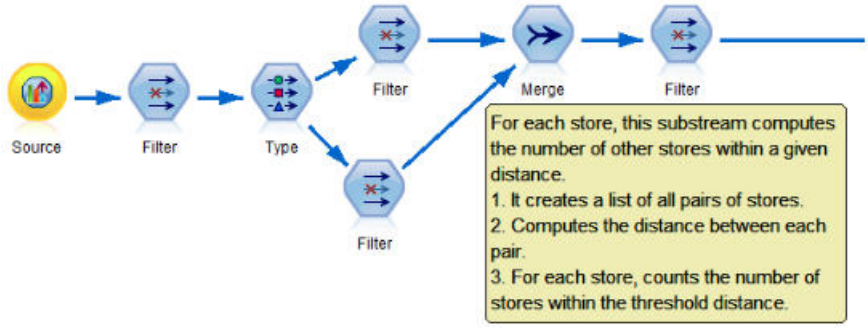

Figure 12. Exemple de flux avec des commentaires

Lorsque vous développez des flux, utilisez des caches SPSS Modeler pour stocker rapidement des résultats intermédiaires

Dans les flux qui s'exécutent sur le noeud Analytic Server, la mise en cache fonctionne en stockant les données d'une partie spécifique du flux dans des fichiers temporaires sur le système de fichiers HDFS (par opposition au stockage sur le serveur SPSS Modeler). Les caches fonctionnent bien avec les mégadonnées et peuvent être utilisés dans des flux exécutés sur Analytic Server.

Chapitre 5. Traitement des incidents

Analytic Server fournit plusieurs outils pratiques pour l'identification des problèmes.

Consignation

Analytic Server crée des fichiers journaux et des fichiers de trace client pouvant être utiles pour diagnostiquer les problèmes. Avec l'installation par défaut de Liberty, vous pouvez rechercher les fichiers journaux dans le répertoire `{RACINE_AS}/ae_wlpserver/usr/servers/aeserver/logs`.

La configuration de consignation par défaut génère deux fichiers journaux qui se renouvellent quotidiennement.

as.log Ce fichier contient le récapitulatif de haut niveau des messages informatifs d'avertissement et d'erreur. Vérifiez d'abord ce fichier lorsque des erreurs se produisent au niveau du serveur ne pouvant pas être résolues à l'aide du message d'erreur s'affichant sur l'interface utilisateur.

as_trace.log
Ce fichier contient toutes les entrées du fichier `ae.log`, ainsi que des informations supplémentaires qui s'adressent essentiellement à l'équipe de support et de développement d'IBM à des fins de débogage.

Analytic Server utilise Apache LOG4J comme application de consignation sous-jacente. A l'aide de LOG4J, vous pouvez ajuster la consignation de manière dynamique en éditant le fichier de configuration `{RACINE_SERVEUR_AS}/configuration/log4j.xml`. Vous serez peut-être amené à le faire à la demande du support pour vous aider à diagnostiquer les problèmes ou pour limiter le nombre de fichiers journaux conservés. Les modifications apportées à ce fichier sont détectées automatiquement en quelques secondes de sorte qu'il n'est pas nécessaire de redémarrer Analytic Server.

Pour plus d'informations sur log4j et le fichier de configuration, voir la documentation sur le site officiel d'Apache à l'adresse http://logging.apache.org/log4j/.

Informations sur la version

Pour savoir quelle version d'Analytic Server est installée, reportez-vous au dossier `{RACINE_AS}/properties/version`. Les fichiers suivants contiennent des informations sur la version.

IBM_SPSS_Analytic_Server-*.swtag
Contient les infos produit détaillées.

version.txt
Version et numéro de compilation du produit installé.

Collecteur de journal

Lorsqu'il est impossible de résoudre des problèmes en consultant directement les fichiers journaux, vous pouvez regrouper tous les journaux et les envoyer au support IBM. Il existe un utilitaire permettant de simplifier la collecte de toutes les données nécessaires.

A l'aide d'un interpréteur de commandes, exécutez les commandes suivantes :

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

Ces commandes créent un fichier compressé sous `{RACINE_AS}/bin`. Le fichier compressé contient tous les fichiers journaux et toutes les informations sur la version du produit.

Problèmes courants

La présente section décrit certains problèmes d'administration courants et la manière d'y remédier.

