IBM SPSS Analytic Server
Version 3.2.1

Verwaltung

Hinweis

Vor Verwendung dieser Informationen und des darin beschriebenen Produkts sollten die Informationen unter „Bemerkungen" auf Seite 25 gelesen werden.

Produktinformation

Diese Ausgabe bezieht sich auf Version 3, Release 2, Modifikation 1 von IBM SPSS Analytic Server und alle nachfolgenden Releases und Modifikationen, bis dieser Hinweis in einer Neuausgabe geändert wird.

Inhaltsverzeichnis

Kapitel 1. Nutzermanagement

Nutzer stellen eine übergeordnete Einteilung der Benutzer, Projekte und Datenquellen bereit, sodass Objekte nicht von mehreren Nutzern gemeinsam genutzt werden können. Jeder Benutzer greift auf das System in dem Kontext eines Nutzers zu, dem er zugewiesen wurde.

Die Verwaltung der Nutzer und das Zuweisen von Benutzern zu Nutzern erfolgt über die Analytic Server-Konsole. Die Ansicht der Nutzerseite hängt von der Rolle des Benutzers ab, der an der Konsole angemeldet ist:

- "Superuser-Administrator", der während der Installation des Nutzermanagers definiert wird. Nur dieser Benutzer kann neue Nutzer erstellen und die Eigenschaften der Nutzer bearbeiten.
- Benutzer mit Administratorrolle können die Eigenschaften der Nutzer bearbeiten, als die sie angemeldet sind.
- Benutzer mit Benutzerrolle können Nutzereigenschaften nicht bearbeiten. Die Nutzerseite wird ihnen nicht angezeigt.
- Benutzer mit Leserrolle können weder Datenquellen bearbeiten noch sich an der Analytic Server-Konsole anmelden.

Administratoren können auf die Projekt- und die Datenquellenseite zugreifen und alle Projekte oder Datenquellen zur Bereinigung und Administration verwalten. Weitere Informationen finden Sie im Benutzerhandbuch zu IBM® SPSS Analytic Server.

Nutzerliste

Auf der Hauptseite mit den Nutzern werden die vorhandenen Nutzer in einer Tabelle angezeigt. Nur der Administrator mit Superuserberechtigung darf diese Seite bearbeiten.

- Klicken Sie auf den Namen eines Nutzers, um die zugehörigen Details anzuzeigen und die Eigenschaften zu bearbeiten.
- Klicken Sie auf die URL eines Nutzers, um die Konsole im Kontext des betreffenden Nutzers zu öffnen.

  Anmerkung: Sie werden von der Konsole abgemeldet und Sie müssen sich mit den für den Nutzer gültigen Berechtigungsnachweisen anmelden.
- Geben Sie einen Suchbegriff in den Suchbereich ein, um die Liste zu filtern, damit nur Nutzer angezeigt werden, deren Name den Suchbegriff enthält.
- Klicken Sie auf **New**, um einen neuen Nutzer mit dem Namen zu erstellen, den Sie im Dialogfeld **Add new tenant** angeben. In „Benennungsregeln" auf Seite 2 finden Sie Informationen zu den Beschränkungen bei Namen, die Sie für Nutzer vergeben können.
- Klicken Sie auf **Delete**, um den/die ausgewählten Nutzer zu entfernen.
- Klicken Sie auf **Refresh**, um die Liste zu aktualisieren.

Details zu einzelnen Nutzern

Der Inhaltsbereich ist in mehrere ausblendbare Abschnitte unterteilt.

Details

Name Ein 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, über die Benutzer sich mithilfe der Analytic Server-Konsole am Nutzer anmelden und SPSS Modeler Server konfigurieren. Details zum Konfigurieren von SPSS Modeler finden Sie im Handbuch *IBM SPSS Analytic Server Installation und Konfiguration*.

Status Nutzer mit dem Status **Active** werden zurzeit verwendet. Wenn der Status des Nutzers auf **Inactive** gesetzt wird, wird verhindert, dass sich Benutzer bei diesem Nutzer anmelden können. Es werden jedoch keine der zugrunde liegenden Informationen gelöscht.

Principals

Principals sind Benutzer und Gruppen, die von dem Sicherheitsprovider übernommen werden, der während der Installation konfiguriert wird. Sie können Principals einem Nutzer als Administratoren, Benutzer oder Leser hinzufügen.

- Durch Eingeben eines Suchbegriffs in das Textfeld wird nach Benutzern und Gruppen gefiltert, deren Name den Suchbegriff enthält. Wählen Sie **Administrator**, **User** oder **Reader** in der Dropdown-Liste aus, um die Rolle der Benutzer innerhalb des Nutzers festzulegen. Klicken Sie auf **Add participant**, um die Benutzer der Liste der Autoren hinzuzufügen.
- Zum Entfernen eines Teilnehmers wählen Sie einen Benutzer oder eine Gruppe in der Mitgliedsliste aus und klicken Sie auf **Remove participant**.

Metrics

Ermöglicht es 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 Dateigröße in Megabyte festlegen. Dieser Grenzwert wird beim Hochladen einer Datei geprüft.

Security provider configuration

Hier können Sie den Provider für die Benutzerauthentifizierung angeben. Bei Angabe von **Default** wird der während der Installation und Konfiguration konfigurierte Standardprovider des Nutzers verwendet. Bei Angabe von **LDAP** können Sie Benutzer über einen externen LDAP-Server wie beispielsweise Active Directory oder OpenLDAP authentifizieren. Geben Sie die Einstellungen für den Provider und optional Filtereinstellungen an, um die im Abschnitt **Principals** verfügbaren Benutzer und Gruppen zu steuern.

