IBM SPSS Analytic Server
Version 3.0.1

Guide d'utilisation

Important
Avant d'utiliser le présent document et le produit associé, prenez connaissance des informations générales figurant à la section «Remarques», à la page 33.

Informations sur le produit

Cette édition s'applique à la version 3.0.1 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. Console Analytic Server

Analytic Server fournit une interface de client léger pour gérer les sources de données et les projets.

Connexion

1. Entrez l'URL d'Analytic Server dans la barre d'adresse de votre navigateur. Cette URL peut être obtenue auprès de l'administrateur de votre serveur.
2. Entrez le nom d'utilisateur avec lequel vous allez vous connecter au serveur.
3. Entrez le mot de passe associé au nom d'utilisateur spécifié.

Après la connexion, la page d'accueil de la console s'affiche.

Navigation dans la console

- L'en-tête affiche le nom du produit, le nom de l'utilisateur actuellement connecté, et un lien vers le système d'aide. Le nom de l'utilisateur actuellement connecté figure à la tête d'une liste déroulante incluant le lien de déconnexion.
- La zone de contenu affiche les actions que vous pouvez effectuer depuis la page d'accueil de la console.

Sources de données

Une source de données est une collection d'enregistrements, plus un modèle de données définissant un jeu de données pour analyse. La source des enregistrements peut être un fichier (texte délimité, texte à largeur fixe, Excel) sur HDFS, une base de données relationnelle ou un contenu HCatalog ou Geospatial. Le modèle de données définit toutes les métadonnées (noms de zones, stockage, niveau de mesure, etc) nécessaires pour l'analyse des données. Les propriétaires des sources de données peuvent accorder ou restreindre l'accès aux sources de données.

Liste de sources de données

La page principale Data sources fournit une liste des sources de données dont l'utilisateur actuel est membre.

- Cliquez sur un nom de source de données pour afficher ses détails et modifier ses propriétés.
- Renseignez la zone de recherche afin de filtrer la liste en n'affichant que les sources de données dont le nom contient la chaîne de recherche.
- Cliquez sur **New** pour créer une source de données avec le nom et le type de contenu que vous spécifiez dans la boîte de dialogue **Add new data source**.
  - Voir «Règles de dénomination», à la page 23 pour examiner les restrictions s'appliquant aux noms de source de données.
  - Les types de contenus disponibles sont File, Database, HCatalog et Geospatial.

    Remarques :

    - L'option HCatalog est uniquement disponible si Analytic Server a été configuré pour fonctionner avec ces sources de données.
    - Une fois sélectionné, le type de contenu ne peut pas être édité.
    - Vous pouvez importer ou exporter plusieurs sources de données en une même opération.
- Cliquez sur **Delete** pour supprimer la source de données. Cette action laisse tous les fichiers associés à la source de données intacts.
- Cliquez sur **Refresh** pour mettre à jour la liste.

- La liste déroulante **Actions** réalise l'opération sélectionnée.
  1. Sélectionnez l'option d'exportation pour créer une archive à partir des sources de données sélectionnées et enregistrez l'archive sur le système de fichiers local. L'archive comprend tous les fichiers ajoutés aux sources de données sélectionnés en mode **Projects** ou **Data source**.

     Remarque : Lorsqu'une seule source de données est sélectionnée, le nom du fichier d'archive est celui de la source de données sélectionnée. Si plusieurs sources de données sont sélectionnées, le fichier d'archive a par défaut le nom `datasources.zip`.
  2. Sélectionnez l'option d'importation pour importer des archives créées par l'action d'exportation.

     Remarque : Les fichiers d'archive incluant des informations provenant de plusieurs sources de données ne peuvent pas être importés. Il est alors nécessaire d'extraire tout d'abord de l'archive `datasources.zip` les archives individuelles de source de données.
  3. Sélectionnez **Duplicate** pour créer une copie de la source de données.

Détails d'une source de données spécifiques

La zone de contenu est divisée en plusieurs sections qui peuvent dépendre du type de contenu de la source de données.

Details

Ces paramètres sont communs à tous les types de contenu.

Name Zone de texte modifiable qui indique le nom de la source de données.

Display name
Zone de texte modifiable indiquant le nom de la source de données tel qu'affiché dans d'autres applications. Si cette zone est vide, la valeur Name est utilisée comme nom d'affichage.

Description
Zone de texte éditable permettant d'ajouter un texte explicatif sur la source de données.

Is public
Case à cocher indiquant si tout le monde peut voir la source de données (case cochée) ou si les utilisateurs et les groupes doivent être explicitement ajoutés en tant que membres (case désélectionnée).

Custom attributes
Les applications peuvent associer des propriétés à des sources de données (par exemple, pour indiquer s'il s'agit d'une source de données temporaire) par le biais d'attributs personnalisés. Ces attributs sont exposés dans la console Analytic Server pour fournir plus d'informations sur la manière dont les applications utilisent les sources de données.

Cliquez sur **Save** pour conserver l'état actuel des paramètres.

Sharing

Ces paramètres sont communs à tous les types de contenu.

Vous pouvez partager la propriété d'une source de données en ajoutant des utilisateurs et des groupes en tant qu'auteurs ou lecteurs.

- Les entrées saisies dans le champ de texte filtrent les utilisateurs et groupes dont le nom coïncide avec la chaîne de recherche. Sélectionnez **Author** ou **Reader** dans la liste déroulante pour affecter leur rôle au sein de la source de données. Cliquez sur **Add member** pour les ajouter à la liste des membres.
- Pour supprimer un participant, sélectionnez un utilisateur ou groupe dans la liste des membres et cliquez sur **Remove member**.

Remarque : Les utilisateurs qui disposent du rôle **Administrator** ont un accès en lecture et en écriture à toutes les sources de données, qu'ils soient ou non mentionnés spécifiquement comme membres de celles-ci.

File Input

Paramètres permettant de définir des sources de données avec un type de contenu de fichier.

File Viewer

Affiche les fichiers disponibles pour leur inclusion dans la source de données. Sélectionnez le mode **Projects** pour visualiser les fichiers dans la structure de projet Analytic Server, **Data source** pour visualiser les fichiers hébergés dans une source de données, ou **File system** pour afficher le système de fichiers (généralement HDFS). Vous pouvez naviguer dans l'une ou l'autre des structures de dossiers, mais HDFS n'est pas éditable du tout. En d'autres termes, vous ne pouvez pas ajouter des fichiers, créer des dossiers ou supprimer des éléments au niveau racine du mode **Projects**, sauf dans les projets définis. Pour créer, modifier ou supprimer un projet, sélectionnez le mode Projects.

- Cliquez sur **Upload** pour charger un fichier vers la source de données en cours ou le projet/sous-dossier. Vous pouvez rechercher et sélectionnez plusieurs fichiers dans un même répertoire.

  Remarque : Les fichiers sont transférés dans le système de fichiers distribué. Vous pouvez trouver des fichiers transférés dans la structure du répertoire `/analytic-root`, sous le titulaire, la source de données ou le projet (en fonction du mode choisi) et le sous-dossiers appropriés. Par exemple, si vous :

  1. Vous connectez à un titulaire `ibm`
  2. Créez une source de données appelée `fraudDetection`
  3. Sélectionnez le mode **Source de données**
  4. Créez un sous-dossier appelé `historicalData`
  5. Transférez un fichier `charges2015.csv`

  Le fichier peut alors être localisé dans le système de fichiers distribué dans `/analytic-root/ibm/.datasource/fraudDetection/historicalData/charges2015.csv`. D'autre part, si vous :

  1. Vous connectez à un titulaire `ibm`
  2. Créez une source de données appelée `fraudDetection`
  3. Sélectionnez le mode **Projet**
  4. Sélectionnez un projet existant appelé `creditProcessing`
  5. Créez un sous-dossier appelé `historicalData`
  6. Transférez un fichier `charges2015.csv`

  Le fichier peut alors être localisé dans le système de fichiers distribué dans `/analytic-root/ibm/creditProcessing/historicalData/charges2015.csv`.
- Cliquez sur **New folder** pour créer un dossier sous le dossier en cours, avec le nom indiqué dans la boîte de dialogue New Folder Name.
- Cliquez sur **Download** pour télécharger les fichiers sélectionnés vers le système de fichier local.
- Cliquez sur **Delete** pour retirer les dossiers/fichiers sélectionnés.

Files included in data source definition

Utilisez le bouton de déplacement pour ajouter les fichiers et dossiers sélectionnés à la source de données ou les en retirer. Pour chaque fichier ou dossier sélectionné dans la source de données, cliquez sur Settings pour définir les spécifications pour la lecture du fichier.

Lorsque plusieurs fichiers sont inclus dans une source de données, ils doivent partager une métadonnée commune ; en d'autres termes, chaque fichier doit avoir le même nombre de zones, les zones doivent être analysées dans le même ordre dans chaque fichier et chaque zone doit avoir le même stockage dans tous les fichiers. Des non concordances entre les fichiers peuvent entraîner l'échec de la création par la console de Preview and Metadata ou l'interprétation de valeurs valides comme non valides (Null) lors de la lecture du fichier par Analytic Server.

Database Selections

Spécifiez les paramètres de connexion pour la base de données comprenant le contenu de l'enregistrement.

Database
Sélectionnez le type de base de données à laquelle se connecter. Choisissez parmi : DB2, Greenplum, Amazon Redshift, MySQL, Netezza, Oracle, SQL Server, Sybase IQ, TeraData, Hive, DashDB ou BigSQL. Si le type de base de données que vous recherchez n'est pas listé, demandez à votre administrateur de configurer Analytic Server avec le pilote JDBC approprié.

