IBM SPSS Analytic Server
Version 2

Benutzerhandbuch

Hinweis

Vor Verwendung dieser Informationen und des darin beschriebenen Produkts sollten die Informationen unter „Bemerkungen" auf Seite 37 gelesen werden.

Produktinformation

Diese Ausgabe bezieht sich auf Version 2, Release 0, Modifikation 0 von IBM SPSS Analytic Server und alle nachfolgenden Releases und Modifikationen, bis dieser Hinweis in einer Neuausgabe geändert wird.

Diese Veröffentlichung ist eine Übersetzung des Handbuchs
IBM SPSS Analytic Server, Version 2, User's Guide,
herausgegeben von International Business Machines Corporation, USA

Herausgegeben von:
TSC Germany
Kst. 2877
Dezember 2014

Inhaltsverzeichnis

Kapitel 1. Neuerungen für Benutzer in Version 2

Analytic Server-Konsole

Neues Layout
Das Layout wurde geändert, damit der Zugriff auf die Seiten über eine Homepage anstatt über Akkordeons erfolgt.

Datenquellen

- Sie können angepasste Attribute für die Datenquelle definieren und von anderen Anwendungen erstellte angepasste Attribute anzeigen.
- Wenn Sie Metadaten für eine Datenquelle erstellen, können Sie einen Scan aller Datenwerte einleiten, um die Kategorienwerte und Bereichsgrenzwerte zu ermitteln. Das Scannen aller Datenwerte stellt sicher, dass die Metadaten korrekt sind, aber es kann einige Zeit in Anspruch nehmen, wenn die Datenquelle viele Felder und Datensätze enthält.
- Es werden mehr Datenquellentypen unterstützt.

  Dateiinhaltstyp

  Die Unterstützung weiterer Dateiinhaltstypen schließt zusätzliche Einstellungen und Parserformate ein. Sie können außerdem die geparste Reihenfolge von Feldern für jede Datei in einer Datenquelle definieren. Beim Hinzufügen eines Verzeichnisses zu einer Datenquelle können Sie Regeln für die Auswahl von Dateien in dem betreffenden Verzeichnis oder in seinen Unterverzeichnissen angeben.

  Semistrukturierte Dateien

  Hierbei handelt es sich um Dateien, z. B. Webprotokolle, die weniger Struktur aufweisen als eine Textdatei mit Trennzeichen, aber Daten enthalten, die über reguläre Ausdrücke in Datensätze und Felder extrahiert werden können.

  Komprimierte Dateien
  Unterstützte Komprimierungsformate sind unter anderem Gzip, Deflate, Bz2, Snappy und IBM CMX. Zusätzlich werden Sequenzdateien mit allen zuvor genannten Komprimierungsformaten unterstützt.

  Textbasierte Dateien in unterschiedlichen Formaten

  Eine einzelne textbasierte Datenquelle kann jetzt Dokumente mit unterschiedlichen Formaten (PDF, Microsoft Word usw.) für die Textanalyse enthalten.

  SPSS Statistics-Dateien
  SPSS Statistics-Dateien (*.sav, *.zsav) sind Binärdateien, die ein Datenmodell enthalten.

  Aufteilbare Dateien im Binärformat (*.asbf)
  Dieser Dateityp wird manchmal von Analytic Server ausgegeben, beispielsweise wenn die Analyse die Verwendung von Feldern mit Listenwerten erfordert.

  Sequenzdateien
  Sequenzdateien (*.seq) sind Textdateien, die als Schlüssel/Wert-Paare strukturiert sind. Sie werden im Allgemeinen als intermediäres Format in MapReduce-Jobs verwendet.

  Datenbankinhaltstyp
  Sie können Datenquellen für Greenplum, MySQL und Sybase IQ definieren, wenn Analytic Server für die Verwendung dieser Datenquellen konfiguriert wurde.

HCatalog-Inhaltstyp
Sie können Datenquellen für Apache Cassandra, MongoDB und Oracle NoSQL definieren, wenn Analytic Server für die Verwendung dieser Datenquellen konfiguriert wurde.

Typ für georäumliche Inhalte
Sie können Datenquellen für geografische Daten definieren, die Shapefiles oder Online-Kartenservices verwenden.

Analyse

Neue SPSS Modeler-Funktionalität

Zusammenführung
Unterstützung für Zusammenführungen nach Rangordnung wurde hinzugefügt.

Zeitreihen
Unterstützung für die Verarbeitung von Zeitreihen und darüber hinaus für verteiltes Erstellen und Scoring von temporalen kausalen Modellen (TCM – Temporal Causal Model) wurde hinzugefügt. Siehe Knoten für AS-Zeitintervalle, Streaming-TCM und TCM in SPSS Modeler.

Geodaten
Unterstützung für die Verarbeitung von geografischen Koordinatensystemen und darüber hinaus für verteiltes Erstellen und Scoring von georäumlichen Assoziationsregeln (GSAR) und räumlich-zeitlichen Punktprozessmodellen (STP-Modelle) wurde hinzugefügt. Siehe Reprojizierungs-, Assoziationsregel- und STP-Knoten in SPSS Modeler.

Clustering
Unterstützung für verteiltes Erstellen und Scoring von zweistufigen Clustermodellen wurde hinzugefügt. Siehe TwoStep-AS-Knoten in SPSS Modeler.

Verbesserte Unterstützung für vorhandene SPSS Modeler-Funktionalität.

Aggregat
Zeichenfolgefelder können mithilfe von Minimal- und Maximalwerten sowie mithilfe der Anzahl Werte ungleich null aggregiert werden. Auf der Registerkarte **Optimization** werden näherungsweise berechnete Rangwerte (Median, Quartile) für numerische Felder unterstützt.

Zusammenführung
Unterstützung für Zusammenführungen nach Bedingung und für Zusammenführungen nach Schlüsseln ohne Angabe eines Schlüsselfelds wurde hinzugefügt, beispielsweise zum Erstellen eines globalen Mittelwerts.

Ensemblemodellierung

Der Algorithmus für das Erstellen von Ensemblemodellen für Baummodelle, lineare Modelle und Modelle mit neuronalen Netzen wurde verbessert, um Daten, die nicht zufällig über gleich große Blöcke verteilt sind, besser handhaben zu können.

Kapitel 2. Analytic Server-Konsole

Analytic Server stellt eine Thin Client-Schnittstelle für die Verwaltung von Datenquellen und Projekten bereit.

Anmeldung

1. Geben Sie die URL von Analytic Server in die Adressleiste Ihres Browsers ein. Die URL können Sie von Ihrem Serveradministrator erhalten.
2. Geben Sie den Benutzernamen ein, mit dem die Anmeldung am Server erfolgen soll.
3. Geben Sie das Kennwort ein, das dem angegebenen Benutzernamen zugeordnet ist.

Nach der Anmeldung wird die Startseite der Konsole angezeigt.

Navigieren in der Konsole

- In der Kopfzeile werden der Produktname und der Name des zurzeit angemeldeten Benutzers sowie der Link zum Hilfesystem angezeigt. Der Name des zurzeit angemeldeten Benutzers steht an oberster Stelle in einer Dropdown-Liste, die auch den Link für die Abmeldung enthält.
- Im Inhaltsbereich werden die Aktionen angezeigt, die Sie über die Startseite der Konsole ausführen können.

Datenquellen

Eine Datenquelle besteht aus einer Sammlung von Datensätzen und einem Datenmodell, die ein Dataset für die Analyse definieren. Die Quelle von Datensätzen kann eine Datei (Text mit Trennzeichen, Text mit fester Breite, Excel) in HDFS, eine Datenbank oder HCatalog sein. Das Datenmodell definiert alle Metadaten (Feldnamen, Speicher, Messniveau usw.), die für die Analyse von Daten erforderlich sind. Datenquelleneigner können Zugriff auf Datenquellen erteilen oder einschränken.

Datenquellenliste

Die Hauptseite mit den Datenquellen enthält eine Liste mit Datenquellen, deren Mitglied der aktuelle Benutzer ist.

- Klicken Sie auf den Namen einer Datenquelle, um die zugehörigen Details anzuzeigen und die Eigenschaften zu bearbeiten.
- Geben Sie einen Suchbegriff in den Suchbereich ein, um die Liste zu filtern, damit nur Datenquellen angezeigt werden, deren Name den Suchbegriff enthält.
- Klicken Sie auf die Schaltfläche **New**, um eine neue Datenquelle mit dem Namen und dem Inhaltstyp zu erstellen, den Sie im Dialogfeld **Add new data source** angeben.
  - Informationen zu den Einschränkungen bei Namen, die Sie für Datenquellen vergeben können, finden Sie in „Benennungsregeln" auf Seite 29.
  - Die verfügbaren Inhaltstypen sind **File**, **Database**, **HCatalog** und **Geospatial**.

    Anmerkung: Der Typ "HCatalog" ist nur verfügbar, wenn Analytic Server für das Arbeiten mit diesen Datenquellen konfiguriert wurde.

    Anmerkung: Wenn der Inhaltstyp ausgewählt wurde, kann er nicht mehr bearbeitet werden.
- Klicken Sie auf **Delete**, um die Datenquelle zu entfernen. Bei dieser Aktion bleiben alle Dateien, die der Datenquelle zugeordnet sind, intakt.
- Klicken Sie auf **Refresh**, um die Liste zu aktualisieren.

- Über die Dropdown-Liste **Actions** wird die ausgewählte Aktion ausgeführt.
  1. Wählen Sie **Export** aus, um ein Archiv der Datenquelle zu erstellen und es auf dem lokalen Dateisystem zu speichern.
  2. Wählen Sie **Import** aus, um ein von der Exportaktion erstelltes Archiv zu importieren.
  3. Wählen Sie **Duplicate** aus, um eine Kopie der Datenquelle zu erstellen.

Individuelle Datenquellendetails

Der Inhaltsbereich ist in mehrere Abschnitte unterteilt, die vom Inhaltstyp der Datenquelle abhängen können.

Details

Diese Einstellungen sind für alle Inhaltstypen gleich.

Name Ein bearbeitbares Textfeld, in dem der Name der Datenquelle angezeigt wird.

Display name

Ein bearbeitbares Textfeld, in dem der Name der Datenquelle so wie in anderen Anwendungen angezeigt wird. Wenn dieses Feld leer ist, wird der in **Name** angegebene Name als Anzeigename verwendet.

Description

Ein bearbeitbares Textfeld, in dem Sie einen erläuternden Text zur Datenquelle angeben können.

Is public

Ein Kontrollkästchen, das angibt, ob jeder die Datenquelle sehen kann (Kästchen ist ausgewählt) oder ob Benutzer und Gruppen explizit als Mitglieder hinzugefügt werden müssen (Kästchen ist abgewählt).

Custom attributes

Anwendungen können durch Verwendung von angepassten Attributen Eigenschaften an Datenquellen anhängen und dadurch beispielsweise angeben, ob die Datenquelle temporär ist. Diese Attribute werden in der Analytic Server-Konsole verfügbar gemacht, um einen tieferen Einblick zu ermöglichen, wie Anwendungen die Datenquelle verwenden.

Klicken Sie auf **Save**, um den aktuellen Status der Einstellungen zu speichern.

Sharing

Diese Einstellungen sind für alle Inhaltstypen gleich.

Sie können das Eigentumsrecht für eine Datenquelle freigeben, indem Sie Benutzer und Gruppen als Autoren hinzufügen.

- Durch Eingeben eines Suchbegriffs in das Textfeld wird nach Benutzern und Gruppen gefiltert, deren Name den Suchbegriff enthält. Klicken Sie auf **Add member**, um sie der Liste der Autoren hinzuzufügen.
- Zum Entfernen eines Autors wählen Sie einen Benutzer oder eine Gruppe in der Mitgliedsliste aus und klicken Sie auf **Remove member**.

Anmerkung: Administratoren verfügen über Lese- und Schreibzugriff auf jede Datenquelle, unabhängig davon, ob sie namentlich als Mitglied aufgelistet sind.

File Input

Einstellungen, die für die Definition von Datenquellen mit dem Inhaltstyp **File** spezifisch sind.

File Viewer

Hier werden für den Einschluss in die Datenquelle verfügbare Dateien angezeigt. Wählen Sie den Modus **Projects** aus, um Dateien in der Analytic Server-Projektstruktur anzuzeigen, oder **Data source**, um in einer Datenquelle gespeicherte Dateien anzuzeigen, oder

File system, um das Dateisystem anzuzeigen (normalerweise HDFS). Sie können beide Ordnerstrukturen durchsuchen, HDFS ist jedoch in keinem Fall bearbeitbar. Im Modus **Projects** können Sie auf der Stammebene keine Dateien hinzufügen, keine Ordner erstellen und keine Elemente löschen. Dies ist nur innerhalb der definierten Projekte möglich. Verwenden Sie zum Erstellen, Bearbeiten oder Löschen eines Projekts die Option Projects.

- Klicken Sie auf **Upload**, um eine Datei in die aktuelle Datenquelle oder das aktuelle Projekt bzw. den aktuellen Unterordner hochzuladen. Sie können in einem einzelnen Verzeichnis nach mehreren Dateien suchen und mehrere auswählen.
- Klicken Sie auf **New folder**, um unter dem aktuellen Ordner einen neuen Ordner mit dem Namen zu erstellen, den Sie im Dialogfeld **New Folder Name** angeben.
- Klicken Sie auf **Download**, um die ausgewählten Dateien in das lokale Dateisystem herunterzuladen.
- Klicken Sie auf **Delete**, um die ausgewählten Dateien/Order zu entfernen.

Files included in data source definition

Verwenden Sie die Pfeilschaltfläche, um der Datenquelle ausgewählte Dateien oder Ordner hinzuzufügen oder daraus zu entfernen. Klicken Sie für jede ausgewählte Datei bzw. für jeden ausgewählten Ordner in der Datenquelle auf Settings, um die Spezifikationen für das Lesen der Datei zu definieren.

Wenn mehrere Dateien in einer Datenquelle vorhanden sind, müssen sie allgemeine Metadaten gemeinsam nutzen; das heißt, jede Datei muss dieselbe Anzahl Felder aufweisen, die Felder müssen in jeder Datei in derselben Reihenfolge geparst werden und jedes Feld muss über alle Dateien hinweg denselben Speicher belegen. Abweichungen zwischen Dateien können zur Folge haben, dass die Konsole die Vorschau und Metadaten (Preview and Metadata) nicht erstellen kann oder ansonsten gültige Werte als ungültig (null) geparst werden, wenn Analytic Server die Datei liest.