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:

- Innerhalb eines Nutzers müssen Namen 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 2. Aufnehmen neuer Benutzer

Bitten Sie die Benutzer, zu `http://<Host>:<Port>/<Kontextstammverzeichnis>/admin/<Nutzer>` zu navigieren und ihren Benutzernamen und ihr Kennwort für die Anmeldung an der Analytic Server-Konsole einzugeben.

Anmerkung: Die Eingabe des Benutzernamens während der Anmeldeaufforderung der Analytic Server-Konsole erfolgt ohne das Suffix des Realmnamens. Infolgedessen wird Benutzern bei Verwendung mehrerer Realms die Dropdown-Liste **Realms** angezeigt, aus der sie den entsprechenden Realm auswählen können. Wenn nur ein Realm definiert ist, wird Benutzern bei der Anmeldung bei Analytic Server die Dropdown-Liste **Realms** nicht angezeigt.

<Host>
Die Adresse des Analytic Server-Hosts.

<Port>
Der Port, an dem Analytic Server empfangsbereit ist. Der Standardwert ist **9080**.

<Kontextstammverzeichnis>
Das Kontextstammverzeichnis von Analytic Server. Der Standardwert ist **analyticserver**.

<Nutzer>
In einer Multi-Tenant-Umgebung der Nutzer, zu dem Sie gehören. In einer Umgebung mit einem einzelnen Nutzer lautet der Nutzer **`ibm`**.

Wenn die Hostmaschine beispielsweise die IP-Adresse 9.86.44.232 hat, Sie einen Nutzer "mycompany" erstellt haben, diesem Nutzer Benutzer hinzugefügt haben und für die anderen Einstellungen die Standardwerte übernommen haben, sollten Benutzer zu `http://9.86.44.232:9080/analyticserver/admin/mycompany` navigieren, um auf die Analytic Server-Konsole zuzugreifen.

Kapitel 3. Analytic Server-Jobnamen

Analytic Server erstellt MapReduce- und Spark-Jobs, die über die Benutzerschnittstelle des Ressourcenmanagers Ihres Hadoop-Clusters überwacht werden können.

Der MapReduce-Jobname hat die folgende Struktur:

```
AS/{Nutzername}/{Benutzername}/{Algorithmusname}
```

{Nutzername}
: Der Name des Nutzers, unter dem der Job ausgeführt wird.

{Benutzername}
: Der Benutzer, der den Job angefordert hat.

{Algorithmusname}
: Der primäre Algorithmus im Job. Beachten Sie, dass ein einzelner Datenstrom möglicherweise mehrere MapReduce-Jobs generiert; ebenso können mehrere Operationen innerhalb eines Datenstroms in einem einzelnen MapReduce-Job enthalten sein.

Alle MapReduce-Jobs werden in der Benutzerschnittstelle des Ressourcenmanagers angezeigt. Eine einzelne Spark-Anwendung wird für jeden Analytic Server gestartet. Öffnen Sie die Benutzerschnittstelle der Spark-Anwendung, um die Spark-Jobs zu überwachen. (Die Jobnamen werden in der Beschreibungsspalte angezeigt.)

Kapitel 4. Best Practices und Empfehlungen für IBM SPSS Analytic Server

In den folgenden Abschnitten werden Best Practices und Empfehlungen für Analytic Server hinsichtlich Datenquellen, Clusterkonfiguration und IBM SPSS Modeler-Datenströme bereitgestellt.

Datenquellen

Analytic Server unterstützt die folgenden Datenquellentypen:

- Dateibasierte Datenquellen wie z. B. Dateien mit Trennzeichen, Dateien mit festgelegtem Text und Microsoft Excel-Dateien.
- Relationale Datenbanken wie z. B. Db2, Oracle, Microsoft SQL Server, Teradata, Postgres, Netezza, MySQL und Amazon Redshift.
- Hive/HCatalog-Datenquellen, die alle integrierten Datentypen (z. B. ORC und Parquet) sowie jeden angepassten Datentyp enthalten, für den eine entsprechende Implementierung des Hive-Parallel-Seriell- und Hive-Seriell-Parallel-Umsetzers verfügbar ist. Außerdem kann Analytic Server für den Zugriff auf NoSQL-Datenbanken konfiguriert werden, wie z. B. HBase, MongoDB, Accumulo, Cassandra, Oracle NoSQL und andere Datenbanken, für die eine entsprechende Hive-Speicherhandlerimplementierung verfügbar ist.
- Datenquellen des georäumlichen Typs (auf der Basis von Shapefiles und Kartenservices).

Analytic Server-Einschränkungen bei Hive/HCatalog-Datenquellen

- Wenn Hive-Pushback für den Auswahlknoten von SPSS Modeler erforderlich ist, kann der Filterausdruck nur partitionierte Spalten des Typs STRING referenzieren. Ab Analytic Server 3.0 ist Datentypunterstützung für die folgenden partitionierten Spalten verfügbar: TINYINT, SMALLINT, INT, BIGINT. Der statische Filterausdruck, der für die Hive-Datenquelle angegeben wird, kann Filterausdrücke für partitionierte Spalten jeden Datentyps enthalten.
- Analytic Server unterstützt nicht Datenquellen, die auf Hive-Ansichten basieren.