Remarque : Analytic Server prend en charge les bases de données MySQL se trouvant sur des systèmes distants.

Server address
Entrez l'URL du serveur hébergeant la base de données.

Server port
Le numéro de port que la base de données écoute.

Database name
Le nom de la base de données auquel vous souhaitez vous connecter.

Username
Si la base de données est protégée par un mot de passe, entrez votre nom d'utilisateur.

Password
Si la base de données est protégée par un mot de passe, entre votre mot de passe.

Table name
Entrez le nom d'une table de la base de données que vous souhaitez utiliser.

Maximum concurrent reads
Entrez la limite de requêtes parallèles pouvant être envoyées depuis Analytic Server à la base de données pour lecture de la table spécifiée dans la source de données.

HCatalog Selections
Spécifiez les paramètres d'accès aux données gérées sous Apache HCatalog.

Database
Le nom de la base de données HCatalog.

Table name
Entrez le nom d'une table de la base de données que vous souhaitez utiliser.

Filter Le filtre de partition de la table, si la table a été créée comme table partitionnée. Le filtrage HCatalog n'est pris en charge que sur les clés de partition Hive de type chaîne.

Remarque : Il semble que les opérateurs !=, <> et LIKE ne fonctionnent pas avec certaines distributions Hadoop. Ceci est dû à un problème de compatibilité entre HCatalog et ces distributions.

HCatalog Field Mappings
Affiche le mappage d'un élément dans HCatalog à un champ dans la source de données. Cliquez sur Edit pour modifier les mappages de champs.

Remarque : Après la création d'une source de données basée HCatalog qui expose des données d'une table Hive, vous constaterez peut-être que la table Hive est constituée d'un grand nombre de fichiers de données et qu'un délai substantiel s'écoule chaque fois que Analytic Server tente de lire des données depuis la source de données. Si vous êtes confronté à de tels délais, reconstruisez la table Hive en utilisant un plus petit nombre de fichiers de données plus volumineux et limitez le nombre de fichiers à 400, ou moins.

Geospatial Selections
Permet d'indiquer les paramètres d'accès aux données géographiques.

Type Geospatial
Les données géographiques peuvent provenir de service de mappage en ligne ou d'un fichier de forme.

Si vous utilisez un service de mappage, indiquez l'URL du service et sélectionnez la couche de carte à utiliser.

Si vous utilisez un fichier shapefile, vous devez le sélectionner ou le charger. Un fichier shapefile est en réalité un ensemble de fichiers portant le même nom et stockés dans le même répertoire. Sélectionnez le fichier portant le suffixe SHP. Analytic Server recherche et utilise les autres fichiers. Deux fichiers supplémentaires avec les suffixes SHX et DBF doivent toujours être présents. Selon le fichier shapefile, un certain nombre d'autres fichiers sont aussi nécessaires.

Preview and Metadata
Après avoir spécifié les paramètres pour la source de données, cliquez sur Preview and Metadata pour vérifier et confirmer les spécifications de la source de données.

Output
Des sources de données avec type de contenu fichier ou base de données peuvent être ajoutées à la sortie de flux exécutés sur Analytic Server. Sélectionnez **Make writeable** pour autoriser ces ajouts et :

- Pour les sources de données avec base de données comme type de contenu, sélectionnez une table de base de données de sortie dans laquelle les données en sortie peuvent être écrites.
- Pour les sources de données avec fichiers comme type de contenu :
  1. Sélectionnez un dossier de sortie dans lequel écrire les nouveaux fichiers.

     Conseil : Utilisez un dossier distinct pour chaque source de données afin de faciliter le suivi des associations entre fichiers et sources de données.
  2. Sélectionnez un format de fichier ; soit **CSV**, soit **Splittable binary format**.
  3. Vous avez aussi la possibilité de sélectionner l'option **Make sequence file**. Ceci est utile si vous désirez créer des fichiers compressés morcelables pouvant être utilisés par des travaux MapReduce en aval.
  4. Sélectionnez **Newlines can be escaped** si votre sortie est CSV et si des zones de chaîne de caractères contiennent des caractères de retour à la ligne ou de retour chariot. Cette option remplace les caractères de retour à la ligne par une barre oblique inversée suivie de la lettre “n”, les caractères de retour chariot par une barre oblique inversée suivie de la lettre “r”, et les barres obliques inversées par deux barres obliques inversées consécutives. Ces données doivent être lues avec le même paramétrage. Il est fortement conseillé d'utiliser le format binaire Splittable avec les donnée de type chaîne contenant des caractères de retour à la ligne ou de retour chariot.
  5. Sélectionnez un format de compression. La liste inclut tous les formats ayant été configurés pour leur utilisation avec votre installation d'Analytic Server.

Remarque : Certaines combinaisons de format de compression et de format de fichier débouchent sur une sortie qui ne peut pas être fractionnée et qui ne convient donc pas à des traitements MapReduce ultérieurs. Analytic Server génère un avertissement dans la section Output si vous effectuez une telle sélection.

Settings (sources de données de fichier)

Cette boîte de dialogue vous permet de définir les spécifications pour la lecture de données basées fichiers. Ces paramètres s'appliquent à tous les fichiers sélectionnés et à tous les fichiers dans les dossiers sélectionnés conformes aux critères spécifiés dans l'onglet **Folder**.

Des paramètres d'analyseur incorrects pour un fichier peuvent entraîner l'échec de la création par la console de l'élément Preview and Metadata, ou l'interprétation de valeurs valides comme non valides (Null) lors de la lecture du fichier par Analytic Server.

Onglet Settings

Cet onglet vous permet de spécifier le type de fichier et les paramètres d'analyseur spécifiques à ce type de fichier.

Vous pouvez définir des sources de données en utilisant des fichiers compressés sous n'importe quel format de fichier pris en charge. Les formats de compression pris en charge incluent Gzip, Deflate, Bz2, Snappy et IBM CMX.

Type de fichier délimité

Les fichiers délimités sont des fichiers texte avec contenu de champ libre dont les enregistrements contiennent un nombre constant de champs mais un nombre varié de caractères par champ. Les fichiers délimités ont généralement des extensions de fichier *.csv ou *.tab. Pour plus d'informations, voir «Paramètres de type de fichier délimité», à la page 7.

Type de fichier fixe

Les fichiers texte à zone fixe sont des fichiers dont les champs ne sont pas délimités mais qui commencent à la même position et ont une longueur fixe. Ces fichiers sont généralement associés à une extension de fichier *.dat. Pour plus d'informations, voir «Paramètres de type de fichier fixe», à la page 8.

Type de fichier semi-structuré

Les fichiers semi-structurés (tels que *.log) sont des fichiers texte avec une structure prévisible pouvant être mappée à des champs via des expressions régulières, mais moins structurés que les fichiers délimités. Pour plus d'informations, voir «Paramètres de type de fichier semi-structuré», à la page 9.

Type de fichier Text Analytics

Les fichiers Text Analytics sont des documents (tels que *.doc, *.pdf ou *.txt) qui peuvent être analysés à l'aide de SPSS Text Analytics.

Skip empty lines
Indique si les lignes vides dans le contenu texte extrait doivent être ignorées. Valeur par défaut : **No**.

Line separator
Spécifie la chaîne définissant une nouvelle ligne. Par défaut, il s'agit du caractère de retour à la ligne "\n".

Type de fichier SPSS Statistics

Les fichiers SPSS Statistics (*.sav, *.zsav) sont des fichiers binaires qui contiennent un modèle de données. Aucun autre paramètre n'est requis dans l'onglet Settings pour ce type de fichier.

Type de fichier SBF (Splittable Binary Format)

Spécifie que le fichier est de type binaire divisible (*.asbf). Ce type de fichier peut représenter tous les types de zone Analytic Server (contrairement à CSV, qui ne peut jamais représenter des zones de liste et nécessite un paramétrage particulier pour gérer les retours à la ligne et les retours chariot imbriqués). Aucun autre paramètre n'est requis dans l'onglet Settings pour ce type de fichier.

Type de fichier séquence

Les fichiers séquence (*.seq) sont des fichiers texte structurés en paires clé/valeur. Ils sont généralement utilisés comme format intermédiaire dans les travaux MapReduce.

Type de fichier Excel

Spécifie que le fichier est du type Microsoft Excel (*.xls, *.xlsx). Pour plus d'informations, voir «Paramètres de type de fichier Excel», à la page 10.

Paramètres de type de fichier délimité :

Vous pouvez spécifier les paramètres suivants pour les types de fichier délimité.

Character set encoding
Codage de caractères du fichier. Sélectionnez ou spécifiez un nom de jeu de caractères Java, tel que "UTF-8", "ISO-8859-2" ou "GB18030". Valeur par défaut : **UTF-8**.

Field delimiters
Un ou plusieurs caractères marquant les contours de zones. Chaque caractère est pris comme délimiteur indépendant. Par exemple, si vous sélectionnez **Comma** et **Tab** (ou sélectionnez **Other** et entrez `,\t`), ceci signifie que soit une virgule, soit une tabulation marque les limites de la zone. Si les caractères de contrôle délimitent les zones, les caractères figurant ici sont traités comme délimiteurs, en plus des caractères de contrôle. La valeur par défaut est "," si des caractère de contrôle ne délimitent pas les zones ; sinon, la valeur par défaut est la chaîne vide.