Database Selections

Geben Sie die Verbindungsparameter für die Datenbank an, die den Datensatzinhalt enthält.

Database

Wählen Sie den Datenbanktyp aus, zu dem Sie eine Verbindung herstellen wollen. Folgendes steht zur Auswahl: DB2, Greenplum, MySQL, Netezza, Oracle, SQL Server, Sybase IQ oder TeraData. Wenn der von Ihnen gesuchte Typ nicht aufgeführt ist, bitten Sie Ihren Serveradministrator, Analytic Server mit dem entsprechenden JDBC-Treiber zu konfigurieren.

Server address

Geben Sie die URL des Servers an, auf dem sich die Datenbank befindet.

Server port

Die Nummer des Ports, an dem die Datenbank empfangsbereit ist.

Database name

Der Name der Datenbank, zu der Sie eine Verbindung herstellen wollen.

Username

Wenn die Datenbank kennwortgeschützt ist, geben Sie Ihren Benutzernamen ein.

Password

Wenn die Datenbank kennwortgeschützt ist, geben Sie Ihr Kennwort ein.

Table name

Geben Sie den Namen einer Tabelle aus der Datenbank ein, die Sie verwenden wollen.

Maximum concurrent reads

Geben Sie den Grenzwert für die Anzahl paralleler Abfragen ein, die von Analytic Server zur Datenbank gesendet werden können, um aus der in der Datenquelle angegebenen Tabelle zu lesen.

HCatalog Selections

Geben Sie die Parameter für den Zugriff auf Daten an, die unter Apache HCatalog verwaltet werden.

Database

Der Name der HCatalog-Datenbank.

Table name

Geben Sie den Namen einer Tabelle aus der Datenbank ein, die Sie verwenden wollen.

Filter Der Partitionsfilter für die Tabelle, wenn die Tabelle als partitionierte Tabelle erstellt wurde. HCatalog-Filterung wird nur für Hive-Partitionsschlüssel mit dem Zeichenfolgetyp (string) unterstützt.

Anmerkung: Die Operatoren !=, <> und LIKE scheinen in bestimmten Hadoop-Verteilungen nicht zu funktionieren. Hierbei handelt es sich um ein Kompatibilitätsproblem zwischen HCatalog und den betreffenden Verteilungen.

HCatalog Field Mappings

Zeigt die Zuordnung eines Elements in HCatalog zu einem Feld in der Datenquelle an. Klicken Sie auf Edit, um die Feldzuordnungen zu modifizieren.

Anmerkung: Nach der Erstellung einer HCatalog-basierten Datenquelle, die Daten aus einer Hive-Tabelle bereitstellt, stellen Sie möglicherweise fest, dass Analytic Server das Lesen von Daten aus einer Datenquelle immer dann mit erheblicher Verzögerung beginnt, wenn die Hive-Tabelle aus einer großen Anzahl Dateien erstellt wird. Wenn Sie solche Verzögerungen feststellen, müssen Sie die Hive-Tabelle mit einer kleineren Anzahl von umfangreichen Datendateien erneut erstellen und die Anzahl Dateien dabei auf 400 oder weniger reduzieren.

Geospatial Selections

Geben Sie die Parameter für den Zugriff auf geografische Daten an.

Geospatial type

Die geografischen Daten können aus einem Online-Kartenservice oder einer Shapefile stammen.

Wenn Sie einen Kartenservice verwenden, geben Sie die URL des Service an und wählen Sie den gewünschten Kartenlayer aus.

Wenn Sie eine Shapefile verwenden, laden Sie die Shapefile hoch.

Preview and Metadata

Nachdem Sie die Einstellungen für die Datenquelle angegeben haben, klicken Sie auf Preview and Metadata, um die Datenquellenspezifikationen zu prüfen und zu bestätigen.

Output

Datenquellen mit Datei- oder Datenbankinhaltstyp können über die Ausgabe von Datenströmen angehängt werden, die in Analytic Server ausgeführt werden. Wählen Sie **Make writeable** aus, um das Anhängen zu aktivieren, und gehen Sie dann wie folgt vor:

- Wählen Sie für Datenquellen mit Datenbankinhaltstyp eine Ausgabedatenbanktabelle aus, in die die Ausgabedaten geschrieben werden.
- Für Datenquellen mit Dateiinhaltstyp:
  1. Wählen Sie den Ausgabeordner aus, in den die neuen Dateien geschrieben werden.

     Tipp: Verwenden Sie für jede Datenquelle einen separaten Ordner, damit die Zuordnungen zwischen Dateien und Datenquellen leichter verfolgt werden können.
  2. Wählen Sie ein Dateiformat aus: entweder **CSV** (durch Kommas getrennte Variable) oder **Splittable binary format**.

3. Wählen Sie optional **Make sequence file** aus. Dies ist hilfreich, wenn Sie aufteilbare komprimierte Dateien erstellen wollen, die in nachfolgenden MapReduce-Jobs verwendet werden können.
4. Wählen Sie **Newlines can be escaped** aus, damit Zeilenumbrüche in den Daten als Zeichenfolge "\n" in der Ausgabedatei geschrieben werden und die Zeichenfolge "\n" als "\\ n" in die Ausgabedatei geschrieben wird. Wird diese Option nicht ausgewählt, wird "\n" als "\n" in die Ausgabedatei geschrieben und das Vorliegen eines Zeilenumbruchs wird einen Fehler verursachen.
5. Wählen Sie ein Komprimierungsformat aus. Die Liste enthält alle Formate, die zur Verwendung mit Ihrer Analytic Server-Installation konfiguriert wurden.

   Anmerkung: Manche Kombinationen aus Komprimierungsformat und Dateiformat führen dazu, dass die Ausgabe nicht aufgeteilt werden kann und die Ausgabe daher nicht für die weitere MapReduce-Verarbeitung geeignet ist. Analytic Server gibt eine Warnung im Abschnitt für die Ausgabe aus, wenn Sie eine solche Auswahl treffen.

Einstellungen (Dateidatenquellen)

Im Dialogfeld mit den Einstellungen (Settings) können Sie die Spezifikationen für das Lesen dateibasierter Daten definieren. Die Einstellungen gelten für alle ausgewählten Dateien und für alle Dateien in den ausgewählten Ordnern, die die Kriterien auf der Registerkarte **File selection** erfüllen.

Die Angabe falscher Parsereinstellungen für eine Datei kann zur Folge haben, dass Vorschau und Metadaten von der Konsole nicht erstellt werden können oder ansonsten gültige Werte als ungültig (null) geparst werden, wenn Analytic Server die Datei liest.

Registerkarte "Settings"

Auf der Registerkarte **Settings** können Sie den Dateityp und die für den Dateityp spezifischen Parsereinstellungen angeben.

Sie können Datenquellen mithilfe von komprimierten Dateien für ein beliebiges unterstütztes Dateiformat definieren. Unterstützte Komprimierungsformate sind unter anderem Gzip, Deflate, Bz2, Snappy und IBM CMX.

Typ für Dateien mit Trennzeichen

Dateien mit Trennzeichen sind Textdateien mit freien Feldern, deren Datensätze eine konstante Anzahl von Feldern, aber eine variable Anzahl von Zeichen pro Feld enthalten. Dateien mit Trennzeichen haben normalerweise die Dateierweiterung *.csv oder *.tab. Weitere Informationen finden Sie in „Einstellungen für Dateitypen mit Trennzeichen” auf Seite 8.

Typ für Dateien mit festem Format

Textdateien mit festen Feldern sind Dateien, deren Felder nicht begrenzt sind, aber an derselben Position beginnen und eine feste Länge aufweisen. Textdateien mit festen Feldern haben normalerweise die Dateierweiterung *.dat. Weitere Informationen finden Sie in „Einstellungen für unveränderliche Dateitypen” auf Seite 10.

Typ für semistrukturierte Dateien

Semistrukturierte Dateien (z. B. *.log) sind Textdateien, die eine vorhersehbare Struktur aufweisen, die über reguläre Ausdrücke Feldern zugeordnet werden kann. Diese Dateien sind jedoch nicht in dem hohen Maße strukturiert wie Dateien mit Trennzeichen. Weitere Informationen finden Sie in „Einstellungen für semistrukturierte Dateitypen” auf Seite 10.

Text Analytics-Dateityp

Text Analytics-Dateien sind Dokumente (z. B. *.doc, *.pdf oder *.txt), die mit SPSS Text Analytics analysiert werden können.

Skip empty lines
Gibt an, ob leere Zeilen im extrahierten Textinhalt ignoriert werden sollen. Standardwert ist **No**.

Line separator
Gibt die Zeichenfolge an, mit der eine neue Zeile definiert wird. Standardwert ist das Zeilenvorschubzeichen "\n".

SPSS Statistics-Dateityp

SPSS Statistics-Dateien (*.sav, *.zsav) sind Binärdateien, die ein Datenmodell enthalten. Für diesen Dateityp sind keine weiteren Einstellungen auf der Registerkarte **Settings** erforderlich.

Typ für aufteilbare Binärformatdateien

Gibt an, dass es sich beim Dateityp um eine aufteilbare Datei im Binärformat (*.asbf) handelt. Dieser Dateityp wird manchmal von Analytic Server ausgegeben, beispielsweise wenn die Analyse die Verwendung von Feldern mit Listenwerten erfordert. Für diesen Dateityp sind keine weiteren Einstellungen auf der Registerkarte **Settings** erforderlich.

Typ für Sequenzdateien

Sequenzdateien (*.seq) sind Textdateien, die als Schlüssel/Wert-Paare strukturiert sind. Sie werden im Allgemeinen als intermediäres Format in MapReduce-Jobs verwendet.

Excel-Dateityp

Gibt an, dass es sich bei dem Dateityp um eine Microsoft Excel-Datei (*.xls, *.xlsx) handelt. Weitere Informationen finden Sie in „Einstellungen für Excel-Dateitypen” auf Seite 12.

Einstellungen für Dateitypen mit Trennzeichen:

Sie können die folgenden Einstellungen für Dateitypen mit Trennzeichen angeben.

Character set encoding
Die Zeichencodierung der Datei. Wählen Sie einen Java-Zeichensatznamen wie "UTF-8", "ISO-8859-2", "GB18030" usw. aus oder geben Sie diesen an. Der Standardwert ist **UTF-8**.

Field delimiters
Mindestens ein Zeichen, das Feldgrenzen markiert. Jedes Zeichen wird als unabhängiges Trennzeichen gesehen. Wenn Sie beispielsweise **Comma** und **Tab** auswählen (oder wenn Sie **Other** auswählen und `,\t` eingeben), bedeutet dies, dass entweder ein Komma oder ein Tabulator Feldgrenzen markiert. Wenn Steuerzeichen als Feldtrennzeichen fungieren, werden die hier angegebenen Zeichen zusätzlich zu den Steuerzeichen als Trennzeichen betrachtet. Wenn Steuerzeichen nicht als Feldtrennzeichen dienen, ist "," der Standardwert; andernfalls ist der Standardwert eine leere Zeichenfolge.

Control characters delimit fields
Legt fest, ob ASCII-Steuerzeichen, außer LF und CR, als Feldtrennzeichen betrachtet werden. Standardwert ist **No**.

First row contains field names
Legt fest, ob die erste Zeile für die Festlegung der Feldnamen verwendet werden soll. Standardwert ist **No**.

Number of initial characters to skip
Die Anzahl der Zeichen am Anfang der Datei, die übersprungen werden sollen. Eine nicht negative Ganzzahl. Der Standardwert ist "0" (null).

Merge white space
Legt fest, ob mehrere benachbarte Vorkommen eines Leerzeichens und/oder Tabulators als ein einziges Feldtrennzeichen betrachtet werden. Hat keine Auswirkung, wenn weder das Leerzeichen noch der Tabulator ein Feldtrennzeichen ist. Der Standardwert ist **Yes**.

End-of-line comment characters
Mindestens ein Zeichen, das Zeilenendekommentare markiert. Das Zeichen und alles, was im Datensatz darauf folgt, wird, ignoriert. Jedes Zeichen wird als unabhängige Kommentarmarkierung gesehen. "/*" bedeutet z. B., dass ein Kommentar entweder mit einem Schrägstrich oder einem Stern beginnt. Es ist nicht möglich, Kommentarmarkierungen aus mehreren Zeichen zu definieren, z. B. "//". Die leere Zeichenfolge signalisiert, dass keine Kommentarzeichen definiert sind. Wenn Kommentarzeichen definiert sind, werden diese überprüft, bevor Anführungszeichen verarbeitet oder zu überspringende Zeichen am Anfang übersprungen werden. Der Standardwert ist die leere Zeichenfolge.

Invalid characters
Legt fest, wie ungültige Zeichen (Bytesequenzen, die nicht Zeichen in der Codierung entsprechen) behandelt werden sollen. Eine leere Zeichenfolge gibt an, dass sie gelöscht werden sollen. Eine nicht leere Zeichenfolge (in der Regel ein einzelnes Zeichen) gibt an, dass sie durch den Inhalt der Zeichenfolge ersetzt werden sollen. Der Standardwert ist die leere Zeichenfolge.

Single quotes
Gibt die Verarbeitung von einfachen Anführungszeichen (Hochkommas) an. Der Standardwert ist **Keep**.

Keep Hochkommas haben keine besondere Bedeutung und werden wie jedes andere Zeichen behandelt.

Drop Hochkommas werden gelöscht, wenn sie nicht in Anführungszeichen stehen

Pair Hochkommas werden als Anführungszeichen betrachtet und Zeichen zwischen zwei Hochkommas verlieren ihre besondere Bedeutung (sie werden als Zeichen in Anführungszeichen betrachtet). Ob Hochkommas selbst innerhalb von Zeichenfolgen in Hochkommas vorkommen können, wird durch die Einstellung **Quotes can be quoted by doubling** festgelegt.

Double quotation marks
Gibt die Handhabung von Anführungszeichen an. Der Standardwert ist **Pair**.

Keep Anführungszeichen haben keine besondere Bedeutung und werden wie jedes andere Zeichen behandelt.