Exécution de flux

Les travaux R traduisent en Unicode les mots qui ne sont pas en anglais

Dans les clusters Cloudera, si le codage système des serveurs Hadoop n'est pas UTF-8, R traduit en Unicode les mots qui ne sont pas en anglais.

1. Accédez à l'onglet de configuration YARN dans la console Cloudera Manager.
2. Ajoutez le paramètre ci-après dans la zone "NodeManager Environment Advanced Configuration Snippet (Safety Valve)".

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

Echec de l'exécution des travaux PySpark

Vérifiez que le service Spark est déployé sur tous les noeuds Analytic Server et les gestionnaires de noeud.

Echec de l'exécution des travaux PySpark dans les environnements activés pour Kerberos

Vous devez exécuter la commande `kinit`, puis redémarrer Analytic Server, pour que les tests PySpark puissent aboutir. Exemple :

HDP Kerberos

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

CDH Kerberos

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

Erreurs de mémoire

Configuration de YARN après des erreurs de mémoire du programme d'exécution

L'erreur suivante peut survenir lorsque la mémoire du programme d'exécution requise est supérieure au seuil maximal :

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

La procédure suivante fournit les paramètres de configuration YARN requis pour résoudre le problème.

Pour Ambari

1. Dans l'interface utilisateur Ambari, accédez à **YARN** > **Configs** > **Settings**.
2. Augmentez le noeud de mémoire **(mémoire allouée pour tous les conteneurs YARN)** en indiquant la valeur `8192MB`.
3. Augmentez les valeurs de conteneur :
   - Attribuez la valeur `682MB` à **Minimum Container Size (Memory)**
   - Attribuez la valeur `8192MB` à **Maximum Container Size (Memory)**
4. Augmentez la valeur **Maximum Container Size (VCores)** à 3.
5. Redémarrez YARN, Spark et le service Analytic Server.

Pour Cloudera

1. Augmentez la valeur de `yarn.nodemanager.resource.memory-mb` à 8 Go
   - Dans l'interface utilisateur de Cloudera Manager, accédez à **Yarn service** > **Configurations** > **Search Container Memory** et augmentez la valeur à `8GB`.

2. Dans l'interface utilisateur de Cloudera Manager, accédez à **YARN service** > **Quick Links** puis sélectionnez **Dynamic Resource Pools**.
3. Sous **Configuration**, cliquez sur **edit** pour chaque pool disponible et sous **YARN**, attribuez la valeur 4 à **Max Running Apps**.
4. Redémarrez YARN, Spark et le service Analytic Server.

Hadoop avec Apache Spark 2.x

- La plupart des travaux `forcespark` et `forcehadoop` échouent lorsque Hadoop et Apache Spark 2.x coexistent dans le même environnement. L'erreur apparaît dans le journal d'application Yarn comme suit : `java.lang.NoClassDefFoundError: org/apache/hadoop/fs/FSDataInputStream`.

  Le problème peut être résolu manuellement en modifiant le fichier `/etc/spark2/conf/spark-defaults.conf` comme suit :

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

- Lorsque deux versions du JDK sont installées sur le même système, Cloudera utilise le JDK 1.7 alors que Spark 2.x utilise le JDK 1.8. L'exécution de travaux `forcespark` ou `forcehadoop` avec Apache Spark 2.x peut entraîner l'échec de tous ces travaux avec le message d'erreur suivant :

  ```
  L'exécution a échoué. Motif : org/apache/spark/api/java/function/PairFunction :
  Version majeure.mineurs 52.0 non prise en charge
  ```

  Pour Cloudera, ajoutez la ligne suivante dans la section **Analytic Server Advanced Configuration Snippet (Safety Valve) for server.env** de Cloudera Manager :

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

Accord des droits `admin` aux utilisateurs d'Apache Hive UDF

Une erreur `Invalid function` peut de produire après l'enregistrement d'Analytic Server Apache Hive UDF. Par défaut, il existe deux rôles Hive (`admin` et `public`). Les utilisateurs Hive appartiennent au rôle `public`. Hive UDF nécessite que les utilisateurs enregistrés disposent du privilège `admin` (la sécurité Hive est activée).