Control characters delimit fields
Définit si les caractères de contrôle ASCII, à l'exception de LF et CR, sont traités comme délimiteurs de zone. La valeur par défaut est **No**.

First row contains field names
Définit si la première ligne doit être utilisée pour déterminer les noms de zones. La valeur par défaut est **No**.

Number of initial characters to skip
Le nombre de caractères à ignorer en début de fichier. Un entier non négatif. Valeur par défaut : 0.

Merge white space
Définit si l'occurrence de plusieurs espaces et/ou tabulations de manière adjacente doit être traitée comme un délimiteur de zone unique. Elle n'a aucun effet si l'espace et la tabulation ne sont pas des délimiteurs de zone. La valeur par défaut est **Yes**.

End-of-line comment characters
Un ou plusieurs caractères marquant les commentaires de fin de ligne. Le caractère est ignoré, ainsi que tout ce qui le suit dans l'enregistrement. Chaque caractère est interprété comme une marque de commentaire indépendante. Par exemple, "/*" signifie qu'une barre oblique ou un astérisque commencent un commentaire. Il n'est pas possible de définir des marques de

commentaires à plusieurs caractères comme "//". Une chaîne vide signale qu'aucun caractère de commentaire n'est défini. Si défini, les caractères de commentaires sont vérifiés avant que les guillemets soient traités ou que les caractères initiaux à ignorer soient ignorés. La valeur par défaut est la chaîne vide.

Invalid characters
Détermine comment les caractères non valides (séquences d'octets qui ne correspondent pas à des caractères dans le codage) doivent être traités.

Discard
Supprime les séquences d'octets non valides.

Replace with
Remplace chaque séquence d'octets non valide par le caractère indiqué.

Single quotes
Indique le mode de traitement des guillemets simples (apostrophes). La valeur par défaut est **Keep**.

Keep Les guillemets simples n'ont aucune signification particulière et sont traités comme tout autre caractère.

Drop Les guillemets simples sont supprimés, à moins qu'ils apparaissent entre guillemets.

Pair Les guillemets simples sont traités comme guillemets et les caractères entre paires de guillemets simples perdent toute signification particulière (on considère qu'ils apparaissent entre guillemets). Le paramètre **Quotes can be quoted by doubling** détermine si les guillemets simples peuvent eux-mêmes apparaître à l'intérieur des chaînes entre guillemets simples.

Double quotation marks
Spécifie comment traiter les guillemets. La valeur par défaut est **Pair**.

Keep Les guillemets n'ont pas de signification particulière et sont traités comme tous les autres caractères.

Drop Les guillemets sont supprimés à moins de figurer entre apostrophes

Pair Les guillemets sont traités comme des apostrophes et les caractères entre les guillemets perdent leur signification spéciale éventuelle (sont considérés être encadrés par des apostrophes). Le paramètre **Quotes can be quoted by doubling** détermine si des guillemets peuvent figurer dans des chaînes encadrées elles-mêmes par des guillemets.

Quotes can be quoted by doubling
Indique si les guillemets peuvent être représentés dans les chaînes entre guillemets et les apostrophes représentées dans les chaînes entre apostrophes lorsque **Pair** a été sélectionné. Si sa valeur est **Yes**, les guillemets sont assortis d'un caractères d'échappement en les doublant et les apostrophes sont assorties sont assorties d'un caractère d'échappement en les doublant dans les chaînes entre apostrophes. Si la valeur définie est **No**, il n'y a aucun moyen de mettre entre guillemets un guillemet double à l'intérieur d'une chaîne entre guillemets doubles ou un guillemet simple à l'intérieur d'une chaîne entre guillemets simples. La valeur par défaut est **Yes**.

Newlines can be escaped

Indique si l'analyseur interprète une barre oblique inversée suivie par la lettre "n", la lettre "r" ou une autre barre oblique inversée comme un caractère retour à la ligne, un caractère de retour chariot ou une barre oblique inversée, respectivement. Si les retours à la ligne ne sont pas précédés d'un caractère d'échappement, ces séquences de caractères sont lues de manière littérales comme une barre oblique inversée suivie de la lettre "n", et ainsi de suite. Valeur par défaut : **No**.

Paramètres de type de fichier fixe :

Vous pouvez spécifier les paramètres suivants pour les types de fichier fixes.

Character set encoding
Codage de caractères du fichier. Sélectionnez ou spécifiez un nom de jeu de caractères Java, tel que "UTF-8", "ISO-8859-2" ou "GB18030". Valeur par défaut : **UTF-8**.

Invalid characters
Détermine comment les caractères non valides (séquences d'octets qui ne correspondent pas à des caractères dans le codage) doivent être traités.

Discard
Supprime les séquences d'octets non valides.

Replace with
Remplace chaque séquence d'octets non valide par le caractère indiqué.

Longueur de l'enregistrement
Indique comment les enregistrements sont définis. Si **Newline delimited** est sélectionné, les enregistrements sont définis (délimités) par les retours à la ligne, le début du fichier, ou la fin du fichier. Si **Specific length** est sélectionné, les enregistrements sont définis par une longueur d'enregistrements, en octets. Indiquez une valeur positive.

Initial records to skip
Le nombre d'enregistrements à ignorer au début du fichier. Spécifiez un nombre entier non négatif. Valeur par défaut : 0.

Fields Cette section définit les zones du fichier. Cliquez sur **Add Field** et indiquez le nom du champ, la colonne à laquelle commencent les valeurs du champ, et la longueur de ces valeurs. Les colonnes d'un fichier sont numérotées à partir de 0.

Paramètres de type de fichier semi-structuré :

Les paramètres de fichiers semi-structurés consistent de règles pour le mappage du contenu du fichier à des champs.

Rules Table
Les règles individuelles extraient des informations depuis un enregistrement afin de créer un champ ; conjointement, dans la table de règles, elles définissent tous les champs pouvant être extraits de chaque enregistrement dans une source de données.

Les règles de la table sont appliquées selon leur ordre à chaque enregistrement ; si toutes les règles de la table correspondent à l'enregistrement, aucune autre table de règles n'est requise pour traiter l'enregistrement et l'enregistrement suivant est alors traité. Si une règle quelconque de la table ne correspond pas, toutes les valeurs de champ extraites par les règles précédentes sont ignorées ; s'il existe une autre table de règles, ses règles sont appliquées à l'enregistrement. Si aucune table ne correspond à l'enregistrement, la règle de non concordance est appliquée.

Mismatch
Vous pouvez choisir d'ignorer (**Skip**) les enregistrements ne correspondant à aucune des tables de règles ou définir la valeur de toutes les zones de l'enregistrement à **Missing** (valeur Null).

Export Rules
Vous pouvez enregistrer toutes les règles visibles actuellement en vue d'une réutilisation. La table exportée est enregistrée sur le serveur.

Import Rules
Vous pouvez importer une table de règles enregistrée dans la table de règles visible actuellement. Ceci écrasant toutes les règles que vous aviez défini pour cette table, il est préférable de créer une nouvelle table, puis d'importer une table de règles.

Rule Editor

L'éditeur de règles vous permet de créer une règle d'extraction pour un champ unique.

Anonymous capture group
Une règle de capture de champ commence généralement à extraire des données depuis un enregistrement à la position où la règle précédente s'est arrêtée. Lorsque sont présentes des informations exogènes entre deux champs dans une structure de données semi-structurée, il peut s'avérer utile de définir un groupe de capture anonyme afin de positionner l'analyseur au début du champ suivant. Lorsque vous sélectionnez **Anonymous capture group**, les contrôles destinés à l'attribution d'un nom et d'un label au groupe de capture sont désactivés, mais le reste des fonctions de la boîte de dialogue fonctionne normalement.

Field name
Attribuez un nom au champ. Celui-ci sera utilisé pour définir les métadonnées de la source de données. Les noms de champ doivent être uniques au sein d'une table de règles.

Rule name
Vous pouvez entrer un libellé de description pour la règle.

Description
Vous pouvez entrer une description plus longue pour la règle.

Defining a rule
Vous pouvez définir des règles via deux méthodes.

Utilisation de contrôles pour règles d'extraction
Ceci simplifie la création de règles d'extraction.

1. Spécifiez le point auquel commencer à extraire les données de champ ; **Current position** indique de débuter là où la règle précédente s'est arrêtée et **Skip until** de commencer au début de l'enregistrement et d'ignorer tous les caractères jusqu'à ce que celui spécifié dans la zone de saisie soit atteint. Sélectionnez **Include** si vous désirez que les données de champ incluent le caractère à la position de départ.
2. Sélectionnez un groupe de champs dans la liste déroulante **Capture**.
3. Vous pouvez sélectionner le point auquel arrêter l'extraction des données de champ ; **Whitespace** indique d'arrêter lorsque des blancs (tels que des espaces ou des tabulations) sont rencontrés et **At character(s)** d'arrêter à la chaîne spécifiée. Sélectionnez **Include** si vous désirez que les données de champ incluent le caractère à la position d'arrêt.

Définition manuelle de règles d'expression régulière
Sélectionnez cette méthode si vous êtes familier avec la syntaxe de rédaction d'expressions régulières. Entrez une expression régulière dans la zone de texte **Regexp**.

Add Field Capture Group
Permet d'enregistrer l'expression régulière pour son utilisation ultérieure. La groupe de capture enregistré apparaît dans la liste déroulante **Capture**.