Drop Anführungszeichen werden gelöscht, wenn sie nicht in Anführungszeichen stehen

Pair Anführungszeichen werden als Anführungszeichen betrachtet und Zeichen zwischen Paaren von Anführungszeichen verlieren ihre besondere Bedeutung (sie werden als Zeichen in Anführungszeichen betrachtet). Ob Anführungszeichen selbst innerhalb von in Anführungszeichen gesetzten Zeichenfolgen vorkommen können, wird durch die Einstellung **Quotes can be quoted by doubling** festgelegt.

Quotes can be quoted by doubling
Gibt an, ob Anführungszeichen innerhalb von in Anführungszeichen gesetzten Zeichenfolgen dargestellt werden können und ob Hochkommas innerhalb von in Hochkommas gesetzten Zeichenfolgen dargestellt werden können, wenn **Pair** festgelegt ist. Bei Angabe von **Yes** werden Anführungszeichen innerhalb von Zeichenfolgen in Anführungszeichen verdoppelt und Hochkommas innerhalb von Zeichenfolgen in Hochkommas verdoppelt. Bei Angabe von **No** gibt es keine Möglichkeit, ein Anführungszeichen innerhalb einer Zeichenfolge in Anführungszeichen oder ein Hochkomma innerhalb einer Zeichenfolge in Hochkommas zu setzen. Der Standardwert ist **Yes**.

Newlines can be escaped
Gibt an, ob der Parser beim Lesen einer Datei die Zeichenfolge "\n" als Zeilenumbruch interpretiert. Wenn Zeilenumbrüche nicht mit einem Escapezeichen versehen sind, wird "\n" einfach als Zeichenfolge gelesen. Wenn Zeilenumbrüche mit Escapezeichen versehen sind, wird "\n" als ASCII-Zeilenvorschubzeichen und "\\n" als Zeichenfolge "\n" gelesen. Standardwert ist **No**.

Einstellungen für unveränderliche Dateitypen:

Sie können die folgenden Einstellungen für unveränderliche Dateitypen angeben:

Character set encoding
Die Zeichencodierung der Datei. Wählen Sie einen Java-Zeichensatznamen wie "UTF-8", "ISO-8859-2", "GB18030" usw. aus oder geben Sie diesen an. Der Standardwert ist **UTF-8**.

Invalid characters
Legt fest, wie ungültige Zeichen (Bytesequenzen, die nicht Zeichen in der Codierung entsprechen) behandelt werden sollen. Eine leere Zeichenfolge gibt an, dass sie gelöscht werden sollen. Eine nicht leere Zeichenfolge (in der Regel ein einzelnes Zeichen) gibt an, dass sie durch den Inhalt der Zeichenfolge ersetzt werden sollen. Der Standardwert ist die leere Zeichenfolge.

Record length
Gibt an, wie Datensätze definiert werden. Bei Auswahl von **Newline delimited** werden Datensätze durch Zeilenumbrüche, den Dateianfang oder das Dateiende definiert (begrenzt). Bei Auswahl von **Specific length** werden Datensätze durch eine Satzlänge in Byte definiert. Geben Sie einen positiven Wert an.

Initial records to skip
Die Anzahl der Datensätze am Anfang der Datei, die übersprungen werden sollen. Geben Sie eine nicht negative ganze Zahl an. Der Standardwert ist 0.

Fields In diesem Abschnitt werden die Felder in der Datei definiert. Klicken Sie auf **Add Field** und geben Sie den Feldnamen, die Spalte, in denen Feldwerte beginnen, und die Länge der Feldwerte an. Spalten werden in einer Datei mit null beginnend nummeriert.

Einstellungen für semistrukturierte Dateitypen:

Einstellungen für semistrukturierte Dateien bestehen aus Regeln für die Zuordnung des Dateiinhalts zu Feldern.

Rules Table
Einzelne Regeln extrahieren Informationen aus einem Datensatz, um ein Feld zu erstellen. Kombiniert in einer Regeltabelle definieren Regeln alle Felder, die aus jedem Datensatz in einer Datenquelle extrahiert werden können.

Die Regeln in der Tabelle werden der Reihe nach auf jeden Datensatz angewendet. Wenn alle Regeln in der Tabelle mit dem Datensatz übereinstimmen, werden keine anderen Regeltabellen für die Verarbeitung des Datensatzes benötigt und es wird der nächste Datensatz verarbeitet. Wenn eine Regel in der Tabelle nicht übereinstimmt, werden alle durch vorherige Regeln in der Tabelle extrahierten Feldwerte verworfen. Falls eine andere Regeltabelle vorhanden ist, werden die Regeln in der betreffenden Tabelle auf den Datensatz angewendet. Wenn keine Tabelle mit dem Datensatz übereinstimmt, wird die Regel für Abweichungen (Mismatch) angewendet.

Mismatch
Sie können auswählen, dass Datensätze, die mit keiner der Regeltabellen übereinstimmen, übersprungen werden sollen (**Skip**), oder Sie können den Wert aller Felder im Datensatz auf fehlend (**Missing**) setzen (null).

Export Rules
Sie können die zurzeit sichtbare Regeltabelle zwecks Wiederverwendung speichern. Die exportierte Tabelle wird auf dem Server gespeichert.

Import Rules
Sie können eine gespeicherte Regeltabelle in die zurzeit sichtbare Regeltabelle importieren. Dadurch werden alle von Ihnen für die betreffende Tabelle definierten Regeln überschrieben. Die beste Vorgehensweise besteht darin, eine neue Tabelle zu erstellen und dann eine Regeltabelle zu importieren.

Regeleditor:

Mit dem Regeleditor können Sie eine Extraktionsregel für ein einzelnes Feld erstellen.

Anonymous capture group
Eine Felderfassungsregel beginnt mit der Extraktion von Daten aus einem Datensatz normalerweise bei der Position, an der die vorherige Regel endete. Wenn zwischen zwei Feldern in einer semistrukturierten Datenquelle irrelevante Informationen vorhanden sind, kann es deshalb sinnvoll sein, eine anonyme Erfassungsgruppe zu definieren, die den Parser an der Stelle positioniert, an der das nächste Feld beginnt. Wenn Sie **Anonymous capture group** auswählen, werden die Steuerelemente für die Benennung und Kennzeichnung der Erfassungsgruppe inaktiviert, aber der verbleibende Teil des Dialogs funktioniert normal.

Field name
Geben Sie einen Namen für das Feld ein. Er wird zum Definieren der Datenquellenmetadaten verwendet. Feldnamen müssen innerhalb einer Regeltabelle eindeutig sein.

Rule name
Geben Sie optional eine Beschriftung für die Regel ein.

Description
Geben Sie optional eine längere Beschreibung für die Regel ein.

Defining a rule
Es gibt zwei Methoden zum Definieren von Regeln.

Use controls for extraction rules
Die Verwendung von Steuerelementen vereinfacht die Erstellung von Extraktionsregeln.

1. Geben Sie den Ausgangspunkt für das Extrahieren von Felddaten an. Mit der Option **Current position** wird an der Stelle begonnen, an der die vorherige Regel stoppte. Mit der Option **Skip until** wird am Anfang des Datensatzes begonnen und es werden alle Zeichen ignoriert, bis das im Textfeld angegebene Zeichen erreicht wird. Wählen Sie **Include** aus, wenn die Felddaten das Zeichen an der Anfangsposition einschließen sollen.
2. Wählen Sie eine Felderfassungsgruppe in der Dropdown-Liste **Capture** aus.
3. Wählen Sie optional den Punkt aus, an dem das Extrahieren von Felddaten gestoppt werden soll. Bei Auswahl von **Whitespace** wird gestoppt, wenn Leerraum (z. B. Leerzeichen oder Tabulatoren) gefunden wird, bei Auswahl von **At character(s)** wird bei der angegebenen Zeichenfolge gestoppt. Wählen Sie **Include** aus, wenn die Felddaten das Zeichen an der Endposition einschließen sollen.

Manually define regexp rules
Wählen Sie diese Option aus, wenn Sie die Syntax für reguläre Ausdrücke selber schreiben wollen. Geben Sie einen regulären Ausdruck in das Textfeld **Regexp** ein.

Add Field Capture Group
Ermöglicht Ihnen, den regulären Ausdruck zur späteren Verwendung zu speichern. Die gespeicherte Erfassungsgruppe wird in der Dropdown-Liste **Capture** angezeigt.

Der Regeleditor zeigt eine Vorschau der von dieser Regel aus dem ersten Datensatz extrahierten Daten an, nachdem alle vorherigen Regeln in der Regeltabelle angewendet wurden.

Einstellungen für Excel-Dateitypen:

Sie können die folgenden Einstellungen für Excel-Dateien angeben.

Worksheet selection
Wählt das Excel-Arbeitsblatt als zu verwendende Datenquelle aus. Geben Sie entweder einen numerischen Index (der Index des ersten Arbeitsblatts ist 0) oder den Namen des Arbeitsblatts an. Standardmäßig wird das erste Arbeitsblatt verwendet.

Data range selection for import.
Sie können den Datenimport mit der ersten nicht leeren Zeile oder mit einem expliziten Zellenbereich beginnen.

- **Range starts on first non-blank row**. Sucht die erste nicht leere Zelle und verwendet sie als linke obere Ecke des Datenbereichs.
- Geben Sie alternativ einen expliziten Zellenbereich nach Zeile und Spalte an. Wenn Sie beispielsweise den Excel-Bereich A1:D5 angeben wollen, können Sie A1 in das erste Feld und D5 in das zweite Feld eingeben (oder alternativ R1C1 und R5C4). Alle Zeilen im angegebenen Bereich werden zurückgegeben, einschließlich leerer Zeilen.

First row contains field names
Gibt an, ob die erste Zeile des ausgewählten Zellenbereichs die Feldnamen enthält. Der Standardwert ist **No**.

Stop reading after encountering blank rows
Gibt an, ob das Lesen von Datensätzen gestoppt wird, nachdem mehr als eine leere Zeile festgestellt wurde, oder ob das Lesen aller Daten bis an das Ende des Arbeitsblatts fortgesetzt wird, einschließlich leerer Zeilen. Der Standardwert ist **No**.

Registerkarte "Formats"

Auf der Registerkarte **Formats** können Sie Formatinformationen für die geparsten Felder definieren.

Feldkonvertierungseinstellungen

Trim white space
Entfernt Leerzeichen am Anfang und/oder Ende der Zeichenfolgefelder. Standardwert ist **None**. Die folgenden Werte werden unterstützt:

None Entfernt Leerzeichen nicht.

Left Entfernt Leerzeichen am Anfang der Zeichenfolge.

Right Entfernt Leerzeichen am Ende der Zeichenfolge.

Both Entfernt Leerzeichen am Anfang und Ende der Zeichenfolge.

Locale Definiert eine Ländereinstellung. Standardmäßig die Ländereinstellung des Servers. Die Ländereinstellungszeichenfolge sollte wie folgt angegeben werden: <Sprache>[_Land[_Variante]], wobei Folgendes gilt:

Sprache
Ein gültiger, aus zwei Buchstaben bestehender Code in Kleinbuchstaben gemäß ISO-639-Definition.

Land Ein gültiger, aus zwei Buchstaben bestehender Code in Großbuchstaben gemäß ISO-639-Definition.

Variante
Ein für den Anbieter oder Browser spezifischer Code.

Decimal separator
Legt das als Dezimalzeichen verwendete Zeichen fest. Standardmäßig wird die für die Ländereinstellung spezifische Einstellung verwendet.

Grouping symbols
Legt fest, ob das für die Ländereinstellung spezifische Zeichen, das als Tausendertrennzeichen verwendet wird, verwendet werden soll.

Default date format
Definiert ein Standarddatumsformat. Es werden alle durch die Unicode-LDML-Spezifikation (LDML - Locale Data Markup Language) definierten Formatmuster unterstützt.

Default time format
Definiert ein Standardzeitformat.

Default timestamp
Definiert ein Standardzeitmarkenformat.

Default time zone
Legt die Zeitzone fest. Standardmäßig wird UTC verwendet. Die Einstellung gilt für die Zeit- und Zeitmarkenfelder, für die nicht explizit eine Zeitzone angegeben ist.

Field Overrides

In diesem Abschnitt können Sie Formatierungsanweisungen für einzelne Felder zuweisen. Wählen Sie ein Feld aus dem Datenmodell aus oder geben Sie einen Feldnamen ein und klicken Sie auf **Add**, um das Feld der Liste mit den Feldern mit individuellen Anweisungen hinzuzufügen. Klicken Sie auf **Remove**, um es aus der Liste zu entfernen. Für ein in der Liste ausgewähltes Feld können Sie die folgenden Feldeigenschaften festlegen.

Storage
Legt den Speicherort des Felds fest.

Decimal separator
Legt für Realspeicherfelder das Zeichen fest, das als Dezimalzeichen verwendet wird. Standardmäßig wird die für die Ländereinstellung spezifische Einstellung verwendet.

Grouping symbols
Legt für Ganzzahl- oder Realspeicherfelder fest, ob das für die Ländereinstellung spezifische Tausendertrennzeichen verwendet werden soll.

Formats
Legt das Format für Datums-, Zeit- oder Zeitmarkenspeicherfelder fest. Wählen Sie ein Format in der Dropdown-Liste aus.

Registerkarte "Field Order"

Für Dateitypen mit Trennzeichen oder Excel-Dateitypen können Sie auf der Registerkarte **Field Order** die geparste Reihenfolge von Feldern für die Datei definieren. Dies ist wichtig, wenn mehrere Dateien in einer Datenquelle vorhanden sind, da die tatsächliche Reihenfolge der Felder über die Dateien hinweg abweichen kann, aber die geparste Reihenfolge der Felder identisch sein muss, um ein konsistentes Datenmodell zu erstellen.

Für unveränderliche und semistrukturierte Dateitypen wird die Reihenfolge auf der Registerkarte **Settings** definiert.

Wenn eine einzelne Datei in der Datenquelle vorhanden ist oder wenn alle Dateien dieselbe Feldreihenfolge aufweisen, können Sie die Standardeinstellung **Field order matches data model** verwenden. Wenn mehrere Dateien in der Datenquelle vorhanden sind und die Reihenfolge der Felder in den Dateien nicht identisch ist, müssen Sie eine bestimmte Feldreihenfolge (**Specific field order**) für das Parsen der Datei definieren.