Clusterkonfiguration - Sicherheit

Kerberos-Identitätswechsel

Vor Version 3.0.1 verwendeten Analytic Server-Instanzen einen Benutzerprincipalnamen im Analytic Server-Chiffrierschlüssel zum Authentifizieren von HDFS-Operationen, wenn Kerberos-Sicherheit aktiviert war. Ab Version 3.0.1 verwendet Analytic Server einen Benutzerprincipalnamen im Analytic Server-Chiffrierschlüssel zusammen mit dem anfordernden Benutzernamen (des Benutzers, der die REST-Anforderung ausgibt), um HDFS-Operationen zu authentifizieren, die Kerberos-Identitätswechsel verwenden. Analytic Server 3.0.1 oder höher muss dem HDFS (oder den Hive-Servicekonfigurationen) bei Ausführung in einem Kerberos-aktivierten Cluster Konfigurationsattribute für Identitätswechsel hinzufügen. Bei HDFS müssen der HDFS-Datei `core-site.xml` die folgenden Eigenschaften hinzugefügt werden:

```
hadoop.proxyuser.<Analytic_Server-Service-Principal-Name> .hosts = *
hadoop.proxyuser.<Analytic_Server-Service-Principal-Name> .groups = *
```

Dabei ist `<Analytic_Server-Service-Principal-Name>` der Standardwert von `as_user`, der im Konfigurationsfeld `Analytic_Server_User` von Analytic Server angegeben ist.

Die folgenden Eigenschaften müssen der HDFS-Datei `core-site.xml` hinzugefügt werden, wenn von HDFS über Hive/HCatalog auf Daten zugegriffen wird:

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

Realmübergreifende Kerberos-Authentifizierung

Analytic Server unterstützt realmübergreifende Kerberos-Authentifizierung. Um diese Funktion zu aktivieren, müssen Sie zuerst sicherstellen, dass die realmübergreifende KDC-Authentifizierung aktiviert ist, und dann dem Ambari-Konfigurationsabschnitt **Custom analytics.cfg** von Analytic Server die folgende Einstellung hinzufügen:

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

Clusterkonfiguration - Leistungsoptimierungseinstellungen und -ergebnisse

Spark-Konfiguration

Analytic Server verwendet den Modus `yarn-client` für die Interaktion mit YARN und für die Ausführung von Spark-Jobs im Hadoop-Cluster.

Angepasste Analytic Server-Konfiguration:

- Ambari-Einstellungen werden im Ambari-Konfigurationsabschnitt **Custom analytics.cfg** von Analytic Server definiert.
- Cloudera-Einstellungen befinden sich im Cloudera Manager-Abschnitt **Analytic Server Advanced Configuration Snippet (Safety Valve) for analyticserver-conf/config.properties**.

1. Ziehen Sie eine Erhöhung des Werts der Konfigurationseinstellung **spark.driver.memory** durch Hinzufügen eines Konfigurationselements in der angepassten Analytic Server-Konfiguration in Erwägung (der Standardwert ist `1g`, falls die Einstellung nicht explizit angegeben ist). Beispiel:

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

2. Wählen Sie eine der folgenden Ressourcennutzungsoptionen für die die Verwendung von Analytic Server mit Spark aus.

   - **Option A: Konfiguration mit statischer Ressourcenzuordnung**

     Es gibt 3 Parameter, die in der angepassten Analytic Server-Konfiguration konfiguriert werden müssen:

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

     Im Folgenden werden die Schritte zum Bestimmen der Parameterwerte beschrieben.

     a. Ermitteln Sie den Prozentsatz für CPU und Speicher, den Analytic Server Spark permanent zuordnen kann. Dies führt zu einer bestimmten Anzahl Kerne (C) und einem Festbetrag für den auf jedem Computer verwendbaren Speicher (M).

     b. Ermitteln Sie die Anzahl Executor (E), die jeder Computer ausführen kann. Diese Executor werden als separate Hadoop-Container (Prozesse) auf jedem Clusterknoten ausgeführt. In der Regel ist ein Wert größer als 2 angemessen, er muss jedoch unter der Gesamtzahl Kerne liegen. Der Speicher, der für Spark zugeordnet wird, wird zwischen diesen Executor aufgeteilt. Daher senkt die Auswahl eines hohen Werts für diesen Parameter die Speichermenge, die für jeden Container zugeordnet wird.

     c. Ermitteln Sie die Anzahl Kerne, die für jeden Executor verwendet werden (CE). In der Regel ist dieser Wert C/E (die Anzahl Kerne von jedem Computer, die für die Spark-Anwendung zugeordnet werden, dividiert durch die Gesamtzahl Executor).

     d. Ermitteln Sie die Speichermenge, die für jeden Executor verwendet wird (ME). Dies ist in der Regel M/E.

     Anmerkung: Die Anzahl verwendeter Executor und Kerne muss so ausgewogen sein, dass die Menge für jeden Executorspeicher größer als 3G * CE ist. Jeder Kern von jedem Executor muss über mindestens 3 G Speicher verfügen, der als Speicher oder Berechnungsspeicher verwendet wird.

```
spark.executor.instances = <E>*N /<E> //in Schritt b festgelegter Wert; N ist die Anzahl Rechenknoten
spark.executor.cores = <CE> //in Schritt c festgelegter Wert
spark.executor.memory = <ME> //in Schritt d festgelegter Wert
```

Abbildung 1. Spark-Einstellungen 'Custom analytics.cfg'

- **Option B: Konfiguration mit dynamischer Ressourcenzuordnung**

  Bei Verwendung dieser Option werden alle von YARN zugeordneten Executor entsprechend den im gesamten Cluster tatsächlich verfügbaren Ressourcen dynamisch vergrößert/verkleinert.

  Mindestkonfiguration:

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

  Standardkonfiguration:

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

  Hinweise:

  – `spark.executor.instances = <E>` sollte nicht verwendet werden, andernfalls wird eine statische Ressourcenzuordnung verwendet.

  – Die unter Option A erörterten Überlegungen hinsichtlich Kernen und Speicherwerten des Executors gelten auch hier.

3. Sie können den Spark-Cache in der angepassten Analytic Server-Konfiguration mit den folgenden Einstellungen inaktivieren:

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

Abbildung 2. Spark-Cacheeinstellungen 'Custom analytics.cfg'

Der Spark-Cache darf nicht inaktiviert werden, wenn große IBM SPSS Modeler-Datenströme verwendet werden. Die Inaktivierung des Spark-Cache führt in diesem Fall zu einer langsameren Ausführung von Datenströmen, vermeidet jedoch abnormale Speicherbedingungen, die auftreten können, wenn die pro Executor angegebene Speichermenge klein ist.