L'Editeur de règles présente un aperçu des données extraites depuis le premier enregistrement par cette règle, une fois que toutes les règles précédentes dans la table de règles ont été appliquées.

Paramètres de type de fichier Excel :

Vous pouvez spécifier les paramètres suivants pour les fichiers Excel.

Worksheet selection
Sélectionne la feuille de calcul Excel à utiliser comme source de données. Vous devez spécifier un index numérique (celui de la première feuille de calcul est 0) ou le nom de la feuille de calcul. Par défaut, celle utilisée est la première feuille de calcul.

Data range selection for import.
Vous pouvez importer des données en partant de la première ligne renseignée ou en indiquant un intervalle de cellules explicite.

- **La plage débute à la première ligne non vide**. Repère la première cellule renseignée et l'utilise comme angle supérieur gauche de l'intervalle de données.
- Vous pouvez également spécifier une plage de cellules explicite en indiquant une ligne et colonne. Par exemple, pour spécifier la plage Excel A1:D5, entrez A1 dans la première zone et D5 dans la seconde (ou bien, R1C1 et R5C4). Toutes les lignes de l'intervalle indiqué sont renvoyées, y compris les lignes vides.

First row contains field names
Indique si la première ligne de la plage de cellules sélectionnée contient les noms de zones. Valeur par défaut : **No**.

Stop reading after encountering blank rows
Indique si la lecture doit être arrêtée si plusieurs lignes vides sont rencontrées ou s'il convient de continuer la lecture jusqu'à la fin de la feuille de travail, y-compris des lignes vides. Valeur par défaut : **No**.

Formats

L'onglet Formats vous permet de définir les informations de formatage des champs analysés.

Paramètres de conversion de champ

Trim white space
Supprime les espaces en début et/ou en fin de chaîne. La valeur par défaut est **None**. Les valeurs suivantes sont prises en charge :

None Ne supprime pas les espaces.

Left Supprime les espaces en début de chaîne.

Right Supprime les espaces en fin de chaîne.

Both Supprime les espaces en début et en fin de chaîne.

Locale Définit les paramètres régionaux. Les paramètres régionaux du serveur sont sélectionnés par défaut. La chaîne des paramètres régionaux doit être spécifiée comme suit : <langue>[_pays[_variante]], où :

langue
Code valide, composé de deux lettres en minuscules, tel que défini par la norme ISO-639.

pays Code valide, composé de deux lettres en majuscules, tel que défini par la norme ISO-3166.

variante
Un code fournisseur ou navigateur spécifique.

Decimal separator
Définit le caractère utilisé comme signe décimal. Reçoit par défaut celui spécifique à l'environnement local.

Grouping symbols
Définit si le caractère local doit ou non être utilisé pour le séparateur de milliers.

Default date format
Définit un format de date par défaut. Tous les schémas de format définis par la spécification Unicode LDML (Locale Data Markup Language) sont pris en charge.

Default time format
Définit un format d'heure par défaut.

Default timestamp
Définit un format d'horodatage par défaut.

Default time zone
Définit le fuseau horaire. Valeur par défaut : UTC. Ce paramètre s'applique aux zones d'heure et d'horodatage pour lesquelles un fuseau horaire spécifique n'a pas été indiqué.

Field Overrides

Cette section vous permet d'affecter des instructions de formatage à des champs individuels. Sélectionnez un champ dans le modèle de données ou entrez un nom de champ, et cliquez sur **Add** pour l'ajouter à la liste de champs dotés d'instructions individuelles. Cliquez sur **Remove** pour le retirer de la liste. Vous pouvez définir les propriétés suivantes pour le champ sélectionné dans la liste.

Storage
Définit le stockage du champ.

Decimal separator
Pour les champs de stockage réel, définit le caractère utilisé comme signe décimal. Reçoit par défaut celui spécifique à l'environnement local.

Grouping symbols
Pour les champs avec stockage Entier ou réel, indique si le caractère utilisé comme séparateur de milliers spécifique à l'environnement local doit être utilisé.

Formats
Pour les champs avec stockage Date, Heure ou Horodatage, définit le format. Sélectionnez un format dans la liste déroulante.

Onglet Field Order

Pour les fichiers de type délimité et Excel, cet onglet vous permet de définir l'ordre d'analyse des zones du fichiers. Ceci est important quand une source de données comprend plusieurs fichiers vu que l'ordre concret des zones peut être différent entre les fichiers mais que l'ordre d'analyse des zones doit être identique pour créer un modèle de données cohérent.

Pour les fichiers de type fixe ou semi-structuré, cet ordre est défini dans l'onglet Settings.

Lorsque la source de données comporte un seul fichier, ou que tous les fichiers ont le même ordre de zones, vous pouvez utiliser l'option par défaut, **Field order matches data model**. Si la source de données comporte plusieurs fichiers et que l'ordre des zones ne correspond pas dans les fichiers, utilisez l'option **Specific field order** pour l'analyse des fichiers.

1. Pour ajouter une zone à la liste ordonnée, entrez le nom de zone ou sélectionnez-le dans la liste fournie par le modèle de données. Vous pouvez ajouter simultanément toutes les zones du modèle de données en cliquant sur **Add all**. Les noms de zone ne seront ajoutés qu'une seule fois à la liste ordonnée.
2. Utilisez les boutons fléchés pour organiser les zones comme souhaité.

Lorsque **Specific field order** est utilisé, les zones n'ayant pas été ajoutées à la liste ne font pas partie de l'ensemble de résultats pour ce fichier. S'il existe dans le modèle de données des zones qui ne sont pas répertoriées dans cette boîte de dialogue, leurs valeurs sont Null dans l'ensemble de résultats.

Onglet Folder

Lorsque vous spécifiez les paramètres de l'analyseur pour un dossier, l'onglet Folder vous permet de sélectionner les fichiers à inclure dans la source de données.

Match all files in the selected folder
La source de données inclut tous les fichiers au premier niveau du dossier ; les fichiers des sous-dossiers ne sont pas inclus.

Match files using a regular expression
La source de données inclut tous les fichiers au premier niveau du dossier qui correspondent à l'expression régulière spécifiée ; les fichiers des sous-dossiers ne sont pas inclus.

Match files using a Unix globbing expression (potentiellement récursive)
La source de données inclut tous les fichiers correspondants à l'expression d'expansion de nom de fichier Unix ; l'expression peut inclure des fichiers situés dans des sous-dossiers du dossier sélectionné.

HCatalog Field Mappings

HCatalog Schema
Affiche la structure de la table spécifiée. HCatalog peut prendre en charge un jeu de données hautement structuré. Pour définir une source de données Analytic Server sur ce type de données, la structure doit être aplatie sous forme de lignes et de colonnes simples. Sélectionnez un élément dans le schéma et cliquez sur le bouton de déplacement pour le mapper à une zone pour analyse.

Tous les noeuds d'arborescence ne peuvent pas être mappés. Par exemple, un tableau ou une mappe de types complexes sont considérés comme un "parent" et ne peuvent pas être mappés directement ; chaque élément simple dans un tableau ou une mappe HCatalog dit être ajouté séparément. Ces noeuds peuvent être identifiés par l'étiquette de l'arborescence terminant par ...:array:struct, ou ...:map:struct.

Par exemple :

- Pour un tableau d'entiers, vous pouvez affecter une zone à une valeur dans le tableau : bigintarray[45], mais non pas le tableau lui-même : bigintarray
- Pour une mappe, vous pouvez affecter une zone à une valeur dans la mappe : datamap["clé"], mais non pas la mappe elle-même : datamap
- Pour un tableau d'un tableau d'entiers, vous pouvez affecter une zone à une valeur bigintarrayarray[45][2], mais non pas le tableau lui-même, bigintarrayarray[45].

  Par conséquent, lorsque vous affectez une zone à un élément d'un tableau ou d'une mappe, la définition de l'élément doit inclure l'index ou la clé : bigintarray[index] ou bigintmap["clé"].

Field Mappings

HCatalog Element
Cliquez deux fois sur une cellule pour l'éditer. Vous devez éditer la cellule lorsque l'élément HCatalog est une matrice ou une carte. Dans le cas d'une matrice, indiquez l'entier correspondant au membre de la matrice que vous souhaitez mapper à un champ. Dans le cas d'une carte, indiquez une chaîne entre guillemets correspondant à la clé que vous souhaitez mapper à un champ.

Mapping Field
Le champ tel qu'il apparaît dans la source de données Analytic Server. Cliquez deux fois sur une cellule pour l'éditer. Les valeurs dupliquées dans la colonne Mapping Field ne sont pas autorisées et génèrent une erreur.

Storage
Le stockage du champ. Le stockage est dérivé de HCatalog et il ne peut pas être édité.

Remarque : Lorsque vous cliquez sur Preview and Metadata pour finaliser une source de données HCatalog, il n'existe aucune option d'édition.

Raw Data
Affiche les enregistrements tels qu'ils sont stockés dans HCatalog ; ceci peut vous aider à déterminer comment mapper le schéma HCatalog à des zones.

Remarque : Tout filtrage spécifié dans HCatalog Selections est appliqué à la vue des données brutes.

Utilisation des sources de données HCatalog

Analytic Server prend en charge les sources de données HCatalog. Cette section décrit comment configurer diverses bases de données NoSQL sous-jacentes.

Dans la plupart des cas, vous devez consulter la documentation du fournisseur pour l'intégration Hive.