Pour accorder les droits `admin` aux utilisateurs de Hive UDF :

1. Connectez-vous à Beeline en tant que Hive :

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

2. Exécutez la commande suivante dans Beeline :

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

   Remarque : Autres commandes SQL utiles :

   Afficher les rôles déjà affectés à l'utilisateur `hive`

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

   Afficher les utilisateurs affectés au rôle `public`

   ```
   show principals public;
   ```

3. Redémarrez Hive et réenregistrez Analytic Server Hive UDF.

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

Erreur HiveDB

L'erreur suivante peut se produire lors de l'écriture sur HiveDB :

```
(AEQAE4805E) Execution failed. Reason: com.google.common.io.Closeables.closeQuietly(Ljava/io/Closeable;)
```

Cette erreur est due à la présence de plusieurs versions du fichier `guava-*.jar` sur le cluster Hadoop. Cette erreur peut être résolue en procédant de la manière suivante (dans cet exemple, on utilise HDP 3.1) :

1. Ouvrez la console Ambari et arrêtez le service Analytic Server.
2. Copiez `/usr/hdp/3.1.0.0-78/spark2/jars/guava-14.0.1.jar` dans `{AS_ROOT}/ae_wlpserver/usr/servers/aeserver/apps/AE_BOOT.war/WEB-INF/lib`.
3. Dans la console Ambari, actualisez le service Analytic Server, puis démarrez-le.

Réglage des performances

Cette section explique comment optimiser les performances de votre système.

Analytic Server est un composant de l'infrastructure Ambari qui utilise d'autres composants tels que HDFS, Yarn et Spark. Les techniques de réglage des performances courantes pour Hadoop, HDFS et Spark s'appliquent aux charges de travail d'Analytic Server. Chaque charge de travail d'Analytic Server est différente ; par conséquent, vous devez expérimenter les réglages en fonction de votre charge de travail de déploiements spécifique. Les propriétés et les conseils de réglage ci-après constituent des modifications clés qui ont eu un impact sur les résultats des tests de benchmarking et de mise à l'échelle d'Analytic Server.

Lorsque le premier travail s'exécute sur Analytic Server, le serveur démarre une application Spark persistante qui reste active jusqu'à ce qu'Analytic Server soit arrêté. L'application Spark persistante alloue et conserve toutes les ressources de cluster qui lui sont allouées pendant toute la durée d'exécution d'Analytic Server, même si un travail Analytic Server n'est pas en cours d'exécution. Réfléchissez bien à la quantité de ressources allouées aux applications Spark d'Analytic Server. Si toutes les ressources de cluster sont allouées à l'application Spark d'Analytic Server, il se peut que les autres travaux soient différés ou qu'ils ne soient pas exécutés. Ils peuvent être mis en file d'attente jusqu'à ce que des ressources suffisantes soient disponibles, mais ces ressources seront consommées par l'application Spark d'Analytic Server.

Si plusieurs services Analytic Server sont configurés et déployés, chaque instance de service peut potentiellement allouer sa propre application Spark persistante. Par exemple, si deux services Analytic Server sont déployés pour prendre en charge la reprise en ligne à haute disponibilité, deux applications Spark persistantes peuvent être actives et allouer chacune des ressources de cluster.

Certaines situations sont encore plus complexes. Par exemple, Analytic Server peut démarrer un travail MapReduce requérant des ressources de cluster. Ces travaux MapReduce nécessitent des ressources qui ne sont pas allouées à l'application Spark. Les composants spécifiques qui requièrent des travaux MapReduce sont des générations de modèle PSM.

Les propriétés ci-après peuvent être configurées pour allouer des ressources à l'application Spark. Si elles sont définies dans le fichier spark-defaults.conf de l'installation Spark, elles sont appliquées à tous les travaux Spark exécutés dans l'environnement. Si elles sont définies dans la configuration d'Analytic Server en tant que propriétés personnalisées sous la section “Custom analytic.cfg”, elles sont appliquées à l'application Spark d'Analytic Server seulement.

spark.executor.memory

Quantité de mémoire à utiliser par processus de programme d'exécution.

spark.executor.instances

Nombre de processus de programme d'exécution à démarrer.

spark.executor.cores

Nombre d'unité d'exécution de tâche de programme d'exécution par processus de programme d'exécution. Cette valeur doit être comprise entre 1 et 5.

Exemple de définition des trois propriétés principales de Spark. Il existe 10 noeuds de données dans un cluster HDFS et chaque noeud de données possède 24 coeurs et 48 Go de mémoire, et n'exécute que des processus HDFS. Voici une manière de configurer les propriétés pour cet environnement, si l'on suppose que vous n'exécutez que des travaux Analytic Server dans cet environnement et que vous envisagez une allocation maximale à une seule application Spark d'Analytic Server.