JVM-Konfiguration

Ambari-Einstellungen:

1. Legen Sie die Speichermenge, die der Server für die lokale Verarbeitung verwenden kann, in der Ambari-Konfiguration von Analytic Server fest. Der Standardwert (2 GB) kann für kleine bis mittlere Datenströme sicher verwendet werden, bei größeren Datenströmen sollte jedoch ein höherer Wert für die Heapspeichergröße (z. B. 10 GB) verwendet werden.

   Analytic Server > **Konfiguration** > **Advanced analytic-jvm-options**
2. Ersetzen Sie `-Xmx2048M` durch `-Xmx10G`, speichern Sie die Konfiguration und starten Sie Analytic Server erneut.

Abbildung 3. Einstellung 'Advanced analytic-jvm-options'

Cloudera-Einstellungen:

1. Navigieren Sie in Cloudera Manager zur Registerkarte **Configuration** des Analytic Server-Service und aktualisieren Sie das Steuerelement `jvm-options`, um die Speichermenge festzulegen, die der Server für die lokale Verarbeitung verwenden kann. Der Standardwert (2 GB) kann für kleine bis mittlere Datenströme sicher verwendet werden, bei größeren Datenströmen sollte jedoch ein höherer Wert für die Heapspeichergröße (z. B. 10 GB) verwendet werden.

   Analytic Server-Service > **Konfiguration** > **jvm-options**
2. Ersetzen Sie `-Xmx2048M` durch `-Xmx10G`, speichern Sie die Konfiguration und starten Sie Analytic Server erneut.

Konfiguration von Yarn MapReduce2:

- Wenn MapReduce-Jobs parallel zu Spark-Jobs für die Analytic Server-Ausführung erforderlich sind, muss der Yarn-Cluster so konfiguriert werden, dass mindestens 4 GB Speicher pro Yarn-Container vorhanden sind.

Zookeeper-Konfiguration:

- Cloudera erfordert die manuelle Aktualisierung der Zookeeper-Konfiguration. Weitere Informationen finden Sie unter https://www.cloudera.com/documentation/enterprise/5-4-x/topics/cdh_ig_zookeeper_server_maintain.html.
- Wenn Sie komplexe SPSS Modeler-Datenströme oder Wide Data (eine große Anzahl Felder) verwenden, schlagen Jobs möglicherweise aufgrund einer unterbrochenen Analytic Server-Zookeeper-Verbindung fehl. Das Problem resultiert aus der hohen Programmgröße, die der SPSS Modeler-Server an Analytic Server sendet. Es ist unwahrscheinlicher, dass das Problem in Analytic Server 3.0 (oder höher) auftritt. Beheben Sie das Problem mithilfe der folgenden Schritte:
  1. Navigieren Sie in der Ambari-Konsole zur Zookeeper-Service-Registerkarte **Configs**, fügen Sie der Vorlage `zookeeper-env` unter **Advanced zookeeper-env** die folgende Zeile hinzu und starten Sie den Zookeeper-Service erneut.

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

Abbildung 4. Einstellungen der Vorlage 'zookeeper-env'

2. Navigieren Sie in der Ambari-Konsole zur Analytic Server-Service-Registerkarte **Configs**, fügen Sie Folgendes zu **Advanced analytics-jvm-options** hinzu und starten Sie den Analytic Server-Service anschließend erneut.

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

Abbildung 5. Einstellung 'Advanced analytics-jvm-options'

Anmerkung: Wenn das Problem bestehen bleibt, erhöhen Sie den Wert für `-Djute.maxbuffer` an beiden Stellen von `2097152` auf `4194304`.

IBM SPSS Modeler-Datenstromempfehlungen

Anmerkung: Die meisten der folgenden Empfehlungen gelten auch für Small Data.

Prototyp mit Small Data

Wenn Sie mit einem Datenstrom experimentieren, fügen Sie häufig ein paar Knoten hinzu, testen den Datenstrom bis zu diesem Punkt, fügen möglicherweise einen Knoten zum Prüfen von tabellarischen oder grafischen Ausgaben hinzu und setzen anschließend die Erstellung des Datenstroms fort. In der Regel können Sie es sich nicht leisten, bei jedem Test Ihres Datenstroms einen Durchlauf Ihrer Big Data durchzuführen.

Durch die Erstellung einer geeigneten Datenstichprobe Ihrer Big Data können Sie den Datenstrom anhand tatsächlicher Daten testen, und zwar ohne den Zeitaufwand, der für einen vollständigen Datendurchlauf erforderlich ist. Die Datenstichprobe muss ausreichend Daten enthalten, um Ihren Datenstrom erfolgreich ausführen zu können. Wenn Sie z. B. Transaktionen in Kölner Filialen analysieren wollen, muss Ihre Datenstichprobe Transaktionen aus den entsprechenden Filialen enthalten.

Nach der Stichprobenziehung können Sie Folgendes ausführen:

- Erstellen eines Cache der Datenstichprobe in dem Cluster, in dem sich die Big Data befinden.
  Vorteile: Einfach und Quellenknoten müssen nicht gewechselt werden.
  Nachteile: Der Cache wird gelöscht, wenn die Sitzung beendet ist.
- Erstellen einer neuen Analytic Server-Datenquelle, die die Datenstichprobe enthält.
  Vroteile: Permanente Datenquelle.
  Nachteile: Erfordert die Bearbeitung/den Wechsel von Quellenknoten.
- Herunterladen der Datenstichprobe auf Ihr lokales System und Erstellen einer lokalen Datenquelle.