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

Apache Cassandra
«Apache Cassandra»

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

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

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

Sources de données XML
«Sources de données XML», à la page 16

Apache Cassandra

Analytic Server prend en charge les sources de données HCatalog comportant un contenu sous-jacent dans Apache Cassandra.

Cassandra fournit un magasin clé-valeur structuré. Les clés sont mappées à plusieurs valeurs, lesquelles sont groupées en familles de colonne. Les familles de colonne sont déterminées lorsqu'une base de données est créée, mais des colonnes peuvent être ajoutées à une famille à n'importe quel moment. Par ailleurs, des colonnes ne sont ajoutées qu'aux clés spécifiées et donc des clés différentes peuvent avoir des nombres de colonnes différents dans une famille donnée. Les valeurs d'une famille de colonne pour chaque clé sont conservées ensemble.

Vous pouvez définir des tables Cassandra de deux manières : en utilisant l'interface de ligne de commande Cassandra traditionnelle (cassandra-cli) ou le nouveau shell CQL (csqlsh).

Utilisez la syntaxe suivante pour créer une table Apache Cassandra externe dans Hive si la table a été créée à l'aide de l'interface de ligne de commande traditionnelle (CLI).

```
CREATE EXTERNAL TABLE <nom_table_hive>  (<spécifications colonne>)
STORED BY 'org.apache.hadoop.hive.cassandra.CassandraStorageHandler'
WITH SERDEPROPERTIES("cassandra.cf.name" = "<famille_colonne_cassandra>",
"cassandra.host"="<hôte_cassandra>","cassandra.port" = "<port_cassandra>")
TBLPROPERTIES ("cassandra.ks.name" = "<cassandra_keyspace>");
```

Par exemple, pour la définition de table suivante via l'interface de ligne de commande :

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

create column family users with comparator = UTF8Type;

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

```
assume users keys as utf8;

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

get users['jdoe'];
```

... la DDL de table Hive sera similaire à ceci :

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

Utilisez la syntaxe suivante pour créer une table Apache Cassandra externe dans Hive si la table a été créée via CQL.

```
CREATE EXTERNAL TABLE <nom_table_hive>  (<spécifications colonne>)
STORED BY 'org.apache.hadoop.hive.cassandra.cql.CassandraCqlStorageHandler'
WITH SERDEPROPERTIES("cassandra.cf.name" = "<famille_colonne_cassandra>",
"cassandra.host"="<hôte_cassandra>","cassandra.port" = "<port_cassandra>")
TBLPROPERTIES ("cassandra.ks.name" = "<cassandra_keyspace>");
```

Par exemple, pour la définition de table CQL3 suivante :

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

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

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

... la DDL de la table Hive sera comme la suivante :

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

Sources de données XML

Analytic Server prend en charge les données XML via HCatalog.

Exemple

1. Mappez le schéma XML au types de données Hive via la DDL (Data Definition Language) Hive, d'après les règles suivantes.

```
CREATE [EXTERNAL] TABLE <nom_table> (<spécifications_colonne>)
ROW FORMAT SERDE "com.ibm.spss.hive.serde2.xml.XmlSerDe"
WITH SERDEPROPERTIES (
    ["xml.processor.class"="<nom_classe_processeur_xml>",]
    "column.xpath.<nom_colonne>"="<requête_xpath>",
     ...
    ["xml.map.specification.<nom_élément>"="<spécification_mappe>"
     ...
    ]
)
STORED AS
    INPUTFORMAT "com.ibm.spss.hive.serde2.xml.XmlInputFormat"
    OUTPUTFORMAT "org.apache.hadoop.hive.ql.io.IgnoreKeyTextOutputFormat"
[LOCATION "<emplacement_données>"]
TBLPROPERTIES (
    "xmlinput.start"="<balise_début ",
    "xmlinput.end"="<balise_fin>"
);
```

Remarque : Si vos fichiers XML utilisent la compression Bz2, `INPUTFORMAT` doit être défini à `com.ibm.spss.hive.serde2.xml.SplittableXmlInputFormat`. S'ils utilisent la compression CMX, il doit être défini à to `com.ibm.spss.hive.serde2.xml.CmxXmlInputFormat`.

Par exemple, le code XML suivant :

```
<records>
  <record customer_id="0000-JTALA">
    <demographics>
      <gender>F</gender>
      <agecat>1</agecat>
      <edcat>1</edcat>
      <jobcat>2</jobcat>
      <empcat>2</empcat>
      <retire>0</retire>
      <jobsat>1</jobsat>
      <marital>1</marital>
      <spousedcat>1</spousedcat>
      <residecat>4</residecat>
      <homeown>0</homeown>
      <hometype>2</hometype>
      <addresscat>2</addresscat>
    </demographics>
    <financial>
      <income>18</income>
      <creddebt>1.003392</creddebt>
      <othdebt>2.740608</othdebt>
      <default>0</default>
    </financial>
  </record>
</records>
```

...serait représenté par la DDL Hive suivante.

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

```
TBLPROPERTIES (
    "xmlinput.start"="<record customer",
    "xmlinput.end"="</record>"
);
```

Pour plus d'informations, voir «Mappage des types de données XML vers Hive».

2. Créez une source de données Analytic Server avec type de contenu HCatalog dans la console Analytic Server.

Limitations

- Seule la spécification XPath 1.0 est actuellement prise en charge.
- La partie locale des noms qualifiés pour les éléments et attributs est utilisée lors du traitement des noms de zone Hive. Les préfixes d'espace de nom sont ignorés.

Mappage des types de données XML vers Hive : Les données modélisées en XML peuvent être transformées en types de données Hive à l'aide des conventions documentées ci-dessous.

Structures

L'élément XML peut être mappé directement au type de structure Hive de sorte que tous les attributs deviennent les membres de données. Le contenu de l'élément devient un membre supplémentaire, de type primitif ou complexe.

données XML

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

DDL Hive et données brutes

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

Tableaux

Les séquences XML d'éléments peuvent être représentées sous forme de tableaux Hive de type primitif ou complexe. L'exemple suivant illustre comment l'utilisateur peut définir un tableau de chaînes à partir du contenu de l'élément XML <result>.

données XML

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

DDL Hive et données brutes

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

Mappes

Le schéma XML n'assure pas un support natif des mappes. Trois approches usuelles s'adressent à la modélisation de mappes en XML. Pour concilier les différentes approches, nous utilisons la syntaxe suivante :

```
"xml.map.specification.<nom_élément>"="<clé>-><valeur>"
```

où

nom_élément
: Nom de l'élément XML à considérer comme une entrée de mappe

clé
: Noeud XML de la clé d'entrée de mappe

valeur
: Noeud XML de la valeur d'entrée de mappe

La spécification de mappe pour l'élément XML indiqué doit être définie sous la section SERDEPROPERTIES du DDL de création de table Hive. Les clés et les valeurs peuvent être définies à l'aide de la syntaxe suivante :

@attribute
La spécification @attribute permet à l'utilisateur d'utiliser la valeur de l'attribut comme clé ou valeur de la mappe.

element
Le nom d'élément peut être utilisé comme clé ou comme valeur.

#content
Le contenu de l'élément peut être utilisé comme clé ou comme valeur. Vu que les clés de mappe ne peuvent être que de type primitif, le contenu complexe sera converti en chaîne.

Les approches pour la représentation de mappes en XML, et le DDL Hive et les données correspondants, sont les suivantes.

Nom d'élément à contenu
Le nom de l'élément est utilisé comme clé et le contenu comme valeur. Il s'agit d'une des techniques usuelles et celle-ci est utilisée par défaut lors du mappage XML vers des types de mappe Hive. La limitation évidente de cette approche est que la clé de mappe ne peut être que du type chaîne.

données XML

```
<entrée1>valeur1</entry1>
<entrée2>valeur2</entry2>
<entrée3>valeur3</entry3>
```

Mappage, DDL Hive et données brutes

Dans ce cas, vous n'avez pas besoin de spécifier un mappage puisque, par défaut, le nom de l'élément est utilisé comme clé et son contenu comme valeur.

```
result map<chaîne,chaîne>
{"result":{"entrée1": "valeur1", "entrée2": "valeur2", "entrée3": "valeur3"}}
```

Attribut à contenu d'élément
Utilisez une valeur d'attribut comme clé et le contenu de l'élément comme valeur.

données XML

```
<nom entrée="clé1">valeur1</entry>
<nom entrée="clé2">valeur2</entry>
<nom entrée="clé3">valeur3</entry>
```

Mappage, DDL Hive et données brutes

```
"xml.map.specification.entry"="@name->#content"
result map<chaîne,chaîne>
{"result":{"clé1": "valeur1", "clé2": "valeur2", "clé3": "valeur3"}}
```

Attribut à attribut

données XML

```
<nom entrée="clé1" value="valeur1"/>
<nom entrée="clé2" value="valeur2"/>
<nom entrée="clé3" value="valeur3"/>
```

Mappage, DDL Hive et données brutes

```
"xml.map.specification.entry"="@name->@value"
result map<chaîne,chaîne>
{"result":{"clé1": "valeur1", "clé2": "valeur2", "clé3": "valeur3"}}
```

Contenu complexe

Le contenu complexe utilisé comme type primitif sera converti en chaîne XML valide en ajoutant un élément racine appelé <string>. Examinez le code xml suivant :

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

L'expression XPath /dataset/* entraîne le renvoi d'un certain nombre de noeuds XML <value>. Si le champ racine est de type primitif, l'implémentation transformera le résultat de la requête en XML valide en ajoutant le noeud racine <string>.