- Définissez spark.executor.instances=20. Cette propriété permet d'exécuter deux processus de programme d'exécution Spark par noeud de données.
- Définissez spark.executor.memory=22G. Cette propriété définit la taille de segment de mémoire maximale de 22 Go pour chaque processus de programme d'exécution Spark ; en d'autres termes, elle alloue 44 Go à chaque noeud de données. D'autres machines virtuelles Java, ainsi que le système d'exploitation, ont besoin de la mémoire restante.
- Définissez spark.executor.cores=5. Cette propriété fournit 5 unités d'exécution de tâche à chaque programme d'exécution Spark, pour un total de 10 unités d'exécution de tâche par noeud de données.

Surveillance de l'interface utilisateur Spark pour les travaux en cours d'exécution

Si le message de déversement sur le disque (Spill to disk) apparaît, il se peut que les performances soient affectées. Voici quelques solutions :

- Augmentez la mémoire et allouez-la à des programmes d'exécution Spark via **spark.executor.memory**.
- Réduisez le nombre d'éléments **spark.executor.cores**. Vous réduirez ainsi le nombre d'unité d'exécution de tâche simultanées allouant de la mémoire, mais aussi le nombre de parallélismes pour les travaux.
- Changez les propriétés de mémoire Spark. Pourcentage d'allocation **spark.shuffle.memoryFraction** et **spark.storage.memoryFraction** du segment de mémoire de programme d'exécution Spark pour Spark.

Vérification de la quantité de mémoire pour le noeud de nom

Si le nombre de blocs dans HDFS est important et augmente, assurez-vous que le segment de mémoire du noeud de nom augmente en conséquence. Il s'agit d'une recommandation de réglage HDFS commune.

Modification de la quantité de mémoire utilisée pour la mise en cache

Par défaut, **spark.storage.memoryFraction** possède la valeur 0.6. Vous pouvez l'augmenter et définir la valeur 0.8 si la taille de bloc HDFS des données est 64 Mo. Si la taille de bloc HDFS des données d'entrée est supérieure à 64 Mo, vous ne devez augmenter cette valeur que si la mémoire allouée par tâche est supérieure à 2 Go.

Réglage des performances de l'évaluation du modèle

Vous pouvez améliorer les performances des travaux d'évaluation du modèle pour les ensembles de données volumineux avec le moteur Apache Spark comme suit. Normalement, ces étapes n'ont pas d'impact sur le fonctionnement des services autres que les services Analytic Server dans le cluster.

1. Vérifiez si libtcmalloc_minimal.so{/version} est installé sur chaque noeud dans le cluster.

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

2. Si libtcmalloc_minimal.so n'est pas installé, installez le package propre au système d'exploitation contenant la bibliothèque libtcmalloc_minimal sur chaque noeud de votre cluster ou générez et installez manuellement libtcmalloc_minimal. Exemple :

   Ubuntu :

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

   Red Hat Enterprise Linux 6.x (x64) :

   a. Installez le référentiel EPEL pour RedHat (s'il n'est pas déjà installé)

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

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

Génération manuelle :

a. Téléchargez le fichier gperftools-2.4.tar.gz depuis https://github.com/gperftools/gperftools/releases

b. `tar zxvf gperftools-2.4.tar.gz`

c. `cd gperftools-2.4`

d. `./configure --disable-cpu-profiler --disable-heap-profiler --disable-heap-checker --disable-debugalloc --enable-minimal`

e. `make`

f. `sudo make install`

3. Notez l'un des emplacements du fichier de bibliothèque installé libtcmalloc_minimal.so{.version} figurant dans les résultats d'exécution de la commande suivante sur un ou plusieurs noeuds :

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

   Si des noeuds du cluster exécutent un mélange de systèmes d'exploitation, plusieurs emplacements peuvent exister pour ce fichier.

4. Dans la console Ambari, accédez à la configuration Analytic Server et sous la section Custom analytics.cfg, configurez la clé spark.executorEnv.LD_PRELOAD en indiquant l'emplacement de la bibliothèque comme valeur. Après cette modification, redémarrez le service Analytic Server. Par exemple, si la bibliothèque est installée dans /usr/lib64/libtcmalloc_minimal.so.4, la configuration est :

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

   Si plusieurs emplacements sont requis, séparez-les par un espace, comme dans l'exemple suivant :

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

   Si la bibliothèque `libtcmalloc_minimal.so` n'est pas installée sur l'un des noeuds dans l'un des emplacements configurés, il n'y a pas d'erreur générée mais les performances d'évaluation du modèle risquent d'être ralenties sur les noeuds concernés.