Vorteile: Verbraucht keine Clusterressourcen bei Prototyperstellung; der SPSS Modeler-Client ist effizienter als Analytic Server, wenn Sie mit Small Data arbeiten.

Nachteile: Erfordert den Wechsel von Quellenknoten.

Erstellen von separaten Typ- und Filterknoten aus den Quellenknoten

Jeder SPSS Modeler-Quellenknoten hat auch die kombinierte Funktionalität der Filter- und Typknoten. Dies ist nützlich, um den Erstellungsbereich zu optimieren, erschwert jedoch den Wechsel zu unterschiedlichen Quellenknotentypen. Außerdem lässt sich hierbei nicht erkennen, ob Typ- und Filteroperationen auftreten.

Anordnen von Filter- und Auswahlknoten so nah wie möglich am Quellenknoten

Dies reduziert die Anzahl Datensätze in nachgeordneten Operationen.

Vermeiden des Sortierknotens, wo immer möglich

Analytic Server unterstützt keine Optimierungen in Knoten, die von der Sortierung von Daten abhängen (z. B. der Zusammenführungsknoten). Daher ist ein Sortierknoten in der Mitte des Datenstroms nicht wirklich nützlich. Der Sortierknoten ist jedoch nützlich, wenn sofort auf ihn ein Stichprobenknoten folgt, um die ersten N (oder letzten N) Datensätze abzurufen.

Berechnen nur der Felder, die verwendet werden

Vermeiden Sie es, ein Feld zu berechnen und sofort anschließend zu filtern.

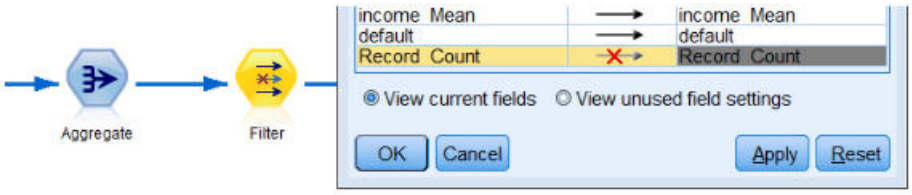

Abbildung 6. Modeler-Feldoptionen

Vermeiden Sie möglichst die Erstellung zahlreicher temporärer Felder, wobei die Ausdrücke aber weiterhin gut verständlich bleiben sollten. Definieren Sie z. B. anstelle des folgenden Beispiels:

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

das folgende Beispiel:

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

Ein solches Umsetzen von temporären Feldern in Inlineausdrücke kann die Leistung steigern, wenn eine große Anzahl Felder transformiert wird.

Festlegen des Speichers in der Datenquelle

Operationen, die den Speichertyp eines Felds in der Mitte des Datenstroms ändern (z. B. von Zeichenfolge in ganze Zahl), können der Gesamtleistung abträglich sein. Sie können den Speicher für Felder beim Definieren von Datenquellen in der Analytic Server-Konsole festlegen, um die Wiederholung dieser Konvertierungen zu vermeiden.

Verwenden von SPSS Modeler bei der Arbeit mit Small Data

Bearbeiten Sie Big Data mit Analytic Server und schließen Sie Berechnungen für Small Data dann mit SPSS Modeler ab.

Wählen Sie die geeigneten Datenstromeigenschaften für Analytic Server aus

Konfigurieren Sie relevante Datenstromeigenschaften (**Tools** > **Options** > **Stream Properties** > **Analytic Server**) und entscheiden Sie, ob die Datenverarbeitung in Analytic Server ausgesetzt und in SPSS Modeler fortgesetzt werden darf (wenn ein Knoten in Analytic Server nicht ausgeführt werden kann).

SPSS Modeler ist standardmäßig so konfiguriert, dass in dieser Situation ein Fehler gemeldet und die Ausführung gestoppt wird. Sie können den Fehler umgehen, indem Sie die Einstellung von `Error` in `Warn` ändern und den Grenzwert für das Datenvolumen anpassen, das in SPSS Modeler verarbeitet werden kann. Beispielsweise können Sie den Standardwert der Datenübertragungsrate (10.000 Datensätze) bei Bedarf aktualisieren. Beachten Sie, dass dieser Grenzwert auch gilt, wenn Sie Ergebnisse anzeigen, die den SPSS Modeler-Tabellenknoten verwenden. Wenn der Grenzwert überschritten wird, meldet SPSS Modeler, dass der Datenabruf den Grenzwert überschritten hat, der in den Datenstromeigenschaften festgelegt ist.

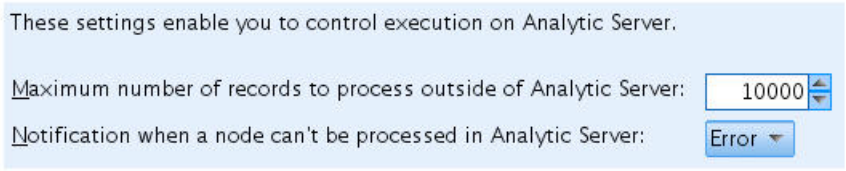

Abbildung 7. Analytic Server-Einstellungen

Verwenden der Analytic Server-Quellenknoten

Analytic Server kann Verbindungen zu verschiedenen Datenbankdatenquellen herstellen, SPSS Modeler erfordert jedoch, dass alle Quellenknoten Analytic Server-Quellenknoten sind (damit der gesamte Datenstrom als Analytic Server-Job ausgeführt wird). Der Datenbankquellenknoten muss in einen Analytic Server-Quellenknoten geändert werden und in der Analytic Server-Konsole muss eine Analytic Server-Datenbankdatenquelle erstellt werden, damit der gesamte Datenstrom in Analytic Server ausgeführt werden kann.