1. Sie können der geordneten Liste ein Feld hinzuzufügen, indem Sie den Feldnamen eingeben oder aus der vom Datenmodell bereitgestellten Liste auswählen. Sie können alle Felder im Datenmodell gleichzeitig hinzufügen, indem Sie auf **Add all** klicken. Feldnamen werden der geordneten Liste nur ein einziges Mal hinzugefügt.

2. Verwenden Sie die Pfeiltasten, um die Felder wie gewünscht zu sortieren.

Bei Verwendung von **Specific field order** sind alle Felder, die nicht der Liste hinzugefügt werden, nicht Teil des Ergebnissets für diese Datei. Wenn im Datenmodell Felder vorhanden sind, die nicht in diesem Dialogfeld aufgelistet sind, werden die Werte im Ergebnisset als null angegeben.

Registerkarte "Folder"

Bei der Angabe von Parsereinstellungen für einen Ordner können Sie auf der Registerkarte **Folder** auswählen, welche Dateien im Ordner in die Datenquelle eingeschlossen werden sollen.

Match all files in the selected folder
: Die Datenquelle schließt alle in der höchsten Ebene des Ordners enthaltenen Dateien ein. Dateien in Unterordnern werden nicht eingeschlossen.

Match files using a regular expression
: Die Datenquelle schließt alle in der höchsten Ebene des Ordners enthaltenen Dateien ein, die mit dem angegebenen regulären Ausdruck übereinstimmen.

Match files using a Unix globbing expression (potentially recursive)
: Die Datenquelle schließt alle Dateien ein, die mit dem angegebenen UNIX-Globbing-Ausdruck übereinstimmen. Der Ausdruck kann Dateien enthalten, die sich in Unterordnern des ausgewählten Ordners befinden.

HCatalog-Feldzuordnungen

HCatalog Schema
: Zeigt die Struktur der angegebenen Tabelle an. HCatalog kann ein hochgradig strukturiertes Dataset unterstützen. Wenn eine Analytic Server-Datenquelle für solche Daten definiert werden soll, muss die Struktur in einfache Zeilen und Spalten abgeflacht werden. Wählen Sie ein Element im Schema aus und klicken Sie auf die Pfeilschaltfläche, um es für die Analyse einem Feld zuzuordnen.

  Nicht alle Baumknoten können zugeordnet werden. Beispielsweise wird ein Array oder eine Zuordnung komplexer Typen als "übergeordnetes Element" angesehen und kann nicht direkt zugeordnet werden; jedes einfache Element in einem HCatalog-Array oder in einer HCatalog-Zuordnung muss separat hinzugefügt werden. Diese Knoten sind durch eine Beschriftung im Baum gekennzeichnet, die auf "...:array:struct" oder auf "...:map:struct" endet.

  Beispiel:

  - Bei einem Array mit ganzen Zahlen können Sie ein Feld einem Wert innerhalb des Arrays zuweisen (bigintarray[45]), aber nicht das Array selbst (bigintarray)
  - Bei einer Zuordnung können Sie ein Feld einem Wert innerhalb der Zuordnung zuweisen (datamap["Schlüssel"]), aber nicht die Zuordnung selbst (datamap
  - Bei einem Array mit einem Array mit ganzen Zahlen können Sie ein Feld einem Wert zuweisen (bigintarrayarray[45][2]), aber nicht das Array selbst (bigintarrayarray[45]).

    Wenn Sie ein Feld einem Array- oder Zuordnungselement zuweisen, muss die Definition des Elements deshalb den Index oder Schlüssel einschließen: bigintarray[Index] oder bigintmap["Schlüssel"].

Feldzuordnungen

HCatalog Element
: Doppelklicken Sie auf eine Zelle, um sie zu bearbeiten. Sie müssen die Zelle bearbeiten, wenn das HCatalog-Element ein Array oder eine Zuordnung ist. Geben Sie mit einem Array die ganze Zahl an, die dem Mitglied des Arrays entspricht, das Sie einem Feld zuordnen wollen. Geben Sie mit einer Zuordnung eine in Anführungszeichen eingeschlossene Zeichenfolge an, die dem Schlüssel entspricht, den Sie einem Feld zuordnen wollen.

Mapping Field
Das Feld, wie es in der Analytic Server-Datenquelle angezeigt wird. Doppelklicken Sie auf eine Zelle, um sie zu bearbeiten. Doppelte Werte in der Spalte **Mapping Field** sind nicht zulässig und führen zu einem Fehler.

Storage
Der Speicherort des Felds. Der Speicherort wird von HCatalog abgeleitet und kann nicht bearbeitet werden.

Anmerkung: Wenn Sie auf Preview and Metadata klicken, um eine HCatalog-Datenquelle fertigzustellen, stehen keine Bearbeitungsoptionen zur Verfügung.

Raw Data
Zeigt die Datensätze an, wie sie in HCatalog gespeichert sind. So können Sie leichter festlegen, wie das HCatalog-Schema Feldern zugeordnet werden soll.

Anmerkung: Jede bei **HCatalog Selections** angegebene Filterung wird auf die Ansicht mit den Rohdaten angewendet.

Aktivieren von HCatalog-Datenquellen

Analytic Server bietet Unterstützung für HCatalog-Datenquellen. In diesem Abschnitt wird beschrieben, wie verschiedene zugrunde liegende NoSQL-Datenbanken aktiviert werden.

Apache Accumulo

Analytic Server bietet Unterstützung für HCatalog-Datenquellen, denen Inhalt in Apache Accumulo zugrunde liegt.

Der verteilte Schlüssel/Wert-Speicher von Apache Accumulo ist ein auf dem BigTable-Design von Google basierendes System zum Speichern und Abrufen von Daten und baut auf Apache Hadoop, Zookeeper und Thrift auf. Apache Accumulo enthält eine Reihe neuartiger Verbesserungen für das BigTable-Design in Form von zellbasierter Zugriffssteuerung und eines serverseitigen Programmierungsmechanismus, mit dem Schlüssel/Wert-Paare an unterschiedlichen Punkten im Datenmanagementprozess modifiziert werden können.

Verwenden Sie die folgende Syntax, um eine externe Apache Accumulo-Tabelle in Hive zu erstellen:

```
set accumulo.instance.id=<Instanzname>;
set accumulo.user.name=<Benutzername>;
set accumulo.user.pass=<Benutzerkennwort>;
set accumulo.zookeepers=<Zookeeper-Host-Port>;

CREATE EXTERNAL TABLE <Hive-Tabellenname>(<Tabellenspaltenspezifikationen>)
STORED BY 'com.ibm.spss.hcatalog.AccumuloStorageHandler'
WITH SERDEPROPERTIES (
'accumulo.columns.mapping' = '<Familien-_und_Qualifikationszuordnungen>',
'accumulo.table.name' = '<Accumulo-Tabellenname>')
TBLPROPERTIES (
    "accumulo.instance.id"="<Instanzname>",
    "accumulo.zookeepers"="<Zookeeper-Host-Port>"
);
```

Beispiel:

```
set accumulo.instance.id=<ID>;
set accumulo.user.name=admin;
set accumulo.user.pass=test;
set accumulo.zookeepers=<Host>:<Port>;

CREATE EXTERNAL TABLE acc_drug1n(rowid STRING,age STRING,sex STRING,bp STRING,
    cholesterol STRING,na STRING,k STRING,drug STRING)
STORED BY 'com.ibm.spss.hcatalog.AccumuloStorageHandler'
WITH SERDEPROPERTIES (
```

```
'accumulo.columns.mapping' = 'rowID,drug|age,drug|sex,drug|bp,drug|cholesterol,
    drug|na,drug|k,drug|drug',
'accumulo.table.name' = 'drug1n')
TBLPROPERTIES (
    "accumulo.instance.id"="<ID>",
    "accumulo.zookeepers"="<Host>:<Port>"
);
```

Anmerkung: Der Accumulo-Benutzername und das zugehörige Kennwort für die angegebene Accumulo-Tabelle müssen mit dem Benutzernamen und dem Kennwort des authentifizierten Analytic Server-Benutzers übereinstimmen.

Apache Cassandra

Analytic Server bietet Unterstützung für HCatalog-Datenquellen, denen Inhalt in Apache Cassandra zugrunde liegt.

Cassandra stellt einen strukturierten Schlüssel/Wert-Speicher bereit. Schlüssel werden mehreren Werten zugeordnet, die als Spaltenfamilien gruppiert werden. Die Spaltenfamilien werden beim Erstellen einer Datenbank festgelegt, einer Familie können aber jederzeit Spalten hinzugefügt werden. Darüber hinaus werden Spalten nur angegebenen Schlüsseln hinzugefügt, sodass unterschiedliche Schlüssel in jeder beliebigen Familie eine unterschiedliche Anzahl Spalten aufweisen können. Die Werte aus einer Spaltenfamilie für jeden Schlüssel werden zusammen gespeichert.

Cassandra-Tabellen können auf zwei Arten definiert werden: mit der traditionellen Cassandra-Befehlszeilenschnittstelle (cassandra-cli) und mit der neuen CQL-Shell (csqlsh).

Verwenden Sie die folgende Syntax, um eine externe Apache Cassandra-Tabelle in Hive zu erstellen, wenn die Tabelle mit der traditionellen Befehlszeilenschnittstelle erstellt wurde.

```
CREATE EXTERNAL TABLE <Hive-Tabellenname>  (<Spaltenspezifikationen>)
STORED BY 'com.ibm.spss.hcatalog.CassandraStorageHandler'
WITH SERDEPROPERTIES("cassandra.cf.name" = "<Cassandra-Spaltenfamilie>",
"cassandra.host"="<Cassandra-Host>","cassandra.port" = "<Cassandra-Port>")
TBLPROPERTIES ("cassandra.ks.name" = "<Cassandra-Schlüsselbereich>");
```

Beispiel: Für die folgende CLI-Tabellendefinition

```
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'];
```

sieht die Hive-Tabellen-DDL wie folgt aus:

```
CREATE EXTERNAL TABLE cassandra_users  (key string, first string, last string, age string)
STORED BY 'com.ibm.spss.hcatalog.CassandraStorageHandler'
WITH SERDEPROPERTIES("cassandra.cf.name" = "users",
"cassandra.host"="<Cassandra-Host>","cassandra.port" = "9160")
TBLPROPERTIES ("cassandra.ks.name" = "test");
```

Verwenden Sie die folgende Syntax, um eine externe Apache Cassandra-Tabelle in Hive zu erstellen, wenn die Tabelle mit CQL erstellt wurde.

```
CREATE EXTERNAL TABLE <Hive-Tabellenname>  (<Spaltenspezifikationen>)
STORED BY 'com.ibm.spss.hcatalog.CassandraCqlStorageHandler'
WITH SERDEPROPERTIES("cassandra.cf.name" = "<Cassandra-Spaltenfamilie>",
"cassandra.host"="<Cassandra-Host>","cassandra.port" = "<Cassandra-Port>")
TBLPROPERTIES ("cassandra.ks.name" = "<Cassandra-Schlüsselbereich>");
```

Beispiel: Für die folgende CQL3-Tabellendefinition

```
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);
```

sieht die Hive-Tabellen-DDL wie folgt aus:

```
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 'com.ibm.spss.hcatalog.CassandraCqlStorageHandler'
WITH SERDEPROPERTIES("cassandra.cf.name" = "bankloan_10","cassandra.host"="<cassandra_host>",
   "cassandra.port" = "9160")
TBLPROPERTIES ("cassandra.ks.name" = "test");
```

Apache HBase

Analytic Server bietet Unterstützung für HCatalog-Datenquellen, denen Inhalt in Apache HBase zugrunde liegt.

Apache HBase ist ein verteilter, versionsgesteuerter, spaltenorientierter Open-Source-Speicher, für den Hadoop und HDFS die Basis bilden.

Verwenden Sie die folgende Syntax, um eine externe HBase-Tabelle in Hive zu erstellen:

```
CREATE EXTERNAL TABLE <Tabellenname>(<Tabellenspaltenspezifikation>)
STORED BY 'com.ibm.spss.hcatalog.HBaseStorageHandler'
WITH SERDEPROPERTIES ("hbase.columns.mapping" = "<Spaltenzuordnungsspezifikation>")
TBLPROPERTIES("hbase.table.name" = "<HBase-Tabellenname>")
```

Beispiel:

```
CREATE EXTERNAL TABLE hbase_drug1n(rowid STRING,age STRING,sex STRING,bp STRING,
    cholesterol STRING,na STRING,k STRING,drug STRING)
STORED BY 'com.ibm.spss.hcatalog.HBaseStorageHandler'
WITH SERDEPROPERTIES ("hbase.columns.mapping" = ":key,drug:age,drug:sex,drug:bp,
    drug:cholesterol,drug:na,drug:k,drug:drug")
TBLPROPERTIES("hbase.table.name" = "drug1n");
```

Anmerkung: Informationen zum Erstellen einer HBase-Tabelle finden Sie im Apache HBase Reference Guide (http://hbase.apache.org/book.html).

Anmerkung: Es hat sich bewährt, dem Datenbanknamen ein Präfix voranzustellen, um den Datenbanktyp anzugeben. Vergeben Sie beispielsweise den Namen 'HB_drug1n' für eine HBase-Datenbank oder 'ACC_drug1n' für eine Accumulo-Datenbank. Dies erleichtert beim Arbeiten mit der Analytic Server-Konsole die Auswahl der HCatalog-Datei.

MongoDB

Analytic Server bietet Unterstützung für HCatalog-Datenquellen, denen Inhalt in MongoDB zugrunde liegt.

MongoDB ist eine Open-Source-Dokumentdatenbank, bei der es sich um eine der führenden in C++ geschriebenen NoSQL-Datenbanken handelt. Die Datenbank speichert im JSON-Stil erstellte Dokumente mit dynamischen Schemas.

Verwenden Sie die folgende Syntax, um eine externe MongoDB-Tabelle in Hive zu erstellen:

```
create external table <Hive-Tabellenname>(<Spaltenspezifikationen>)
stored by "com.ibm.spss.hcatalog.MongoDBStorageHandler"
with serdeproperties ( "mongo.column.mapping" = "<MongoDB-Hive-Zuordnung>" )
tblproperties ( "mongo.uri" = "'mongodb://<Host>:<Port>/<Datenbank>.<Sammlung>" );
```

Beispiel:

```
create external table mongo_bankloan(age bigint,ed bigint,employ bigint, address bigint,income bigint,
    debtinc double, creddebt double,othdebt double,default bigint)
STORED BY 'com.ibm.spss.hcatalog.MongoDBStorageHandler'
with serdeproperties ( 'mongo.column.mapping' = '{"age":"age","ed":"ed","employ":"employ","address":"address",
    "income":"income","debtinc":"debtinc","creddebt":"creddebt","othdebt":"othdebt","default":"default"}' )
tblproperties ('mongo.uri'='mongodb://9.48.11.162:27017/test.bankloan');
```

Oracle NoSQL

Analytic Server bietet Unterstützung für HCatalog-Datenquellen, denen Inhalt in Oracle NoSQL zugrunde liegt.