Jointure map-side de Spark

La mise en oeuvre de jointure de Spark dans Analytic Server ne prend pas en charge la fonction de jointure map-side. (La jointure de Spark est principalement une jointure reduce-side). La mise en oeuvre n'utilise pas des jointures map-side pour optimiser les jointures lorsque les données en entrée ne sont pas très volumineuses. L'absence de jointures map-side génère un travail Spark qui utilise énormément de ressources et qui échoue.

Pour optimiser les jointures lors de l'exécution des jointures map-size de Spark dans Analytic Server (ou des travaux Spark natifs basés sur la taille RDD la plus faible), vous pouvez ajouter la propriété `spark.msj.maxBroadcast` au fichier `analytics.cfg` (`SPSS Analytic Server/Configs/Custom analytics.cfg`) ou à `analytics-meta`.

Remarques

Le présent document a été développé pour des produits et des services proposés aux Etats-Unis. Il peut être disponible dans d'autres langues auprès d'IBM. Toutefois, il peut être nécessaire de posséder une copie du produit ou de la version du produit dans cette langue pour pouvoir y accéder.

Le présent document peut contenir des informations ou des références concernant certains produits, logiciels ou services IBM non annoncés dans ce pays. Pour plus de détails, référez-vous aux documents d'annonce disponibles dans votre pays, ou adressez-vous à votre partenaire commercial IBM. Toute référence à un produit, logiciel ou service IBM n'implique pas que seul ce produit, logiciel ou service IBM puisse être utilisé. Tout autre élément fonctionnellement équivalent peut être utilisé, s'il n'enfreint aucun droit d'IBM. Il est de la responsabilité de l'utilisateur d'évaluer et de vérifier lui-même les installations et applications réalisées avec des produits, logiciels ou services non expressément référencés par IBM.

IBM peut détenir des brevets ou des demandes de brevet couvrant les produits mentionnés dans le présent document. La remise de ce document ne vous donne aucun droit de licence sur ces brevets ou demandes de brevet. Si vous désirez recevoir des informations concernant l'acquisition de licences, veuillez en faire la demande par écrit à l'adresse suivante :

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
U.S.A.

Pour le Canada, veuillez adresser votre courrier à :

IBM Director of Commercial Relations
IBM Canada Ltd.
3600 Steeles Avenue East
Markham, Ontario
L3R 9Z7 Canada

Pour toute demande au sujet des licences concernant les produits utilisant un jeu de caractères codé sur deux octets, contactez le service de propriété intellectuelle IBM de votre pays ou envoyez vos questions par courrier à l'adresse suivante :

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

LE PRESENT DOCUMENT EST LIVRE EN L'ETAT SANS AUCUNE GARANTIE EXPLICITE OU IMPLICITE. IBM DECLINE NOTAMMENT TOUTE RESPONSABILITE RELATIVE A CES INFORMATIONS EN CAS DE CONTREFACON AINSI QU'EN CAS DE DEFAUT D'APTITUDE A L'EXECUTION D'UN TRAVAIL DONNE. Certaines juridictions n'autorisent pas l'exclusion des garanties tacites, auquel cas l'exclusion ci-dessus ne vous sera pas applicable.

Le présent document peut contenir des inexactitudes ou des coquilles. Ce document est mis à jour périodiquement. Chaque nouvelle édition inclut les mises à jour. IBM peut, à tout moment et sans préavis, modifier les produits et logiciels décrits dans ce document.

Les références à des sites Web non IBM sont fournies à titre d'information uniquement et n'impliquent en aucun cas une adhésion aux données qu'ils contiennent. Les documents sur ces sites web ne font pas partie des documents de ce produit IBM et l'utilisation de ces sites web se fait à vos propres risques.

IBM pourra utiliser ou diffuser, de toute manière qu'elle jugera appropriée et sans aucune obligation de sa part, tout ou partie des informations qui lui seront fournies.