Berücksichtigen, wie nicht unterstützte Knoten verwendet werden

Analytic Server unterstützt nicht alle Knoten (der Transponierknoten ist dafür ein gutes Beispiel). Sie können die Ergebnisse einer Transponieroperation mit dem Rest des Datenstroms zusammenführen und ihn in Analytic Server ausführen lassen, indem ein untergeordneter Datenstrom, der einen Transponierknoten enthält, in eine Analytic Server-Datenquelle ausgegeben wird, die einen Analytic Server-Exportknoten verwendet. Sie können dann einen Analytic Server-Quellenknoten anhängen, wo der Datenstrom unterbrochen wurde, um in Analytic Server zu schreiben.

Anmerkung: Die Transponieroperation eignet sich für einmalige oder selten ausgeführte Operationen, sollte jedoch nicht für routinemäßige Datenstromoperationen verwendet werden.

Feststellen, ob ein Datenstrom in Analytic Server weiterhin funktioniert, bevor er ausgeführt wird

Wählen Sie nach der Vorbereitung eines Datenstroms zur Ausführung in Analytic Server einen Endknoten aus und prüfen Sie mithilfe der SPSS Modeler-Vorschaufunktion (das Steuerelement **Preview Run** in der Symbolleiste), ob an der Ausführung des Endknotens beteiligte Knoten in Analytic Server funktionieren (ohne den Datenstrom auszuführen). Probleme werden im Nachrichtenfenster aufgelistet.

Kombinieren von Back-to-Back-Zusammenführungsoperationen

Kombinieren Sie eine Reihe von Zusammenführungsknoten zu einem einzelnen Knoten, wenn sie die gleichen Schlüssel und den gleichen Jointyp haben.

Kombinieren identischer untergeordneter Datenströme

Versuchen Sie, identische untergeordnete Datenströme möglichst zu kombinieren, vor allem, wenn sie kostenintensive Operationen (z. B. Zusammenführung und Sortierung) enthalten. SPSS Modeler führt diese Operationen einmal aus und verwendet den Cache, um die Leistung zu verbessern. Im folgenden Beispiel sind die Datenströme bis zum Knoten **newField** identisch.

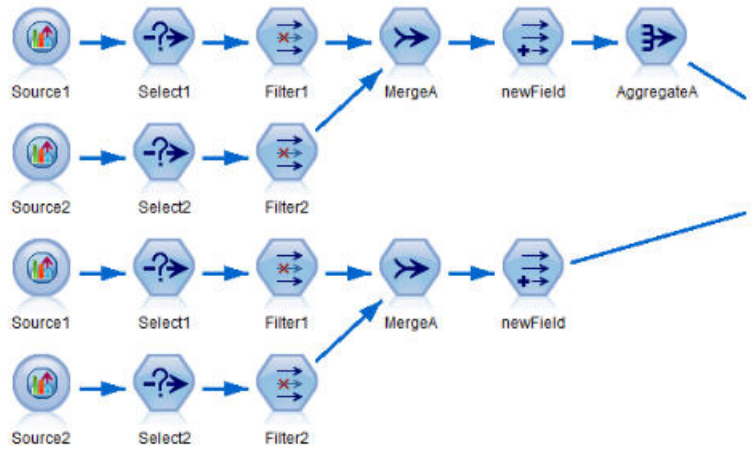

Abbildung 8. Beispieldatenstrom

Verwaltungstechnisch ist es effizienter, wenn der untergeordnete Datenstrom stattdessen wie folgt strukturiert ist:

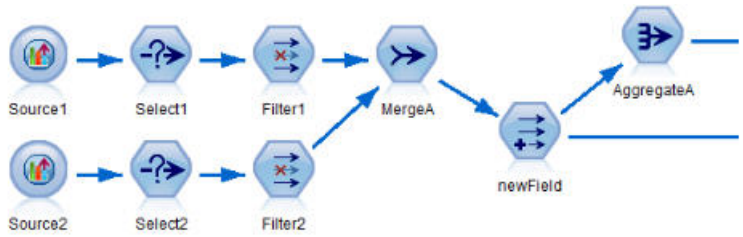

Abbildung 9. Beispieldatenstrom

Entfernen von zusätzlichen Typknoten

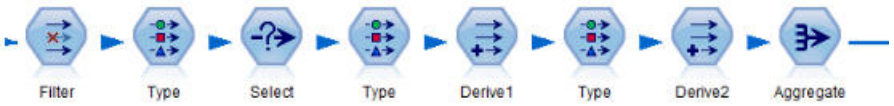

Abbildung 10. Beispieldatenstrom

Vermeiden Sie unnötige Typknoten bei der Ausführung mit Analytic Server. Die Operation `Read Values` des Typknotens startet einen MapReduce-Job. Dies ist in der Regel eine einmalige Einsparung, außer Sie löschen die Typknotenwerte.

Vollständiges Dokumentieren jeden Datenstroms

Das folgende Beispiel zeigt einen komplexen Datenstrom, der eine Anzahl untergeordneter Datenströme enthält.