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

Remarque : L'implémentation n'ajoute pas d'élément racine <string> si le résultat de la requête est un élément XML unique.

Contenu de texte

Le contenu texte composé de blancs uniquement d'un élément XML est ignoré.

Preview and Metadata (sources de données)

Lorsque vous cliquez sur **Preview and Metadata**, un échantillon d'enregistrements et le modèle de données pour la source de données sont affichés. Ceci vous permet d'examiner les informations de métadonnées de base.

Preview
L'onglet Preview montre un petit échantillon d'enregistrements et leurs valeurs de zone.

Edit

Cet onglet affiche les métadonnées de zone de base. Pour les sources de données avec type de contenu Fichiers, le modèle de données est généré depuis un petit échantillon d'enregistrements et vous pouvez modifier manuellement les métadonnées de zone depuis cet onglet. Pour les sources de données avec type de contenu HCatalog, le modèle de données est généré d'après les mappages de zones HCatalog et vous ne pouvez pas modifier le stockage de zone depuis cet onglet.

Field Double-cliquez sur le nom de champ pour l'éditer.

Measurement
Il s'agit du niveau de mesure utilisé pour décrire les caractéristiques des données d'un champ spécifique.

Role Permet d'indiquer aux noeuds de modélisation si les champs sont des champs d'entrée (champs prédicteurs) ou de cible (champs prédits) pour un processus d'apprentissage automatique. Les rôles Both et None sont également disponibles, ainsi que l'option Partition qui signale les champs utilisés pour partitionner les enregistrements en échantillons distincts à des fins d'apprentissage, de test et de validation. La valeur Split indique que des modèles séparés seront générés pour chaque valeur possible du champ. La valeur Frequency indique que les valeurs d'une zone doivent être utilisées pour mesurer la fréquence pour chaque enregistrement. La valeur Record ID est utilisée pour identifier un enregistrement dans la sortie.

Storage
Décrit la façon dont les données sont stockées dans un champ. Par exemple, un champ comportant les valeurs 1 et 0 stocke des données d'entiers. Il convient de le différencier du niveau de mesure qui décrit l'utilisation des données et n'affecte pas le stockage. Par exemple, vous pouvez définir le niveau de mesure afin d'indiquer un champ d'entier comportant les valeurs 1 et 0. En général, 1 correspond à la valeur True et 0 à la valeur False.

Valeurs
Affiche les valeurs individuelles des zones avec niveau de mesure qualitatif ou la plage de valeurs pour les zones avec niveau de mesure continu.

Structure
Indique si les enregistrements dans la zone contiennent une valeur unique (Primitive) ou une liste de valeurs.

Depth Indique la profondeur d'une liste ; 0 correspond à une liste de primitives, 1 à une liste de listes, et ainsi de suite.

Scan all Data Values
Permet de déclencher et d'annuler l'analyse des valeurs des données de la sources de données pour déterminer les valeurs de catégories et les limites de plages. Si une analyse est en cours, cliquez sur le bouton pour **annuler l'analyse des données**. L'analyse de toutes les valeurs de données permet de s'assurer que les métadonnées sont correctes, mais peut prendre un certain temps si la source de données comprend de nombreux champs et enregistrements.

Projets

Les projets sont des espaces de travail permettant de stocker les entrées et d'accéder aux sorties de travaux. Ils offrent une structure organisationnelle de niveau supérieur pour le stockage de fichiers et dossiers. Les projets peuvent être partagés avec des utilisateurs et groupes individuels.

Liste des projets

La page Projects principale fournit une liste des projets dont l'utilisateur actuel est membre.

- Cliquez sur un nom de projet pour afficher ses détails et éditer ses propriétés.
- Renseignez la zone de recherche afin de filtrer la liste en n'affichant que les projets dont le nom contient la chaîne de recherche.
- Cliquez sur **New** pour créer un projet avec le nom que vous spécifiez dans la boîte de dialogue **Add new project**. Voir «Règles de dénomination», à la page 23 pour les restrictions sur les noms que vous pouvez affecter aux projets.
- Cliquez sur **Delete** pour supprimer les projets sélectionnés. Cette action supprime le projet et efface de HDFS toutes les données associées au projet.
- Cliquez sur **Refresh** pour mettre à jour la liste.

Détails de projet individuel

La zone de contenu est divisée en sections **Details**, **Sharing**, **Files** et **Versions** condensables.

Details

Name Zone de texte modifiable affichant le nom du projet.

Display name
Zone de texte modifiable indiquant le nom du projet tel qu'affiché dans d'autres applications. Si cette zone est vide, la valeur Name est utilisée comme nom d'affichage.

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

Versions to keep

Supprime automatiquement la plus ancienne version validée du projet une fois que le nombre de versions indiqué a été atteint. Valeur par défaut : 25.

Remarque : Le processus de nettoyage n'est pas immédiat mais s'effectue en arrière-plan toutes les 20 minutes.

Is public

Case à cocher indiquant (si elle est cochée) que tous les utilisateurs peuvent visualiser le projet ou (si elle est décochée) que des utilisateurs ou des groupes doivent être ajoutés explicitement comme membres.

Cliquez sur **Save** pour conserver l'état actuel des paramètres.

Sharing

Vous pouvez partager un projet en ajoutant des utilisateurs et des groupes en tant qu'auteurs ou lecteurs.

- Les entrées saisies dans le champ de texte filtrent les utilisateurs et groupes dont le nom coïncide avec la chaîne de recherche. Sélectionnez le niveau de partage et cliquez sur **Add member** pour l'ajouter à la liste de membres.
  - Les auteurs sont des membres à part entière du projet et peuvent le modifier, tout comme les dossiers et fichiers qu'il contient. Les utilisateurs et membres de ces groupes disposent de droits d'accès en écriture (noeud Analytic Server Export) sur ce projet lorsqu'ils se connectent à Analytic Server via IBM® SPSS Modeler.
  - Les lecteurs peuvent consulter les dossiers et fichiers d'un projet, de même que définir des sources de données pour les objets d'un projet, mais ne peuvent pas le modifier.
- Pour supprimer un auteur, sélectionnez un utilisateur ou un groupe dans la liste des auteurs et cliquez sur **Remove member**.

Remarque : Les administrateurs disposent d'un accès en lecture et en écriture à tous les projets, qu'ils soient ou non mentionnés spécifiquement comme membres de ceux-ci.

Remarque : Les modifications apportées pour le partage sont appliquées immédiatement et automatiquement.

Files

Project structure pane

Le panneau de droite affiche la structure des projets/dossiers du projet actuellement sélectionné. Vous pouvez parcourir la structure des dossiers, mais elle n'est pas éditable, sauf à travers les boutons.

- Cliquez sur **Download file to the local filesystem** pour télécharger un fichier sélectionné sur le système de fichiers local.
- Cliquez sur **Delete the selected file(s)** pour supprimer le fichier/dossier sélectionné.

File Viewer

Affiche la structure de dossiers du projet en cours. Cette structure n'est modifiable que dans les projets définis. C'est-à-dire que vous ne pouvez pas ajouter des fichiers, créer des dossiers ou supprimer des éléments au niveau racine du mode **Projects**. Pour créer ou supprimer un projet, revenez à la liste Projets.

- Cliquez sur **Upload file to HDFS** pour télécharger un fichier vers le projet/sous-dossier actuel.
- Cliquez sur **Create a new folder** pour créer un nouveau dossier sous le dossier en cours, avec le nom que vous spécifiez dans la boîte de dialogue **New folder name**.
- Cliquez sur **Download file to the local filesystem** pour télécharger les fichiers sélectionnés vers le système de fichiers local.

- Cliquez sur **Delete the selected file(s)** pour supprimer les fichiers/dossiers sélectionnés.

Versions

Les versions des projets correspondent aux modifications apportées au contenu des fichiers et dossiers. Les modifications apportées aux attributs d'un projet, telles que la description, le fait ou non d'être public, les personnes avec lesquelles le projet est partagé, ne requièrent pas une nouvelle version. En revanche, l'ajout, la modification ou la suppression de fichiers et dossiers requiert une nouvelle version.

Project versioning table

La table affiche les versions de projets existantes, leur date de création et de validation, les utilisateurs responsables de chaque version et la version parente. La version parente est la version sur laquelle la version sélectionnée est basée.

- Cliquez sur **Lock** pour modifier le contenu de la version de projet sélectionnée.
- Cliquez sur **Commit** pour enregistrer toutes les modifications apportées à un projet et faire correspondre cette version à l'état visible actuel du projet.
- Cliquez sur **Discard** pour annuler toutes les modifications apportées à un projet verrouillé et renvoyer l'état visible du projet à la dernière version validée.
- Cliquez sur **Delete** pour supprimer la version sélectionnée.

Gestion des utilisateurs

Les administrateurs peuvent gérer les rôles des utilisateurs et des groupes via la page Users.

La zone de contenu est divisée en sections **Details** et **Principals** condensables.

Details

Name Zone non modifiable contenant le nom du titulaire.

Description
Zone modifiable vous permettant de fournir un texte explicatif sur le titulaire.

URL URL à communiquer aux utilisateur pour se connecter au titulaire via la console Analytic Server.

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

Utilisateurs et groupes dérivés du fournisseur de sécurité mis en place lors de la configuration. Vous pouvez modifier le rôle des principaux en les désignant comme administrateurs, utilisateurs ou lecteurs.