Les licenciés souhaitant obtenir des informations permettant : (i) l'échange des données entre des logiciels créés de façon indépendante et d'autres logiciels (dont celui-ci), et (ii) l'utilisation mutuelle des données ainsi échangées, doivent adresser leur demande à :

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
U.S.A.

Ces informations peuvent être soumises à des conditions particulières, prévoyant notamment le paiement d'une redevance.

Le logiciel sous licence décrit dans ce document et tous les éléments sous licence disponibles s'y rapportant sont fournis par IBM conformément aux dispositions du document intitulé IBM Customer Agreement, des Conditions internationales d'utilisation des logiciels IBM ou de tout autre accord équivalent.

Les données de performance et les exemples client ne sont présentés qu'à des fins d'illustration. Les performances réelles peuvent varier en fonction des configurations et des conditions d'exploitation.

Les informations concernant des produits non IBM ont été obtenues auprès des fournisseurs de ces produits, par l'intermédiaire d'annonces publiques ou via d'autres sources disponibles. IBM n'a pas testé ces produits et ne peut confirmer l'exactitude de leurs performances ni leur compatibilité. Aucune réclamation relative à des produits non IBM ne pourra être reçue par IBM. Toute question concernant les performances de produits non IBM doit être adressée aux fournisseurs de ces produits.

Les instructions relatives aux intentions d'IBM pour ses opérations à venir sont susceptibles d'être modifiées ou annulées sans préavis, et doivent être considérées uniquement comme un objectif.

Tous les tarifs indiqués sont les prix de vente actuels suggérés par IBM et sont susceptibles d'être modifiés sans préavis. Les tarifs appliqués peuvent varier selon les revendeurs.

Ces informations sont fournies uniquement à titre de planification. Elles sont susceptibles d'être modifiées avant la mise à disposition des produits décrits.

Le présent document peut contenir des exemples de données et de rapports utilisés couramment dans l'environnement professionnel. Ces exemples mentionnent des noms fictifs de personnes, de sociétés, de marques ou de produits à des fins illustratives ou explicatives uniquement. Tous ces noms sont fictifs, et toute ressemblance avec des noms de personnes ou de sociétés réelles serait purement fortuite.

LICENCE DE COPYRIGHT :

Le présent document peut contenir des exemples de données et de rapports utilisés couramment dans l'environnement professionnel. Ces exemples mentionnent des noms fictifs de personnes, de sociétés, de marques ou de produits à des fins illustratives ou explicatives uniquement. Tous ces noms sont fictifs, et toute ressemblance avec des noms de personnes ou de sociétés réelles serait purement fortuite.

Toute copie totale ou partielle de ces programmes exemples et des oeuvres qui en sont dérivées doit comprendre une notice de copyright, libellée comme suit :

Marques

IBM, le logo IBM et ibm.com sont des marques d'International Business Machines Corp. dans de nombreux pays. Les autres noms de produits et de services peuvent être des marques d'IBM ou appartenir à des tiers. La liste actualisée de toutes les marques d'IBM est disponible sur la page Web "Copyright and trademark information" à l'adresse www.ibm.com/legal/copytrade.shtml.

Adobe, le logo Adobe, PostScript et le logo PostScript sont des marques d'Adobe Systems Incorporated aux Etats-Unis et/ou dans certains autres pays.

IT Infrastructure Library est une marque de The Central Computer and Telecommunications Agency qui fait désormais partie de The Office of Government Commerce.

Intel, le logo Intel, Intel Inside, le logo Intel Inside, Intel Centrino, le logo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium, et Pentium sont des marques d'Intel Corporation ou de ses filiales aux Etats-Unis et dans certains autres pays.

Linux est une marque de Linus Torvalds aux Etats-Unis et/ou dans certains autres pays.

Microsoft, Windows, Windows NT et le logo Windows sont des marques de Microsoft Corporation aux Etats-Unis et/ou dans certains autres pays.

ITIL est une marque de The Minister for the Cabinet Office et est enregistrée au bureau américain Patent and Trademark Office.

UNIX est une marque enregistrée de The Open Group aux Etats-Unis et/ou dans certains autres pays.

Cell Broadband Engine est une marque de Sony Computer Entertainment, Inc., aux Etats-Unis et/ou dans certains autres pays, et est utilisée sous license.

Linear Tape-Open, LTO, le logo LTO, Ultrium et le logo Ultrium sont des marques de HP, IBM Corp. et Quantum aux Etats-Unis et/ou dans certains autres pays.