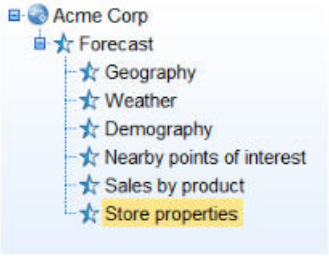

Abbildung 11. Beispiel eines untergeordneten Datenstroms

In solchen Fällen ist es wichtig, die Superknoten ordnungsgemäß zu benennen und den Datenstrom zu dokumentieren (so wie Sie Code dokumentieren). Ein klarer Kommentar kann anderen Analysten, die den Datenstrom lesen oder verwalten, wichtige Informationen bereitstellen. Beispiel:

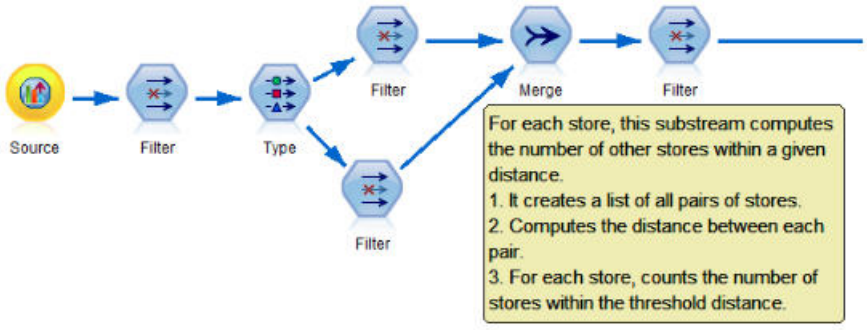

Abbildung 12. Datenstrombeispiel mit Kommentaren

Verwenden von SPSS Modeler-Caches zum schnellen Speichern von Zwischenergebnissen bei der Entwicklung von Datenströmen

In Datenströmen, die für Analytic Server ausgeführt werden, funktioniert das Knotencaching, indem die Daten in einem bestimmten Teil des Datenstroms in temporären Dateien in HDFS gespeichert werden (im Gegensatz zum Speichern auf dem SPSS Modeler-Server). Caches funktionieren gut mit Big Data und können sicher in Datenströmen verwendet werden, die in Analytic Server ausgeführt werden.

Kapitel 5. Fehlerbehebung

Analytic Server stellt eine Reihe von nützlichen Tools für die Problembestimmung bereit.

Protokollierung

Analytic Server erstellt Kundenprotokolldateien und Tracedateien, die beim Diagnostizieren von Problemen hilfreich sind. Bei der Liberty-Standardinstallation finden Sie die Protokolldateien im Verzeichnis `{AS-Stammverzeichnis}/ae_wlpserver/usr/servers/aeserver/logs`.

Bei der Standardkonfiguration werden zwei Protokolldateien erstellt, in die im täglichen Wechsel geschrieben wird.

as.log Diese Datei enthält die übergeordnete Zusammenfassung von Informationswarnungen und Fehlernachrichten. Bei Serverfehlern, die nicht anhand der in der Benutzerschnittstelle angezeigten Fehlernachricht behoben werden können, sollten Sie diese Datei zuerst prüfen.

as_trace.log
Diese Datei enthält neben allen Einträgen aus `ae.log` weitere Informationen, die zu Debugzwecken primär für den IBM Support und Entwickler vorgesehen sind.

Analytic Server verwendet Apache LOG4J als zugrunde liegende Protokollierungseinrichtung. Mithilfe von LOG4J kann die Protokollierung dynamisch angepasst werden, indem die Konfigurationsdatei `{AS-Serverstammverzeichnis}/configuration/log4j.xml` bearbeitet wird. Möglicherweise werden Sie vom IBM Support gebeten, die Datei zu bearbeiten, um das Diagnostizieren von Problemen zu unterstützen. Sie könnten die Datei auch ändern, um die Anzahl der vorhandenen Protokolldateien zu begrenzen. Änderungen an der Datei werden innerhalb von wenigen Sekunden automatisch erkannt, sodass Analytic Server nicht erneut gestartet werden muss.

Weitere Informationen zu log4j und zur Konfigurationsdatei finden Sie in der Dokumentation auf der offiziellen Apache-Website unter http://logging.apache.org/log4j/.

Versionsinformation

Sie können durch Prüfen des Ordners `{AS-Stammverzeichnis}/properties/version` ermitteln, welche Version von Analytic Server installiert ist. Die folgenden Dateien enthalten die Versionsinformation.

IBM_SPSS_Analytic_Server-*.swtag
Enthält die detaillierte Produktinformation.

version.txt
Enthält Version und Buildnummer für das installierte Produkt.

Log Collector

Wenn Probleme nicht durch direktes Prüfen der Protokolldateien gelöst werden können, besteht die Möglichkeit, alle Protokolle zu bündeln und an den IBM Support zu senden. Es wird ein Dienstprogramm bereitgestellt, um die Erfassung aller erforderlichen Daten zu vereinfachen.

Führen Sie über eine Befehlsshell die folgenden Befehle aus:

```
cd  {AS-Stammverzeichnis}/bin
run  >sh ./logcollector.sh
```

Mit diesen Befehlen wird eine komprimierte Datei unter `{AS-Stammverzeichnis}/bin` erstellt. Die komprimierte Datei enthält alle Protokolldateien und die Informationen zur Produktversion.

Allgemeine Probleme

In diesem Abschnitt werden einige allgemeine Verwaltungsprobleme sowie Wege zu deren Lösung beschrieben.

Ausführen von Datenströmen

R-Jobs setzen nicht englische Wörter in Unicode um

In Cloudera-Clustern setzt R nicht englische Wörter in Unicode um, wenn die Systemcodierung von Hadoop-Server nicht UTF-8 ist.

1. Navigieren Sie in der Cloudera Manager-Konsole zur Registerkarte **YARN configuration**.
2. Fügen Sie im Feld **NodeManager Environment Advanced Configuration Snippet (Safety Valve)** die folgenden Einstellungen hinzu.

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

Die Ausführung von PySpark-Jobs schlägt fehl.

Stellen Sie sicher, dass der Spark-Service in allen Analytic Server-Knoten und in allen Knotenmanagern bereitgestellt wird.

Die Ausführung von PySpark-Jobs in Kerberos-aktivierten Umgebungen schlägt fehl.