Die Oracle NoSQL-Datenbank ist eine verteilte Schlüssel/Wert-Datenbank. Daten werden als Schlüssel/Wert-Paare gespeichert, die basierend auf dem Hashwert des Primärschlüssels in bestimmte Speicherknoten geschrieben werden. Speicherknoten werden repliziert, um Hochverfügbarkeit sicherzustellen. Kundenanwendungen werden mit der Java-/C-API geschrieben, um Daten zu lesen und zu schreiben.

SERDEPROPERTIES- und TABLEPROPERTIES-Parameter

Der Oracle NoSQL-Speicherhandler unterstützt die folgenden Parameter.

SERDEPROPERTIES-Parameter

kv.major.keys.mapping
Durch Kommas getrennte Liste der Hauptschlüssel. Erforderlich

kv.minor.keys.mapping
Durch Kommas getrennte Liste der Nebenschlüssel. Optional

kv.parent.key
Gibt den übergeordneten Schlüssel an, dessen untergeordnete Schlüssel/Wert-Paare von der Abfrage zurückgegeben werden sollen. Der Hauptschlüsselpfad muss ein Teilpfad sein und der Nebenschlüsselpfad muss leer sein. Optional.

kv.avro.json.key
Der Name des Nebenschlüssels, der verwendet wird, um den mit dem Avro-Schema definierten Wert aufzunehmen. Wenn der Nebenschlüssel nicht definiert ist, was in der Regel der Fall ist, wird der Standardwert "value" verwendet. Wenn der Parameter nicht definiert ist, wird der Wert als JSON-Zeichenfolge zurückgegeben. Optional.

kv.avro.json.keys.mapping.column
Definiert den Namen der Hive-Spalte für die Schlüssel/Wert-Paare mit den Haupt- und Nebenschlüsseln. Die Hive-Spalte sollte den Typ map<Zeichenfolge,Zeichenfolge> haben. Optional.

TABLEPROPERTIES-Parameter

kv.host.port
Die IP-Adresse und Portnummer der Oracle NoSQL-Datenbank. Erforderlich

kv.name
Der Name des Schlüssel/Wert-Speichers von Oracle NoSQL. Erforderlich.

Beispiel: Einfaches Avro-Schema

Das Datenlayout wird mithilfe des Apache Avro-Serialisierungsframeworks modelliert. Bei diesem Ansatz erstellen Sie ein Avro-Schema, z. B.

```
{"type":"record",
 "name":"DrugSchema",
 "namespace":"avro",
 "fields":[
        {"name":"id","type":"string","default":""},
        {"name":"age","type":"string","default":""},
        {"name":"sex","type":"string","default":""},
        {"name":"bp","type":"string","default":""},
        {"name":"drug","type":"string","default":""}
]}
```

Dieses Schema sollte für die Oracle NoSQL-Datenbank registriert werden und die ausgefüllten Daten sollten wie nachfolgend gezeigt eine Referenz auf das Schema enthalten.

```
put -key /drugstore_avro/1 -value
   "{\"id\":\"1\",\"age\":\"23\",\"sex\":\"F\",\"bp\":\"HIGH\",\"drug\":\"drugY\"}"
   -json avro.DrugSchema
put -key /drugstore_avro/2 -value
   "{\"id\":\"2\",\"age\":\"47\",\"sex\":\"M\",\"bp\":\"LOW\",\"drug\":\"drugC\"}"
   -json avro.DrugSchema
put -key /drugstore_avro/3 -value
   "{\"id\":\"3\",\"age\":\"47\",\"sex\":\"M\",\"bp\":\"LOW\",\"drug\":\"drugC\"}"
   -json avro.DrugSchema
put -key /drugstore_avro/4 -value
```

```
   "{\"id\":\"4\",\"age\":\"28\",\"sex\":\"F\",\"bp\":\"NORMAL\",\"drug\":\"drugX\"}"
   -json avro.DrugSchema
put -key /drugstore_avro/5 -value
   "{\"id\":\"5\",\"age\":\"61\",\"sex\":\"F\",\"bp\":\"LOW\",\"drug\":\"drugY\"}"
   -json avro.DrugSchema
```

Erstellen Sie eine externe Tabelle und geben Sie die zusätzliche Eigenschaft **kv.avro.json.key** im Abschnitt SERDEPROPERTIES an, um die Daten in Hive bereitzustellen. Der Wert der Eigenschaft muss der Name des Nebenschlüssels sein oder der vordefinierte Name **value**, falls der Nebenschlüssel nicht definiert ist.

```
CREATE EXTERNAL TABLE oracle_json(id string, age string, sex string, bp string, drug string)
  STORED BY 'com.ibm.spss.hcatalog.OracleKVStorageHandler'
  WITH SERDEPROPERTIES ("kv.major.keys.mapping" = "drugstore_avro,keyid",
        "kv.parent.key"="/drugstore_avro","kv.avro.json.key" = "value")
  TBLPROPERTIES ("kv.host.port" = "<hostname>:5000", "kv.name" = "kvstore");
```

Die Ausführung von 'select * from oracle_json' gibt die folgenden Ergebnisse zurück.

```
select * from oracle_json;

1 23 F HIGH drugY
5 61 F LOW drugY
3 47 M LOW drugC
2 47 M LOW drugC
4 28 F NORMAL drugX
```

Die Tabelle oracle_json kann in der Analytic Server-Konsole verwendet werden, um eine Oracle NoSQL-Datenquelle zu erstellen.

Beispiel: Komplexe Schlüssel

Betrachten Sie jetzt das folgende Avro-Schema.

```
{"type":"record",
  "name":"DrugSchema",
  "namespace":"avro",
  "fields":[
      {"name": "age", "type": "string", "default": ""}, // age
      {"name": "bp", "type":"string","default":""},     // blood pressure
      {"name": "drug", "type": "int", "default": ""},    // drug administered
]}
```

Gehen Sie außerdem davon aus, dass der Schlüssel wie folgt modelliert ist:

```
/u/<sex (M/F)>/<patient ID>
```

und füllen Sie den Datenspeicher mit den folgenden Befehlen:

```
put -key /u/F/1 -value
  "{\"age\":\"23\",\"bp\":\"HIGH\",\"drug\":\"drugY\"}" -json avro.DrugSchema
put -key /u/M/2 -value
  "{\"age\":\"47\",\"bp\":\"LOW\",\"drug\":\"drugC\"}" -json avro.DrugSchema
put -key /u/M/3 -value
  "{\"age\":\"47\",\"bp\":\"LOW\",\"drug\":\"drugC\"}" -json avro.DrugSchema
put -key /u/F/4 -value
  "{\"age\":\"28\",\"bp\":\"NORMAL\",\"drug\":\"drugX\"}" -json avro.DrugSchema
put -key /u/F/5 -value
  "{\"age\":\"61\",\"bp\":\"LOW\",\"drug\":\"drugY\"}" -json avro.DrugSchema
```

Damit die Informationen zum Geschlecht und zur Benutzer-ID aus den Hauptschlüsseln beibehalten werden, sollte die Tabelle mit einem zusätzlichen SERDEPROPERTIES-Parameter **kv.avro.json.keys.mapping.column** erstellt werden. Der Wert des Parameters sollte der Name der Hive-Spalte des Typs map<Zeichenfolge,Zeichenfolge> sein. Bei den Schlüsseln in der Zuordnung handelt es

sich um die Namen der Datensatzschlüssel, die in den Eigenschaften **kv.*.keys.mapping** angegeben sind, und bei den Werten handelt es sich um die tatsächlichen Schlüsselwerte. Die DDL zur Tabellenerstellung wird nachfolgend gezeigt:

```
CREATE EXTERNAL TABLE oracle_user(keys map<string,string>, age string, bp string, drug string)
  STORED BY 'com.ibm.spss.hcatalog.OracleKVStorageHandler'
  WITH SERDEPROPERTIES ("kv.major.keys.mapping" = "DrugSchema,sex,patientid",
      "kv.parent.key" = "/u",
      "kv.avro.json.key" = "value",
      "kv.avro.json.keys.mapping.column" = "keys")
  TBLPROPERTIES ("kv.host.port" = "<hostname>:5000", "kv.name" = "kvstore");
```

Die Ausführung von 'select * from oracle_user' gibt die folgenden Ergebnisse zurück:

```
select * from
        oracle_user;
      {"user":"u","gender":"m"," userid":"125"} joe smith 77 13
      {"user":"u","gender":"m"," userid":"129"} jeff smith 67 27
      {"user":"u","gender":"m"," userid":"127"} jim smith 78 11
      {"user":"u","gender":"f"," userid":"131"} jen schmitt 70 20
      {"user":"u","gender":"m"," userid":"130"} jed schmidt 60 31
      {"user":"u","gender":"f"," userid":"128"} jan smythe 79 10
      {"user":"u","gender":"f"," userid":"126"} jess smith 76 12
```

Die Tabelle oracle_user kann in der Analytic Server-Konsole verwendet werden, um eine Oracle NoSQL-Datenquelle zu erstellen. Die Schlüssel für das Geschlecht und die Patienten-ID sowie die Spaltennamen aus dem Avro-Schema können verwendet werden, um entsprechende Felder für die Datenquelle zu definieren.

Bereichsscans

Analytic Server unterstützt Bereichsscans auf der Basis des übergeordneten Präfix für die Hauptschlüssel sowie von Unterbereichen, um den Bereich unterhalb des übergeordneten Schlüssels weiter zu beschränken.

Der übergeordnete Schlüssel gibt das Präfix für die untergeordneten Schlüssel/Wert-Paare an, die zurückgegeben werden sollen. Ein leeres Präfix führt dazu, dass alle Schlüssel im Speicher abgerufen werden. Wenn das Präfix nicht leer ist, muss der Hauptschlüsselpfad ein Teilpfad sein und der Nebenschlüsselpfad muss leer sein. Der übergeordnete Schlüssel wird als Datenquellenattribut **com.ibm.spss.ae.hcatalog.range.parent** gespeichert.

Der Unterbereich schränkt den Bereich unterhalb des übergeordneten Schlüssels noch weiter ein, und zwar auf die Hauptpfadkomponenten im Unterbereich. Der Unterbereichsstartschlüssel wird als **com.ibm.spss.ae.hcatalog.range.start** gespeichert und der Unterbereichsendschlüssel als **com.ibm.spss.ae.hcatalog.range.end**. Der Startschlüssel sollte lexikografisch kleiner-gleich dem Endschlüssel sein. Die Unterbereichsparameter sind optional.

XML-Datenquellen

Analytic Server bietet Unterstützung für XML-Daten über HCatalog.

Beispiel

1. XML-Schema den Hive-Datentypen über die Hive Data Definition Language (DDL) gemäß den folgenden Regeln zuordnen:

```
CREATE [EXTERNAL] TABLE <Tabellenname> (<Spaltenspezifikationen>)
ROW FORMAT SERDE "com.ibm.spss.hive.serde2.xml.XmlSerDe"
WITH SERDEPROPERTIES (
    ["xml.processor.class"="<XML-Prozessorklassenname>",]
    "column.xpath.<Spaltenname>"="<XPath-Abfrage>",
     ...
    ["xml.map.specification.<Elementname>"="<Zuordnungsspezifikation>"
     ...
```

```
    ]
)
STORED AS
    INPUTFORMAT "com.ibm.spss.hive.serde2.xml.XmlInputFormat"
    OUTPUTFORMAT "org.apache.hadoop.hive.ql.io.IgnoreKeyTextOutputFormat"
[LOCATION "<Datenposition>"]
TBLPROPERTIES (
    "xmlinput.start"="<Starttag ",
    "xmlinput.end"="<Endtag>"
);
```

Anmerkung: Wenn Ihre XML-Dateien mit Bz2-Komprimierung komprimiert werden, sollte `com.ibm.spss.hive.serde2.xml.SplittableXmlInputFormat` für `INPUTFORMAT` festgelegt werden. Wenn sie mit CMX-Komprimierung komprimiert werden, sollte `com.ibm.spss.hive.serde2.xml.CmxXmlInputFormat` festgelegt werden.

Beispiel: Die folgende XML

```
<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>
```

würde durch die folgende Hive-DLL dargestellt werden:

```
CREATE TABLE xml_bank(customer_id STRING, demographics map<Zeichenfolge,Zeichenfolge>,
financial map<Zeichenfolge,Zeichenfolge>)
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>"
);
```

Weitere Informationen finden Sie in „Zuordnung von XML zu Hive-Datentypen" auf Seite 23.

2. Analytic Server-Datenquelle mit HCatalog-Inhaltstyp in der Analytic Server-Konsole erstellen.

Einschränkungen

- Zurzeit wird nur die XPath 1.0-Spezifikation unterstützt.

- Bei der Handhabung von Hive-Feldnamen wird nur der lokale Teil der qualifizierten Namen für die Elemente und Attribute verwendet. Die Namensbereichspräfixe werden ignoriert.

Zuordnung von XML zu Hive-Datentypen: Die in XML modellierten Daten können anhand der nachfolgend aufgeführten Konventionen in Hive-Datentypen transformiert werden.

Strukturen

Das XML-Element kann dem Hive-Strukturtyp direkt zugeordnet werden, sodass alle Attribute zu Dateneinträgen werden. Der Inhalt des Elements wird zu einem zusätzlichen Eintrag mit primitivem oder komplexem Typ.

XML-Daten

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

Hive-DDL und Rohdaten

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

Arrays

Die XML-Sequenzen von Elementen können als Hive-Arrays mit primitivem oder komplexem Typ dargestellt werden. Das folgende Beispiel zeigt, wie der Benutzer einen Array mit Zeichenfolgen unter Verwendung des Inhalts des XML-Elements <result> definieren kann.