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 le nombre maximum d'enregistrements autorisés pour une source de données. Cette limite est vérifiée dès lors qu'une source de données est créée ou mise à jour. Par exemple, lorsque vous ajoutez un nouveau fichier ou changez les paramètres d'un fichier.
- Vous pouvez défini 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 dénomination

Les règles de dénomination suivantes s'appliquent à tout élément pouvant recevoir un nom unique dans Analytic Server, comme les sources de données ou les projets.

- Au sein d'un titulaire, les noms des objets de même type doivent être uniques. Par exemple, deux sources de données ne peuvent toutes deux être nommées insuranceClaims, mais une source de données et un projet pourraient tous deux se nommer insuranceClaims.
- Les noms sont sensibles à la casse. Par exemple, insuranceClaims et InsuranceClaims sont considérés comme des noms uniques.
- Les noms ignorent les espaces de début et de fin.
- Les caractères suivants ne sont pas admis dans les noms.

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

Chapitre 2. Intégration du SPSS Modeler

SPSS Modeler est un plan de travail d'exploration de données possédant une approche visuelle à l'analyse. Chaque action distincte d'un travail, qu'il s'agisse de l'accès à une source de données, de la fusion d'enregistrements, de l'écriture d'un nouveau fichier ou de la génération d'un modèle, est représentée par un noeud sur le canevas. Nous relions ces actions ensemble pour former un flux analytique. Pour générer un flux analytique qui s'exécute avec Analytic Server :

1. Le flux doit être démarré avec un noeud Source Analytic Server.
2. Générez la moitié du flux dans l'interface Modeler comme vous le faites normalement, en choisissant les noeuds de processus (Options Zone ou Enregistrement) pris en charge par Analytic Server. Il existe un panneau Analytic Server dans la palette Modeler qui affiche les noeuds pris en charge.
3. Pour terminer le flux, vous disposez de deux options :
   - Choisissez un noeud terminal (Output, Graph, Export ou Modeling) pris en charge par Analytic Server. Dans ce cas, Modeler envoie le flux à Analytic Server. Analytic Server orchestre les travaux nécessaires sur le cluster Hadoop et vous permet de consulter les résultats dans Modeler. Modeler prend les résultats et vous les présente tout comme si le flux avait été traité en local.
   - Si vous choisissez un noeud terminal non pris en charge par Analytic Server, Modeler envoie autant d'élements que possible appartenant au flux à Analytic Server, puis commence à extraire des enregistrements depuis Hadoop. Notez que les modèles ne pouvant pas être générés avec Analytic Server peuvent être évalués par Analytic Server. En d'autres termes, vous pouvez structurer un flux permettant de prendre un sous-échantillon de vos big data valide statistiquement avec Analytic Server, puis générer un modèle "en local" dans Modeler. Le nugget de modèle qui en découle peut alors être inclus dans un flux d'évaluation qui s'exécute entièrement dans Analytic Server.

   Remarque : Vous pouvez définir le nombre maximum d'enregistrements que SPSS Modeler téléchargera depuis Hadoop dans les propriétés de flux d'Analytic Server.

Noeuds pris en charge

Un grand nombre de noeuds SPSS Modeler peut être exécuté sur HDFS, mais il peut exister des différences dans l'exécution de certains noeuds et certains ne sont pas actuellement pris en charge. Cette rubrique décrit le niveau de prise en charge actuel.

Remarque : Voir la documentation SPSS Modeler pour plus d'informations sur l'exécution normale de ces noeuds.

Généralités

- Certains caractères normalement acceptables dans un nom de champ Modeler entre guillemets ne seront pas acceptés par Analytic Server.
- Pour qu'un flux Modeler s'exécute dans Analytic Server, il doit commencer par un ou plusieurs noeuds Analytic Server Source et se terminer par un seul noeud de modélisation ou par un noeud Analytic Server Export.
- Il est recommandé de définir le stockage des cibles continues comme réel, plutôt que comme entier. Les modèles d'évaluation écrivent toujours les valeurs réelles dans les fichiers de données de sortie des cibles continues, alors que le modèle de données de sortie des scores suit le stockage de la cible. Par conséquent, si une cible continue a un stockage d'entier, cela provoquera une non-concordance dans les valeurs écrites et le modèle de données des scores et cette non-concordance causera des erreurs lorsque vous tenterez de lire les données évaluées.

Source

- Un flux ne commençant pas par un noeud de source Analytic Server sera exécuté localement.

Record operations

Toutes ces opérations sont prises en charge, à l'exception des noeuds Streaming TS et Boîtes espace-temps. D'autres remarques sur la fonctionnalité de noeud prise en charge suivent.

Select

- Prend en charge le même ensemble de fonctions que le noeud dériver.

Sample

- L'échantillonnage par blocs n'est pas pris en charge.
- Les méthodes d'échantillonnage complexes ne sont pas prises en charge.
- L'échantillonnage des `n` premiers éléments avec `"Discard sample"` n'est pas pris en charge.
- L'échantillonnage des `n` premiers éléments avec `N>20000` n'est pas pris en charge.
- 1 échantillonnage sur n n'est pas pris en charge lorsque l'option `"Maximum sample size"` n'est pas définie.
- 1 échantillonnage sur n n'est pas pris en charge lorsque l'option `N * "Maximum sample size" > 20000`.
- L'échantillonnage par bloc de pourcentage aléatoire n'est pas pris en charge.
- Le pourcentage aléatoire prend actuellement en charge la mise à disposition d'une valeur de départ.

Aggregate

- Les clés contiguës ne sont pas prises en charge. Si vous réutilisez un flux existant configuré pour trier les données, puis utilisez ce paramètre dans le noeud agrégé, changez ce flux afin de retirer le noeud de tri.
- Les statistiques d'ordre (Valeur médiane, 1er quartile, 3e quartile) sont calculées approximativement et prises en charge via l'onglet Optimisation.

Sort

- L'onglet Optimization n'est pas pris en charge.

Dans un environnement distribué, seul un nombre limité d'opérations conserve l'ordre des enregistrements établi par le noeud Trier.

- Un tri suivi d'un noeud exportation génère une source de données triée.
- Un tri suivi d'un noeud échantillon avec échantillonnage du **Premier** enregistrement renvoie les *N* premiers enregistrements.

En général, vous devez placer un noeud Trier aussi près que possible des opérations nécessitant le tri des enregistrements.

Merge

- Merge by Order n'est pas pris en charge.
- L'onglet Optimization n'est pas pris en charge.
- Les opérations de fusion sont relativement lentes. Si vous disposez d'espace disponible sur HDFS, cela vous prendra probablement beaucoup moins de temps de fusionner vos sources de données une fois et d'utiliser la source fusionnée dans les flux suivants que de fusionner les sources de données dans chaque flux.

Transformation R

La syntaxe R dans le noeud doit être composée d'opérations d'enregistrement unique.

Field operations

Toutes les opérations sur les champs sont prises en charge, à l'exception des noeuds Anonymiser, Transposer, Intervalles de temps et Historique. D'autres remarques sur la fonctionnalité de noeud prise en charge suivent.

Auto Data Prep

- La formation de noeud n'est pas pris en charge. L'application à de nouvelles données des transformations figurant dans un noeud Prép. auto. des données formé est prise en charge.

Derive

- Toutes les fonctions Dériver sont prises en charge à l'exception des fonctions de séquences.
- La dérivation d'une nouvelle zone en tant que comptage est principalement une opération de séquence et n'est donc pas prise en charge.
- Les champs de scission ne peuvent pas être dérivés dans le flux qui les utilise comme scissions. Vous devrez donc créer deux flux : un pour dériver le champ de scission et l'autre pour utiliser le champ comme scission.

Filler

- Prend en charge le même ensemble de fonctions que le noeud dériver.

Regroupement par casiers

La fonction suivante n'est pas prise en charge.

- Création d'intervalles optimale
- Rangs
- Quantiles -> Quantiles : Somme des valeurs
- Quantiles -> Ex-aequo : Conserver dans l'élément actuel et Attribuer aléatoirement
- Quantiles -> N personnalisé : Valeurs supérieures à 100 et toute valeur N où 100 % de N n'est pas égal à zéro.

RFM Analysis

- L'option Conserver dans l'élément actuel pour le traitement des valeurs ex-aequo n'est pas prise en charge. Les scores RFM (récence, fréquence, montant) ne correspondront pas toujours à ceux calculés par Modeler à partir des mêmes données. Les plages de scores seront les mêmes mais les affectations de scores (numéros BIN) peuvent différer d'un point.

Graphs

Tous les noeuds Graph sont pris en charge.

Modeling

Les noeuds Modeling suivants sont pris en charge : Séries temporelles, TCM, Tree-AS, C&R Tree, Quest, CHAID, Linear, Linear-AS, Neural Net, GLE, LSVM, TwoStep-AS, Arbres aléatoires, STP et Règles d'association. La fonction de ces noeuds est décrite en détail par la suite.

Linear Lors de la création de modèles basés sur des données volumineuses, vous voudrez généralement modifier l'objectif en Très grands jeux de données, ou spécifier des scissions.

- La formation continue des modèles PSM existants n'est pas prise en charge.
- L'objectif Standard model building est uniquement recommandé si les champs de scission sont définis de telle sorte que le nombre d'enregistrements se trouvant dans chaque scission n'est pas trop élevé. Notez que la définition de "trop élevé" dépend de la puissance des noeuds individuels dans votre cluster Hadoop. Assurez-vous cependant que les scissions ne sont pas définies de façon trop fine et qu'il y a assez d'enregistrements pour générer un modèle.
- L'objectif Boosting n'est pas pris en charge.
- L'objectif Bagging n'est pas pris en charge.
- L'objectif Very large datasets n'est pas recommandé lorsque le nombre d'enregistrements est réduit car le modèle ne sera pas généré dans la plupart des cas, ou le modèle généré sera dégradé.