XML-Daten

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

Hive-DDL und Rohdaten

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

Zuordnungen

Das XML-Schema stellt keine native Unterstützung für Zuordnungen bereit. Es gibt drei allgemeine Ansätze für die Modellierung von Zuordnungen in XML. Um den drei unterschiedlichen Ansätzen Rechnung zu tragen, wird die folgende Syntax verwendet:

```
"xml.map.specification.<Elementname>"="<Schlüssel>-><Wert>"
```

Dabei gilt Folgendes:

Elementname
: Name des XML-Elements, das als Zuordnungseintrag berücksichtigt werden soll

Schlüssel
: XML-Knoten für den Zuordnungseintragsschlüssel

Wert XML-Knoten für den Zuordnungseintragswert

Die Zuordnungsspezifikation für das angegebene XML-Element sollte in der Hive-Tabellenerstellungs-DLL unter dem Abschnitt SERDEPROPERTIES definiert werden. Die Schlüssel und Werte können mithilfe der folgenden Syntax definiert werden:

@attribute
: Mit der Spezifikation @attribute kann der Benutzer den Wert des Attributs als Schlüssel oder Wert für die Zuordnung verwenden.

element
: Der Elementname kann als Schlüssel oder Wert verwendet werden.

#content

Der Inhalt des Elements kann als Schlüssel oder Wert verwendet werden. Da die Zuordnungsschlüssel nur den primitiven Typ haben können, wird der komplexe Inhalt in eine Zeichenfolge konvertiert.

Die Ansätze zur Darstellung von Zuordnungen in XML und die entsprechende Hive-DLL sowie die entsprechenden Rohdaten werden nachfolgend beschrieben.

Elementname zu Inhalt

Der Name des Elements wird als Schlüssel und der Inhalt als Wert verwendet. Dies ist eines der gängigen Verfahren und wird standardmäßig beim Zuordnen von XML zu Hive-Zuordnungstypen verwendet. Die offensichtliche Einschränkung bei diesem Ansatz besteht darin, dass der Zuordnungsschlüssel nur den Zeichenfolgetyp haben kann.

XML-Daten

```
<Eintrag1>Wert1</Eintrag1>
<Eintrag2>Wert2</Eintrag2>
<Eintrag3>Wert3</Eintrag3>
```

Zuordnung, Hive-DDL und Rohdaten

In diesem Fall müssen Sie keine Zuordnung angeben, da standardmäßig der Name des Elements als Schlüssel und der Inhalt als Wert verwendet wird.

```
result map<Zeichenfolge,Zeichenfolge>
```

```
{"result":{"Eintrag1": "Wert1", "Eintrag2": "Wert2", "Eintrag3": "Wert3"}}
```

Attribut zu Elementinhalt

Attributwert als Schlüssel und Elementinhalt als Wert verwenden.

XML-Daten

```
<entry name="Schlüssel1">Wert1</entry>
<entry name="Schlüssel2">Wert2</entry>
<entry name="Schlüssel3">Wert3</entry>
```

Zuordnung, Hive-DDL und Rohdaten

```
"xml.map.specification.entry"="@name->#content"
```

```
result map<Zeichenfolge,Zeichenfolge>
```

```
{"result":{"Schlüssel1": "Wert1", "Schlüssel2": "Wert2", "Schlüssel3": "Wert3"}}
```

Attribut zu Attribut

XML-Daten

```
<entry name="Schlüssel1" value="Wert1"/>
<entry name="Schlüssel2" value="Wert2"/>
<entry name="Schlüssel3" value="Wert3"/>
```

Zuordnung, Hive-DDL und Rohdaten

```
"xml.map.specification.entry"="@name->@value"
```

```
result map<Zeichenfolge,Zeichenfolge>
```

```
{"result":{"Schlüssel1": "Wert1", "Schlüssel2": "Wert2", "Schlüssel3": "Wert3"}}
```

Komplexer Inhalt

Durch Hinzufügen eines Stammelements mit Namen <string> wird komplexer Inhalt, der als primitiver Typ verwendet wird, in eine gültige XML-Zeichenfolge konvertiert. Betrachten Sie die folgende XML:

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

Der XPath-Ausdruck /dataset/* hat zur Folge, dass eine Reihe von XML-Knoten des Typs <value> zurückgegeben werden. Wenn das Zielfeld ein primitiver Typ ist, transformiert die Implementierung das Ergebnis der Abfrage in gültige XML, indem der Stammknoten <string> hinzugefügt wird.

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

Anmerkung: Die Implementierung fügt kein Stammelement <string> hinzu, wenn das Ergebnis der Abfrage ein einzelnes XML-Element ist.

Textinhalt

Wenn ein XML-Element nur Leerzeichen als Text enthält, wird der Text ignoriert.

Vorschau und Metadaten (Datenquellen)

Durch Klicken auf **Preview and Metadata** wird eine Stichprobe von Datensätzen sowie das Datenmodell für die Datenquelle angezeigt. Hier können Sie die grundlegenden Metadateninformationen überprüfen.

Preview

Auf der Registerkarte **Preview** werden eine kleine Stichprobe der Datensätze und ihre Feldwerte angezeigt.

Edit

Auf der Registerkarte **Edit** werden die grundlegenden Feldmetadaten angezeigt. Für Datenquellen mit dem Inhaltstyp "Dateien" wird das Datenmodell anhand einer kleinen Stichprobe mit Datensätzen generiert. Sie können die Feldmetadaten auf dieser Registerkarte manuell bearbeiten. Für Datenquellen mit dem Inhaltstyp "HCatalog" wird das Datenmodell anhand der HCatalog-Feldzuordnungen generiert. Sie können den Feldspeicher auf dieser Registerkarte nicht bearbeiten.

Field Doppelklicken Sie auf den Feldnamen, um ihn zu bearbeiten.

Measurement

Dies ist das Messniveau, mit dem Merkmale der Daten in einem bestimmten Feld beschrieben werden.

Role Wird verwendet, um Modellierungsknoten mitzuteilen, ob Felder für einen Computerlernprozess Eingabefelder (Input, Vorhersagefelder) oder Zielfelder (Target, vorherzusagende Felder) sind. **Both** und **None** sind ebenfalls verfügbare Rollen, ebenso wie die Rolle **Partition**, die ein Feld angibt, mit dem Datensätze zu Schulungs-, Test- und Prüfzwecken in separate Stichproben aufgeteilt werden. Der Wert **Split** gibt an, dass für jeden möglichen Wert des Feldes separate Modelle erstellt werden. **Frequency** gibt an, dass ein Feldwert als Häufigkeitsgewichtung für jeden Datensatz verwendet werden sollte. Mit **Record ID** wird ein Datensatz in der Ausgabe angegeben.

Storage

Mit der Option **Storage** wird beschrieben, wie Daten in einem Feld gespeichert werden. In einem Feld mit den Werten 1 und 0 werden z. B. ganzzahlige Daten gespeichert. Dies darf nicht mit dem Messniveau verwechselt werden, das die Verwendung der Daten beschreibt und sich nicht auf die Speicherung auswirkt. Sie können z. B. das Messniveau für ein Feld für ganze Zahlen mit den Werten 1 und 0 auf **Flag** setzen. Dies gibt normalerweise Folgendes an: 1 = True und 0 = False.

Values

Zeigt die einzelnen Werte für Felder mit kategorialer Messung oder den Wertebereich für Felder mit fortlaufender Messung an.

Structure
Gibt an, ob Datensätze in dem Feld einen einzelnen Wert (primitives Element) oder eine Liste mit Werten enthalten.

Depth Gibt die Tiefe einer Liste an; 0 ist eine Liste mit primitiven Elementen, 1 ist eine Liste mit Listen usw.

Scan all Data Values
Ermöglicht Ihnen, einen Scan für die Datenwerte der Datenquelle einzuleiten und abzubrechen, um die Kategorienwerte und Bereichsgrenzwerte zu ermitteln. Wenn ein Scan in Bearbeitung ist, klicken Sie zum Abbrechen des Scans auf die Schaltfläche **Cancel Data Scan**. Das Scannen aller Datenwerte stellt sicher, dass die Metadaten korrekt sind, aber es kann einige Zeit in Anspruch nehmen, wenn die Datenquelle viele Felder und Datensätze enthält.

Projekte

Projekte sind Arbeitsbereiche, in denen Eingaben gespeichert werden und auf Ausgaben von Jobs zugegriffen wird. Sie stellen die Organisationsstruktur der höchsten Ebene bereit, die Dateien und Ordner enthält. Projekte können mit einzelnen Benutzern und Gruppen gemeinsam genutzt werden.

Projektliste

Die Hauptseite mit den Projekten enthält eine Liste mit Projekten, deren Mitglied der aktuelle Benutzer ist.

- Klicken Sie auf den Namen eines Projekts, um die zugehörigen Details anzuzeigen und die Eigenschaften zu bearbeiten.
- Geben Sie einen Suchbegriff in den Suchbereich ein, um die Liste zu filtern, damit nur Projekte angezeigt werden, deren Name den Suchbegriff enthält.
- Klicken Sie auf **New**, um ein neues Projekt mit dem Namen zu erstellen, den Sie im Dialogfeld **Add new project** angeben. Informationen zu den Einschränkungen bei Namen, die Sie für Projekte vergeben können, finden Sie in „Benennungsregeln" auf Seite 29.
- Klicken Sie auf **Delete**, um das (die) ausgewählte(n) Projekt(e) zu entfernen. Diese Aktion entfernt das Projekt und löscht alle zum Projekt gehörigen Daten aus HDFS.
- Klicken Sie auf **Refresh**, um die Liste zu aktualisieren.

Einzelne Projektdetails

Der Inhaltsbereich ist in die ausblendbaren Abschnitte **Details**, **Sharing**, **Files** und **Versions** unterteilt.

Details

Name Ein bearbeitbares Textfeld, in dem der Name des Projekts angezeigt wird.

Display name
Ein bearbeitbares Textfeld, in dem der Name des Projekts so wie in anderen Anwendungen angezeigt wird. Wenn dieses Feld leer ist, wird der in **Name** angegebene Name als Anzeigename verwendet.

Description
Ein bearbeitbares Textfeld, in dem Sie einen erläuternden Text zum Projekt angeben können.

Versions to keep
Löscht automatisch die älteste festgeschriebene Projektversion, wenn die Anzahl der Versionen die angegebene Anzahl überschreitet. Der Standardwert ist 25.

Anmerkung: Der Bereinigungsprozess wird nicht sofort, sondern alle 20 Minuten im Hintergrund ausgeführt.

Is public

Ein Kontrollkästchen, das angibt, ob jeder das Projekt sehen kann (Kästchen ist ausgewählt) oder ob Benutzer und Gruppen explizit als Mitglieder hinzugefügt werden müssen (Kästchen ist abgewählt).

Klicken Sie auf **Save**, um den aktuellen Status der Einstellungen zu speichern.

Sharing

Sie können ein Projekt gemeinsam nutzen, indem Sie Benutzer und Gruppen als Autoren oder Anzeigeberechtigte hinzufügen.

- Durch Eingeben eines Suchbegriffs in das Textfeld wird nach Benutzern und Gruppen gefiltert, deren Name den Suchbegriff enthält. Wählen Sie die Ebene der gemeinsamen Nutzung aus und klicken Sie auf **Add member**, um die Liste mit den Mitgliedern zu ergänzen.
  - Autoren sind Vollmitglieder eines Projekts und können das Projekt sowie die darin enthaltenen Ordner und Dateien ändern. Die Benutzer und Mitglieder dieser Gruppen haben Schreibzugriff (Analytic Server-Exportknoten) auf dieses Projekt, wenn sie zu Analytic Server eine Verbindung über IBM® SPSS Modeler herstellen.
  - Anzeigeberechtigte können die Ordner und Dateien in einem Projekt anzeigen und Datenquellen über die Objekte in einem Projekt definieren, aber sie können das Projekt nicht ändern.
- Zum Entfernen eines Autors wählen Sie einen Benutzer oder eine Gruppe in der Autorenliste aus und klicken Sie auf **Remove member**.

Anmerkung: Administratoren verfügen über Lese- und Schreibzugriff auf jedes Projekt, unabhängig davon, ob sie namentlich als Mitglied aufgelistet sind.

Anmerkung: Auf der Registerkarte **Sharing** vorgenommene Änderungen werden sofort und automatisch angewendet.

Files

Projektstrukturbereich

Im rechten Bereich wird die Projekt-/Ordnerstruktur für das zurzeit ausgewählte Projekt angezeigt. Sie können die Ordnerstruktur durchsuchen, sie ist jedoch nur über die Schaltflächen bearbeitbar.

- Klicken Sie auf **Download file to the local filesystem**, um eine ausgewählte Datei in das lokale Dateisystem herunterzuladen.
- Klicken Sie auf **Delete the selected file(s)**, um die ausgewählte Datei/den ausgewählten Ordner zu entfernen.

File Viewer

Zeigt die Ordnerstruktur für das aktuelle Projekt an. Die Ordnerstruktur kann nur innerhalb von definierten Projekten bearbeitet werden. Sie können also auf der Stammebene des Modus **Projects** keine Dateien hinzufügen, keine Ordner erstellen und keine Elemente löschen.

Zum Erstellen oder Löschen müssen Sie zur Projektliste zurückkehren.

- Klicken Sie auf **Upload file to HDFS**, um eine Datei in das aktuelle Projekt bzw. in den aktuellen Unterordner hochzuladen.
- Klicken Sie auf **Create a new folder**, um unter dem aktuellen Ordner einen neuen Ordner mit dem Namen zu erstellen, den Sie im Dialogfeld **New folder name** angeben.
- Klicken Sie auf **Download file to the local filesystem**, um die ausgewählten Dateien in das lokale Dateisystem herunterzuladen.
- Klicken Sie auf **Delete the selected file(s)**, um die ausgewählten Dateien bzw. Ordner zu entfernen.

Versions

Projekte werden auf der Basis von Änderungen des Datei- und Ordnerinhalts versioniert. Für Änderungen an den Attributen eines Projekts (z. B. die Beschreibung, ob es öffentlich ist und mit wem es gemeinsam genutzt wird) ist keine neue Version erforderlich. Zum Hinzufügen, Ändern oder Löschen von Dateien oder Ordnern ist eine neue Version erforderlich.

Tabelle für die Projektversionssteuerung

In der Tabelle werden die vorhandenen Projektversionen, ihr Erstellungs- und Festschreibungsdatum, die Benutzer, die für die einzelnen Versionen verantwortlich sind, und die übergeordnete Version angezeigt. Die übergeordnete Version ist die Version, auf der die ausgewählte Version basiert.

- Klicken Sie auf **Lock**, um Änderungen am ausgewählten Projektversionsinhalt vorzunehmen.
- Klicken Sie auf **Commit**, um alle an einem Projekt vorgenommenen Änderungen zu speichern und diese Version zum aktuellen sichtbaren Status des Projekts zu machen.
- Klicken Sie auf **Discard**, um alle an einem gesperrten Projekt vorgenommenen Änderungen zu verwerfen und die zuletzt festgeschriebene Version wieder zum sichtbaren Status des Projekts zu machen.
- Klicken Sie auf **Delete Project**, um die ausgewählte Version zu entfernen.

Benutzermanagement

Administratoren können die Rollen von Benutzern und Gruppen über die Seite **Users** verwalten.

Der Inhaltsbereich ist in die ausblendbaren Abschnitte **Details** und **Principals** unterteilt.

Details

Name Ein nicht bearbeitbares Textfeld, in dem der Name des Nutzers angezeigt wird.

Description
Ein bearbeitbares Textfeld, in dem Sie einen erläuternden Text zum Nutzer angeben können.

URL Die URL, mit der Benutzer sich über die Analytic Server-Konsole als Nutzer anmelden können.

Principals

Principals sind Benutzer und Gruppen, die vom Sicherheitsprovider übernommen werden, der während der Konfiguration konfiguriert wird. Sie können die Rolle von Principals in die Administrator- oder Benutzerrolle ändern.

Metrics

Ermöglicht Ihnen, Ressourcengrenzwerte für einen Nutzer zu konfigurieren. Gibt den zurzeit vom Nutzer belegten Plattenspeicherplatz zurück.

- Sie können eine Quote für den maximalen Plattenspeicherplatz für den Nutzer festlegen. Wenn dieser Grenzwert erreicht wird, können keine weiteren Daten für diesen Nutzer auf Platte geschrieben werden, bis genügend Plattenspeicherplatz freigegeben wird, damit die Plattenspeicherplatzbelegung des Nutzers unter die Quote fällt.
- Sie können eine Warnstufe für den Plattenspeicherplatz des Nutzers festlegen. Wenn die Quote überschritten wird, können von Principals keine Analysejobs für diesen Nutzer übergeben werden, bis genügend Plattenspeicherplatz freigegeben wird, damit die Plattenspeicherplatzbelegung des Nutzers unter die Quote fällt.

- Sie können eine maximale Anzahl paralleler Jobs festlegen, die gleichzeitig für diesen Nutzer ausgeführt werden können. Wenn die Quote überschritten wird, können von Principals keine Analysejobs für diesen Nutzer übergeben werden, bis ein zurzeit ausgeführter Job abgeschlossen ist.
- Sie können die maximale Anzahl Felder festlegen, die eine Datenquelle haben kann. Dieser Grenzwert wird bei jedem Erstellen oder Aktualisieren einer Datenquelle geprüft.
- Sie können die maximale Anzahl Datensätze festlegen, die eine Datenquelle haben kann. Dieser Grenzwert wird bei jedem Erstellen oder Aktualisieren einer Datenquelle geprüft, z. B. wenn Sie eine neue Datei hinzufügen oder Einstellungen für eine Datei ändern.
- Sie können die maximale Dateigröße in Megabyte festlegen. Dieser Grenzwert wird beim Hochladen einer Datei geprüft.

Benennungsregeln

Bei allen Elementen, für die ein eindeutiger Name in Analytic Server vergeben werden kann, z. B. Datenquellen und Projekte, gelten die folgenden Regeln für Namen:

- Namen müssen in Objekten desselben Typs eindeutig sein. Beispielsweise kann nicht für zwei Datenquellen der Name **insuranceClaims** vergeben werden, aber eine Datenquelle und ein Projekt könnten jeweils den Namen **insuranceClaims** erhalten.
- Bei Namen muss Groß-/Kleinschreibung beachtet werden. **insuranceClaims** und **InsuranceClaims** beispielsweise werden als eindeutige Namen betrachtet.
- Bei Namen werden führende und abschließende Leerzeichen ignoriert.
- Die folgenden Zeichen sind in Namen ungültig:
  `~, #, %, &, *, {, }, \\, :, <, >, ?, /, |, ", \t, \r, \n`

Kapitel 3. SPSS Modeler-Integration

SPSS Modeler ist eine Data-Mining-Workbench, die über einen visuellen Ansatz für die Analyse verfügt. Jede einzelne Aktion in einem Job, vom Zugriff auf eine Datenquelle über die Zusammenführung von Datensätzen bis zur Ausgabe einer neuen Datei oder zur Erstellung eines Modells, wird durch einen Knoten im Erstellungsbereich dargestellt. Diese Aktionen werden miteinander verknüpft, damit ein analytischer Datenstrom gebildet wird.

Zum Erstellen eines SPSS Modeler-Datenstroms, der für eine Analytic Server-Datenquelle ausgeführt werden kann, beginnen Sie mit einem Analytic Server-Quellenknoten. SPSS Modeler überträgt einen möglichst großen Teil des Datenstroms mit einer Pushback-Operation zurück zu Analytic Server und führt dann, falls erforderlich, eine Pull-Operation für ein Subset der Datensätze durch, um die Ausführung des Datenstroms "lokal" auf dem SPSS Modeler-Server zu beenden. Sie können die maximale Anzahl Datensätze, die SPSS Modeler herunterlädt, in den Analytic Server-Datenstromeigenschaften festlegen.

Wenn Ihre Analyse beendet ist und die Datensätze wieder in HDFS geschrieben wurden, beenden Sie den Datenstrom mit einem Analytic Server-Exportknoten.

Details zu diesen Knoten finden Sie in der SPSS Modeler-Dokumentation.

Unterstützte Knoten

Viele SPSS Modeler-Knoten werden für die Ausführung in HDFS unterstützt, bei der Ausführung bestimmter Knoten gibt es jedoch möglicherweise einige Unterschiede und einige Knoten werden zurzeit nicht unterstützt. In diesem Thema wird die aktuelle Unterstützungsstufe detailliert beschrieben.

Allgemein

- Einige Zeichen, die normalerweise in einem Modeler-Feldnamen in Anführungszeichen zulässig sind, werden von Analytic Server nicht akzeptiert.
- Damit ein Modeler-Datenstrom in Analytic Server ausgeführt werden kann, muss er mit mindestens einem Analytic Server-Quellenknoten beginnen und mit einem einzelnen Modellierungsknoten oder Analytic Server-Exportknoten enden.
- Es wird empfohlen, den Speicher von stetigen Zielen als Speicher für reelle Zahlen und nicht als Speicher für ganze Zahlen festzulegen. Scoring-Modelle schreiben immer reelle Werte in die Ausgabedatendateien für stetige Ziele, während das Ausgabedatenmodell für die Scores dem Speicher des Ziels folgt. Wenn ein stetiges Ziel über einen Speicher für ganze Zahlen verfügt, gibt es daher eine Diskrepanz zwischen den geschriebenen Werten und dem Datenmodell für die Scores und diese Diskrepanz führt zu Fehlern, wenn Sie versuchen, die gescorten Daten zu lesen.

Quelle

- Ein Datenstrom, der mit etwas anderem als einem Analytic Server-Quellenknoten beginnt, wird lokal ausgeführt.

Datensatzoperationen

Alle Datensatzoperationen werden unterstützt, mit Ausnahme von Streaming-ZR- und Space-Time-Boxes-Knoten. Weitere Hinweise zur Funktion dieser Knoten folgen.

Auswählen

- Unterstützt dieselbe Funktionsgruppe wie der Ableitungsknoten.

Stichprobe

- Stichprobenziehung auf Blockebene wird nicht unterstützt.

- Komplexe Methoden der Stichprobenziehung werden nicht unterstützt.

Aggregieren

- Zusammenhängende Schlüssel werden nicht unterstützt. Wenn Sie einen vorhandenen Datenstrom wiederverwenden, der zum Sortieren von Daten konfiguriert ist, und diese Einstellung dann im Aggregatknoten verwenden, ändern Sie den Datenstrom, sodass der Sortierknoten entfernt wird.
- Reihenfolgestatistiken (Median, 1. Quartil, 3. Quartil) werden näherungsweise berechnet und über die Registerkarte für Optimierung unterstützt.

Sortieren

- Die Registerkarte für die Optimierung wird nicht unterstützt.

In einer verteilten Umgebung gibt es eine begrenzte Anzahl von Operationen, bei denen die vom Sortierknoten erstellte Datensatzreihenfolge beibehalten wird.

- Eine Sortierung, auf die ein Exportknoten folgt, erstellt eine sortierte Datenquelle.
- Eine Sortierung, auf die ein Stichprobenknoten mit der ersten Datensatzstichprobenziehung folgt, gibt die ersten *N* Datensätze zurück.
- Eine Sortierung, auf die ein Modellierungsknoten mit dem Ziel **Für sehr große Datasets optimieren** (Neuronales Netz, Linear, C&R-Baum, Quest oder CHAID) folgt, ist ein hilfreiches Muster für das Anzeigen von Datensätzen in zufälliger Reihenfolge durch das Sortieren nach einem abgeleiteten Zufallszahlschlüssel, um eine Verzerrung zu vermeiden, die im Modellerstellungsalgorithmus auftreten kann, wenn die ursprünglichen Datensätze geordnet werden.

Im Allgemeinen sollten Sie einen Sortierknoten so nah wie möglich bei den Operationen platzieren, die die sortierten Datensätze benötigen.

Zusammenführen

- Das Zusammenführen nach Reihenfolge wird nicht unterstützt.
- Die Registerkarte für die Optimierung wird nicht unterstützt.
- Das Platzieren eines Stichprobenknotens oder eines Modellnuggets zwischen einem Analytic Server-Quellenknoten und einem Zusammenführungsknoten wird zurzeit nicht unterstützt. Normalerweise ist es möglich, einen Auswahlknoten anzugeben, um die Funktion des Stichprobenknotens zu ersetzen.
- Analytic Server führt bei Schlüsseln für leere Zeichenfolgen keinen Join durch. Wenn also einer der Schlüssel, mit dem Sie die Zusammenführung durchführen, leere Zeichenfolgen enthält, werden alle Datensätze, die die leere Zeichenfolge enthalten, aus der zusammengeführten Ausgabe gelöscht.
- Zusammenführungsoperationen sind relativ langsam. Wenn in HDFS Speicherplatz verfügbar ist, ist es unter Umständen weniger zeitintensiv, wenn Sie Ihre Datenquellen einmal zusammenführen und die zusammengeführte Quelle in den folgenden Datenströmen verwenden, anstatt die Datenquellen in jedem Datenstrom zusammenzuführen.

R-Transformation

Die R-Syntax im Knoten sollte aus Operationen bestehen, die jeweils nur für einen einzelnen Datensatz ausgeführt werden.

Feldoperationen

Alle Feldoperationen werden unterstützt, mit Ausnahme der Transponier-, Zeitintervall- und Verlaufsknoten. Weitere Hinweise zur Funktionalität von unterstützten Knoten folgen.

Autom. Datenvorbereitung

- Das Trainieren des Knotens wird nicht unterstützt. Die Anwendung der Transformationen in einem trainierten Knoten des Typs **Autom. Datenvorbereitung** auf neue Daten wird unterstützt.

Typ

- Die Spalte **Überprüfen** wird nicht unterstützt.
- Die Registerkarte **Format** wird nicht unterstützt.

Ableiten

- Alle Ableitungsfunktionen werden unterstützt, mit Ausnahme von Sequenzfunktionen.
- Aufteilungsfelder können nicht in demselben Datenstrom abgeleitet werden, der sie als Aufteilungen verwendet. Sie müssen zwei Datenströme erstellen: einen, der das Aufteilungsfeld ableitet, und einen, der das Feld als Aufteilungen verwendet.
- Ein Flagfeld kann nicht allein in einem Vergleich verwendet werden. Das heißt, dass `if (flagField) then ... endif` einen Fehler verursacht. Als Fehlerumgehung kann `if (flagField=trueValue) then ... endif` verwendet werden.
- Wenn der Operator ** verwendet wird, wird empfohlen, den Exponenten als reelle Zahl anzugeben, z. B. x**2,0 anstelle von x**2, damit die Ergebnisse mit den Ergebnissen in Modeler übereinstimmen.

Füller

- Unterstützt dieselbe Funktionsgruppe, die auch vom Ableitungsknoten unterstützt wird.

Klassierung

Die folgende Funktionalität wird nicht unterstützt:

- Optimales Klassieren
- Ränge
- N-Perzentile -> Perzentilmethode: Summe der Werte
- N-Perzentile -> Bindungen: "In aktuellem beibehalten" und "Zufällig zuweisen"
- N-Perzentile ->Benutzerdefiniert N: Werte über 100 und jeder N-Wert, bei dem 100 % N ungleich null ist.

RFM-Analyse

- Die Option "In aktuellem beibehalten" für die Handhabung von Bindungen wird nicht unterstützt. RFM-Aktualitäts-, Häufigkeits- und Geldwertscores stimmen nicht immer mit denen überein, die von Modeler aus denselben Daten berechnet werden. Die Scorebereiche sind identisch, Scorezuweisungen (Klassennummern) können sich jedoch um 1 unterscheiden.

Grafiken

Alle Diagrammknoten werden unterstützt.

Modellierung

Die folgenden Modellierungsknoten werden unterstützt: Linear, Neuronales Netz, C&RT, CHAID, Quest, TCM, TwoStep-AS, STP und Assoziationsregeln. Weitere Hinweise zur Funktionalität dieser Knoten folgen.

Linear Beim Erstellen von Modellen für große Datenmengen und -vielfalt wird das Ziel normalerweise in "Sehr große Datasets" geändert oder es werden Aufteilungen angegeben.

- Fortlaufendes Training vorhandener PSM-Modelle wird nicht unterstützt.
- Das Modellerstellungsziel **Standard** wird nur empfohlen, wenn Aufteilungsfelder so definiert sind, dass die Anzahl an Datensätzen in den einzelnen Aufteilungen nicht "zu groß" ist, wobei die Definition von "zu groß" von der Leistungsstärke einzelner Knoten in Ihrem Hadoop-Cluster abhängt. Im Gegensatz dazu müssen Sie auch darauf bedacht sein, sicherzustellen, dass Aufteilungen nicht so fein definiert sind, dass zu wenige Datensätze für die Erstellung eines Modells vorhanden sind.
- Das Ziel **Boosting** wird nicht unterstützt.
- Das Ziel **Bagging** wird nicht unterstützt.