- Automatic Data Preparation n'est pas pris en charge. Cela peut causer des problèmes lorsqu'on tente de générer un modèle basé sur des données comportant un grand nombre de valeurs manquantes. Normalement, celles-ci seraient imputées comme partie de la préparation automatique des données. Une solution de contournement consiste à utiliser un modèle d'arbre ou un réseau de neurones avec le paramètre Avancé pour imputer les valeurs manquantes sélectionnées.
- La statistique d'exactitude n'est pas calculée pour les modèles de scission.

Neural Net

Lors de la création de modèles basés sur des données volumineuses, vous voudrez généralement modifier l'objectif en Très grands jeux de données, ou spécifier des scissions.

- La formation continue des modèles standard ou PSM existants n'est pas prise en charge.
- L'objectif Standard model building est uniquement recommandé si les champs de scission sont définis de telle sorte que le nombre d'enregistrements se trouvant dans chaque scission n'est pas trop élevé. Notez que la définition de "trop élevé" dépend de la puissance des noeuds individuels dans votre cluster Hadoop. Assurez-vous cependant que les scissions ne sont pas définies de façon trop fine et qu'il y a assez d'enregistrements pour générer un modèle.
- L'objectif Boosting n'est pas pris en charge.
- L'objectif Bagging n'est pas pris en charge.
- L'objectif Very large datasets n'est pas recommandé lorsque le nombre d'enregistrements est réduit car le modèle ne sera pas généré dans la plupart des cas, ou le modèle généré sera dégradé.
- Lorsque les données comportent un grand nombre de valeurs manquantes, utilisez le paramètre Avancé pour imputer les valeurs manquantes.
- La statistique d'exactitude n'est pas calculée pour les modèles de scission.

C&R Tree, CHAID et Quest

Lors de la création de modèles basés sur des données volumineuses, vous voudrez généralement modifier l'objectif en Très grands jeux de données, ou spécifier des scissions.

- La formation continue des modèles PSM existants n'est pas prise en charge.
- L'objectif Standard model building est uniquement recommandé si les champs de scission sont définis de telle sorte que le nombre d'enregistrements se trouvant dans chaque scission n'est pas trop élevé. Notez que la définition de "trop élevé" dépend de la puissance des noeuds individuels dans votre cluster Hadoop. Assurez-vous cependant que les scissions ne sont pas définies de façon trop fine et qu'il y a assez d'enregistrements pour générer un modèle.
- L'objectif Boosting n'est pas pris en charge.
- L'objectif Bagging n'est pas pris en charge.
- L'objectif Very large datasets n'est pas recommandé lorsque le nombre d'enregistrements est réduit car le modèle ne sera pas généré dans la plupart des cas, ou le modèle généré sera dégradé.
- Interactive sessions n'est pas pris en charge.
- La statistique d'exactitude n'est pas calculée pour les modèles de scission.
- Lorsqu'un champ de scission est présent, les modèles d'arbre créés localement dans Modeler diffèrent légèrement de ceux qui sont créés par Analytic Server, ce qui produit des différences dans les évaluations. Dans les deux cas, les algorithmes sont valides. Ceux qui sont utilisés par Analytic Server sont simplement plus récents. Etant donné que les algorithmes d'arbre ont tendance à contenir de nombreuses règles heuristiques, la différence entre les deux composants est normale.

Model scoring

Tous les modèles pris en charge pour la modélisation sont également pris en charge pour l'évaluation. En outre, les nuggets de modèle construits localement pour les noeuds suivants sont pris en charge pour le scoring : C&RT, Quest, CHAID, Linéaire et Réseau de neurones (qu'il s'agisse de modèle standard, boosted bagged, ou de jeux de données très volumineux), Régression, C5.0, Logistique, Genlin, GLMM, Cox, SVM, Bayes Net, TwoStep, KNN, Liste de décision, Discriminant, Auto-apprentissage, Détection d'anomalies, Apriori, Carma, K-Moyennes, Kohonen, R, et Exploration de texte.

- Aucune propension brute ou ajustée ne sera évaluée. Comme solution de contournement, vous pouvez obtenir le même effet en calculant manuellement la propension brute à l'aide d'un noeud Dériver avec l'expression suivante : if 'predicted-value' == 'value-of-interest' then 'prob-of-that-value' else 1-'prob-of-that-value' endif

R La syntaxe R dans le nugget doit être composée d'opérations d'enregistrement unique.

Output

Les noeuds Matrice, Analyse, Audit des données, Transformer, Valeurs globales, Statistiques, Moyennes et Table sont pris en charge. D'autres remarques sur la fonctionnalité de noeud prise en charge suivent.

Data Audit

Le noeud Data Audit ne peut pas produire le mode des champs continus.

Means

Le noeud Means ne peut pas produire d'erreur standard ni d'intervalle de confiance de 95 %.

Table Le noeud Table est pris en charge par l'écriture d'une source de données Analytic Server temporaire contenant les résultats des opérations en amont. Le noeud Table interroge alors le contenu de cette source de données.

Export Un flux peut commencer avec un noeud de source Analytic Server et terminer avec un noeud d'exportation autre que le noeud d'exportation Analytic Server, mais les données seront déplacées de SPSS Modeler Server, et finalement vers le lieu d'exportation.

Meilleures pratiques

Déplacement vers HCatalog/Hive

Lorsque vous travaillez avec des données dans un tableau partitionné Hive, vous pouvez structurer votre flux Modeler afin de déplacer les partitions souhaitées vers Hive.

1. Démarrez votre flux à l'aide d'un noeud source Analytic Server faisant référence à la source de données HCatalog/Hive.
2. Connectez-vous à un noeud Select permettant de sélectionner des enregistrements UNIQUEMENT pour les zones utilisées comme champs de partition dans le tableau Hive. Si les zones non utilisées en tant que champs de partition sont référencées dans l'expression de ce noeud Select, alors le flux ne sera pas déplacé vers HCatalog/Hive.
3. Connectez-vous à d'autres noeuds comme vous le faites normalement.

Chapitre 3. Traitement des incidents

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

Sources de données

Les filtres définis dans les colonnes partitionnées des sources de données HCatalog ne sont pas pris en compte.

Ce problème affecte certaines versions de Hive, dans les situations suivantes.

- Si vous définissez une source de données HCatalog en spécifiant un filtre dans la définition.
- Si vous créez un flux Modeler avec un noeud Filtre qui fait référence à la colonne de table partitionnée.

La solution palliative consiste à ajouter au flux Modeler un noeud Derive qui crée une nouvelle zone avec des valeurs égales à la colonne partitionnée. Le noeud Filter doit référencer cette nouvelle zone.

Oracle NoSQL

Des erreurs d'échec d'exécution ont été détectées lors de la connexion à une source de données Oracle NoSQL

Ces erreurs sont dues au fait que le gestionnaire de l'espace de stockage est obsolète. Vous devez utiliser un gestionnaire de l'espace de stockage à jour. Le fichier à jour est disponible dans https://github.com/dvasilen/HiveKVStorageHandler3/raw/HADOOP_2.6-HIVE-1.2.0-KV-3.3.4/release/hive-kv-storage-handler-1.2.0-3.3.4.jar

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

1. Copiez le fichier JAR dans le répertoire `{RACINE_HIVE` de Hive et dans le répertoire `{RACINE_AS}/ae_wlpserver/usr/servers/aeserver/apps/AE_BOOT.war/WEB-INF/lib` d'Analytic Server.
2. Exécutez `{RACINE_AS}/bin/hdfsUpdate.sh` pour propager les modifications vers le système HDFS.
3. Redémarrez Analytic Server pour que les modifications soient appliquées.

Remarque : La classe de gestionnaire de l'espace de stockage `oracle.kv.hadoop.hive.table.TableStorageHandler` est recommandée lorsque la base de données Oracle NoSQL 3.0 est utilisée. Pour la classe, il est requis que les utilisateurs organisent les données avec une métaphore de table.

Remarques

Ces informations ont été développées pour les produits et services offerts en France. Ce document 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 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.

Les informations sur les licences concernant les produits utilisant un jeu de caractères double octet peuvent être obtenues par écrit à l'adresse suivante :

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

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 implicites, 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 éléments figurant sur ces sites Web ne font pas partie des éléments du présent produit IBM et l'utilisation de ces sites relève de votre seule responsabilité.

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 de l'ICA, des Conditions internationales d'utilisation des logiciels IBM ou de tout autre accord équivalent.

Les données de performances et les exemples de clients ne sont présentés qu'à des fins d'illustration. Les performances réelles peuvent varier selon les configurations et les conditions de fonctionnement spécifiques.

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. Elle ne peut recevoir aucune réclamation concernant des produits non IBM. Toute question concernant les performances de produits non IBM doit être adressée aux fournisseurs de ces produits.

Toute instruction relative aux intentions d'IBM pour ses opérations à venir est susceptible d'être modifiée ou annulée sans préavis, et doit être considérée 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. 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. 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 :

© (nom de votre société) (année). Des segments de code sont dérivés des exemples de programmes d'IBM Corp.

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 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 licence.

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.