- Das Ziel **Sehr große Datasets** wird nicht empfohlen, wenn wenige Datensätze vorhanden sind. Oft wird dann entweder kein Modell oder ein vermindertes Modell erstellt.
- **Automatische Datenaufbereitung** wird nicht unterstützt. Dies kann Probleme verursachen, wenn versucht wird, anhand von Daten mit vielen fehlenden Werten ein Modell zu erstellen. Normalerweise würden diese als Teil der automatischen Datenaufbereitung imputiert. Als Problemumgehung kann ein Baummodell oder ein neuronales Netz mit der Einstellung **Erweitert** verwendet werden, um fehlende ausgewählte Werte zu imputieren.
- Die Genauigkeitsstatistik wird für aufgeteilte Modelle nicht berechnet.

Neuronales Netz

Beim Erstellen von Modellen für große Datenmengen und -vielfalt wird das Ziel normalerweise in "Sehr große Datasets" geändert oder es werden Aufteilungen angegeben.

- Fortlaufendes Training vorhandener Standard- oder PSM-Modelle wird nicht unterstützt.
- Das Modellerstellungsziel **Standard** wird nur empfohlen, wenn Aufteilungsfelder so definiert sind, dass die Anzahl an Datensätzen in den einzelnen Aufteilungen nicht "zu groß" ist, wobei die Definition von "zu groß" von der Leistungsstärke einzelner Knoten in Ihrem Hadoop-Cluster abhängt. Im Gegensatz dazu müssen Sie auch darauf bedacht sein, sicherzustellen, dass Aufteilungen nicht so fein definiert sind, dass zu wenige Datensätze für die Erstellung eines Modells vorhanden sind.
- Das Ziel **Boosting** wird nicht unterstützt.
- Das Ziel **Bagging** wird nicht unterstützt.
- Das Ziel **Sehr große Datasets** wird nicht empfohlen, wenn wenige Datensätze vorhanden sind. Oft wird dann entweder kein Modell oder ein vermindertes Modell erstellt.
- Wenn in den Daten viele Werte fehlen, verwenden Sie die Einstellung **Erweitert**, um fehlende Werte zu imputieren.
- Die Genauigkeitsstatistik wird für aufgeteilte Modelle nicht berechnet.

C&R-Baum, CHAID, Quest

Beim Erstellen von Modellen für große Datenmengen und -vielfalt wird das Ziel normalerweise in "Sehr große Datasets" geändert oder es werden Aufteilungen angegeben.

- Fortlaufendes Training vorhandener PSM-Modelle wird nicht unterstützt.
- Das Modellerstellungsziel **Standard** wird nur empfohlen, wenn Aufteilungsfelder so definiert sind, dass die Anzahl an Datensätzen in den einzelnen Aufteilungen nicht "zu groß" ist, wobei die Definition von "zu groß" von der Leistungsstärke einzelner Knoten in Ihrem Hadoop-Cluster abhängt. Im Gegensatz dazu müssen Sie auch darauf bedacht sein, sicherzustellen, dass Aufteilungen nicht so fein definiert sind, dass zu wenige Datensätze für die Erstellung eines Modells vorhanden sind.
- Das Ziel **Boosting** wird nicht unterstützt.
- Das Ziel **Bagging** wird nicht unterstützt.
- Das Ziel **Sehr große Datasets** wird nicht empfohlen, wenn wenige Datensätze vorhanden sind. Oft wird dann entweder kein Modell oder ein vermindertes Modell erstellt.
- Interaktive Sitzungen werden nicht unterstützt.
- Die Genauigkeitsstatistik wird für aufgeteilte Modelle nicht berechnet.

Modellscoring

Alle für die Modellierung unterstützten Modelle werden auch für das Scoring unterstützt. Außerdem werden lokal erstellte Modellnuggets für die folgenden Knoten für das Scoring unterstützt: C&RT, Quest, CHAID, Linear, Neuronales Netz (unabhängig davon, ob es ein Standard-, Boosting- oder Bagging-Modell oder ein Modell für sehr umfangreiche Datasets ist), Regression, C5.0,

Logistisch, Genlin, GLMM, Cox, SVM, Bayes-Netz, TwoStep, KNN, Entscheidungsliste, Diskriminanzanalyse, Selbstlernfunktion, Anomalieerkennung, Apriori, Carma, K-Means, Kohonen, R und Textmining.

- Raw Propensity und Adjusted Propensity werden nicht gescort. Als Problemumgehung können Sie denselben Effekt erzielen, indem Sie die Raw Propensity mithilfe eines Ableitungsknotens mit dem folgenden Ausdruck berechnen: if 'predicted-value' == 'value-of-interest' then 'prob-of-that-value' else 1-'prob-of-that-value' endif
- Beim Scoren eines Modells überprüft Analytic Server nicht, ob alle im Modell verwendeten Felder im Dataset vorhanden sind. Stellen Sie daher vor der Ausführung in Analytic Server sicher, dass dies der Fall ist.

R Die R-Syntax im Nugget sollte aus Operationen bestehen, die jeweils nur für einen Datensatz ausgeführt werden.

Ausgabe

Die Knoten "Matrix", "Analyse", "Data Audit", "Transformieren", "Statistik" und "Mittelwert" werden unterstützt.

Der Tabellenknoten wird unterstützt, indem eine temporäre Analytic Server-Datenquelle geschrieben wird, die die Ergebnisse der vorgeordneten Operationen enthält. Der Tabellenknoten blättert dann durch den Inhalt der Datenquelle.

Export Ein Datenstrom kann mit einem Analytic Server-Quellenknoten beginnen und mit einem anderen Exportknoten als dem Analytic Server-Exportknoten enden, die Daten werden jedoch von HDFS in SPSS Modeler Server und schließlich an die Exportposition verschoben.

Bemerkungen

Die vorliegenden Informationen wurden für Produkte und Services entwickelt, die auf dem deutschen Markt angeboten werden.

Möglicherweise bietet IBM die in dieser Dokumentation beschriebenen Produkte, Services oder Funktionen in anderen Ländern nicht an. Informationen über die gegenwärtig im jeweiligen Land verfügbaren Produkte und Services sind beim zuständigen IBM Ansprechpartner erhältlich. Hinweise auf IBM Lizenzprogramme oder andere IBM Produkte bedeuten nicht, dass nur Programme, Produkte oder Services von IBM verwendet werden können. Anstelle der IBM Produkte, Programme oder Services können auch andere, ihnen äquivalente Produkte, Programme oder Services verwendet werden, solange diese keine gewerblichen oder anderen Schutzrechte von IBM verletzen. Die Verantwortung für den Betrieb von Produkten, Programmen und Services anderer Anbieter liegt beim Kunden.

Für in diesem Handbuch beschriebene Erzeugnisse und Verfahren kann es IBM Patente oder Patentanmeldungen geben. Mit der Auslieferung dieses Handbuchs ist keine Lizenzierung dieser Patente verbunden. Lizenzanforderungen sind schriftlich an folgende Adresse zu richten (Anfragen an diese Adresse müssen auf Englisch formuliert werden):

IBM Director of Licensing
IBM Europe, Middle East & Africa
Tour Descartes
2, avenue Gambetta
92066 Paris La Defense
France

Trotz sorgfältiger Bearbeitung können technische Ungenauigkeiten oder Druckfehler in dieser Veröffentlichung nicht ausgeschlossen werden. Die hier enthaltenen Informationen werden in regelmäßigen Zeitabständen aktualisiert und als Neuausgabe veröffentlicht. IBM kann ohne weitere Mitteilung jederzeit Verbesserungen und/oder Änderungen an den in dieser Veröffentlichung beschriebenen Produkten und/oder Programmen vornehmen.

Verweise in diesen Informationen auf Websites anderer Anbieter werden lediglich als Service für den Kunden bereitgestellt und stellen keinerlei Billigung des Inhalts dieser Websites dar. Das über diese Websites verfügbare Material ist nicht Bestandteil des Materials für dieses IBM Produkt. Die Verwendung dieser Websites geschieht auf eigene Verantwortung.

Werden an IBM Informationen eingesandt, können diese beliebig verwendet werden, ohne dass eine Verpflichtung gegenüber dem Einsender entsteht.

Lizenznehmer des Programms, die Informationen zu diesem Produkt wünschen mit der Zielsetzung: (i) den Austausch von Informationen zwischen unabhängig voneinander erstellten Programmen und anderen Programmen (einschließlich des vorliegenden Programms) sowie (ii) die gemeinsame Nutzung der ausgetauschten Informationen zu ermöglichen, wenden sich an folgende Adresse:

IBM Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
USA

Die Bereitstellung dieser Informationen kann unter Umständen von bestimmten Bedingungen - in einigen Fällen auch von der Zahlung einer Gebühr - abhängig sein.

Die Lieferung des in diesem Dokument beschriebenen Lizenzprogramms sowie des zugehörigen Lizenzmaterials erfolgt auf der Basis der IBM Rahmenvereinbarung bzw. der Allgemeinen Geschäftsbedingungen von IBM, der IBM Internationalen Nutzungsbedingungen für Programmpakete oder einer äquivalenten Vereinbarung.

Alle in diesem Dokument enthaltenen Leistungsdaten stammen aus einer kontrollierten Umgebung. Die Ergebnisse, die in anderen Betriebsumgebungen erzielt werden, können daher erheblich von den hier erzielten Ergebnissen abweichen. Einige Daten stammen möglicherweise von Systemen, deren Entwicklung noch nicht abgeschlossen ist. Eine Gewährleistung, dass diese Daten auch in allgemein verfügbaren Systemen erzielt werden, kann nicht gegeben werden. Darüber hinaus wurden einige Daten unter Umständen durch Extrapolation berechnet. Die tatsächlichen Ergebnisse können davon abweichen. Benutzer dieses Dokuments sollten die entsprechenden Daten in ihrer spezifischen Umgebung prüfen.

Alle Informationen zu Produkten anderer Anbieter stammen von den Anbietern der aufgeführten Produkte, deren veröffentlichten Ankündigungen oder anderen allgemein verfügbaren Quellen. IBM hat diese Produkte nicht getestet und kann daher keine Aussagen zu Leistung, Kompatibilität oder anderen Merkmalen machen. Fragen zu den Leistungsmerkmalen von Produkten anderer Anbieter sind an den jeweiligen Anbieter zu richten.

Aussagen über Pläne und Absichten von IBM unterliegen Änderungen oder können zurückgenommen werden und repräsentieren nur die Ziele von IBM.

Alle von IBM angegebenen Preise sind empfohlene Richtpreise und können jederzeit ohne weitere Mitteilung geändert werden. Händlerpreise können u. U. von den hier genannten Preisen abweichen.

Diese Veröffentlichung dient nur zu Planungszwecken. Die in dieser Veröffentlichung enthaltenen Informationen können geändert werden, bevor die beschriebenen Produkte verfügbar sind.

Diese Veröffentlichung enthält Beispiele für Daten und Berichte des alltäglichen Geschäftsablaufs. Sie sollen nur die Funktionen des Lizenzprogramms illustrieren und können Namen von Personen, Firmen, Marken oder Produkten enthalten. Alle diese Namen sind frei erfunden; Ähnlichkeiten mit tatsächlichen Namen und Adressen sind rein zufällig.

Kopien oder Teile der Beispielprogramme bzw. daraus abgeleiteter Code müssen folgenden Copyrightvermerk beinhalten:

Diese Veröffentlichung enthält Beispiele für Daten und Berichte des alltäglichen Geschäftsablaufs. Sie sollen nur die Funktionen des Lizenzprogramms illustrieren und können Namen von Personen, Firmen, Marken oder Produkten enthalten. Alle diese Namen sind frei erfunden; Ähnlichkeiten mit tatsächlichen Namen und Adressen sind rein zufällig.

Kopien oder Teile der Beispielprogramme bzw. daraus abgeleiteter Code müssen folgenden Copyrightvermerk beinhalten:

© (Name Ihrer Firma) (Jahr). Teile des vorliegenden Codes wurden aus Beispielprogrammen der IBM Corp. abgeleitet.

Marken

IBM, das IBM Logo und ibm.com sind Marken oder eingetragene Marken der IBM Corp in den USA und/oder anderen Ländern. Weitere Produkt- und Servicenamen können Marken von IBM oder anderen Unternehmen sein. Eine aktuelle Liste der IBM Marken finden Sie auf der Webseite "Copyright and trademark information" unter www.ibm.com/legal/copytrade.shtml.

Adobe, das Adobe-Logo, PostScript und das PostScript-Logo sind Marken oder eingetragene Marken der Adobe Systems Incorporated in den USA und/oder anderen Ländern.

IT Infrastructure Library ist eine eingetragene Marke der Central Computer and Telecommunications Agency. Die Central Computer and Telecommunications Agency ist nunmehr in das Office of Government Commerce eingegliedert worden.

Intel, das Intel-Logo, Intel Inside, das Intel Inside-Logo, Intel Centrino, das Intel Centrino-Logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium und Pentium sind Marken oder eingetragene Marken der Intel Corporation oder ihrer Tochtergesellschaften in den USA oder anderen Ländern.

Linux ist eine eingetragene Marke von Linus Torvalds in den USA und/oder anderen Ländern.

Microsoft, Windows, Windows NT und das Windows-Logo sind Marken der Microsoft Corporation in den USA und/oder anderen Ländern.

ITIL ist eine eingetragene Marke, eine eingetragene Gemeinschaftsmarke des Cabinet Office und eine eingetragene Marke, die beim US Patent and Trademark Office registriert ist.

UNIX ist eine eingetragene Marke von The Open Group in den USA und anderen Ländern.

Java und alle auf Java basierenden Marken und Logos sind Marken oder eingetragene Marken der Oracle Corporation und/oder ihrer verbundenen Unternehmen.

Cell Broadband Engine wird unter Lizenz verwendet und ist eine Marke der Sony Computer Entertainment, Inc. in den USA und/oder anderen Ländern.

Linear Tape-Open, LTO, das LTO-Logo, Ultrium und das Ultrium-Logo sind Marken von HP, IBM und Quantum in den USA und anderen Ländern.