Sie müssen den Befehl `kinit` ausführen und anschließend Analytic Server erneut starten, damit PySpark-Tests erfolgreich ausgeführt werden. Beispiel:

HDP-Kerberos

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

CDH-Kerberos

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

Speicherfehler

Konfigurieren von YARN nach Executorspeicherfehler

Der folgende Fehler kann auftreten, wenn der erforderliche Executorspeicher über dem maximalen Schwellenwert liegt:

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

Die folgenden Schritte stellen die YARN-Konfigurationseinstellungen bereit, die erforderlich sind, um das Problem zu beheben.

Für Ambari

1. Rufen Sie in der Ambari-Benutzerschnittstelle **YARN** > **Configs** > **Settings** auf.
2. Erhöhen Sie **node (the memory that is allocated for all YARN containers)** auf `8192 MB`.
3. Erhöhen Sie die Containerwerte:
   - **Minimum Container Size (Memory)** auf `682 MB`
   - **Maximum Container Size (Memory)** auf `8192 MB`
4. Erhöhen Sie **Maximum Container Size (VCores)** auf `3`.
5. Starten Sie YARN, Spark und den Analytic Server-Service erneut.

Für Cloudera

1. Erhöhen Sie `yarn.nodemanager.resource.memory-mb` auf 8 GB.
   - Rufen Sie in der Cloudera Manager-Benutzerschnittstelle **Yarn service** > **Configurations** > **Search Container Memory** auf und erhöhen Sie den Wert auf `8 GB`.

2. Rufen Sie in der Cloudera Manager-Benutzerschnittstelle **YARN service** > **Quick Links** auf und wählen Sie **Dynamic Resource Pools** aus.
3. Klicken Sie unter **Configuration** für jeden verfügbaren Pool auf **edit** und setzen Sie unter **YARN** den Wert für **Max Running Apps** auf 4.
4. Starten Sie YARN, Spark und den Analytic Server-Service erneut.

Hadoop mit Apache Spark 2.x

- Die meisten `forcespark-` und `forcehadoop`-Jobs schlagen fehl, wenn Hadoop und Apache Spark 2.x in derselben Umgebung vorhanden sind. Der Fehler wird wie folgt im Yarn-Anwendungsprotokoll angezeigt: `java.lang.NoClassDefFoundError: org/apache/hadoop/fs/FSDataInputStream`.

  Das Problem kann durch manuelle Bearbeitung der Datei `/etc/spark2/conf/spark-defaults.conf` wie folgt behoben werden:

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

- Wenn zwei JDK-Versionen auf demselben System installiert sind, verwendet Cloudera JDK 1.7, während Spark 2.x JDK 1.8 verwendet. Die Ausführung von `forcespark-` oder `forcehadoop`-Jobs mit Apache Spark 2.x kann zum Fehlschlagen aller Jobs mit der folgenden Fehlernachricht führen:

  ```
  Execution failed. Reason: org/apache/spark/api/java/function/PairFunction : Unsupported major.minor version 52.0
  ```

  Fügen Sie für Cloudera die folgende Zeile im Cloudera Manager-Abschnitt **Analytic Server Advanced Configuration Snippet (Safety Valve) for server.env** hinzu:

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

Erteilen einer `admin`-Berechtigung für Benutzer der benutzerdefinierten Funktion von Apache Hive

Nach der Registrierung der benutzerdefinierten Funktion von Apache Hive in Analytic Server tritt unter Umständen ein Fehler vom Typ `Invalid function` (Ungültige Funktion) auf. Es gibt standardmäßig zwei Hive-Rollen (`admin` und `public`). Hive-Benutzer gehören zur Rolle `public`. Für die benutzerdefinierte Funktion von Hive müssen registrierte Benutzer über eine `admin`-Berechtigung verfügen (Hive-Sicherheit ist aktiviert).

Gehen Sie wie folgt vor, um Benutzern der benutzerdefinierten Funktion von Hive eine `admin`-Berechtigung zu erteilen:

1. Melden Sie sich als Hive bei Beeline an:

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

2. Führen Sie in Beeline den folgenden Befehl aus:

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

   Anmerkung: Zu weiteren nützlichen SQL-Befehlen zählen die folgenden:

   Zum Anzeigen, welche Rollen dem Benutzer `hive` bereits zugewiesen wurden:

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

   Zum Anzeigen, welche Benutzer der Rolle `public` zugewiesen wurden:

   ```
   show principals public;
   ```

3. Starten Sie Hive erneut und registrieren Sie die benutzerdefinierte Funktion von Hive in Analytic Server erneut.

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

HiveDB-Fehler

Beim Schreiben in eine HiveDB tritt unter Umständen der folgende Fehler auf:

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

Der Fehler wird durch mehrere Versionen der Datei `guava-*.jar` im Hadoop-Cluster verursacht. Er kann behoben werden, indem die folgenden Schritte ausgeführt werden (im Beispiel wird HDP 3.1 verwendet):

1. Öffnen Sie die Ambari-Konsole und beenden Sie den Analytic Server-Service.
2. Kopieren Sie `/usr/hdp/3.1.0.0-78/spark2/jars/guava-14.0.1.jar` in `{AS_ROOT}/ae_wlpserver/usr/servers/aeserver/apps/AE_BOOT.war/WEB-INF/lib`.
3. Aktualisieren Sie den Analytic Server-Service in der Ambari-Konsole und starten Sie den Analytic Server-Service anschließend erneut.

Leistungsoptimierung

In diesem Abschnitt werden Möglichkeiten zum Optimieren der Leistung Ihres Systems beschrieben.

Analytic Server ist eine Komponente im Ambari-Framework, die andere Komponenten, wie z. B. HDFS, YARN und Spark, verwendet. Für Analytic Server-Arbeitslasten werden allgemeine Leistungsoptimierungsverfahren für Hadoop, HDFS und Spark angewendet. Jede Analytic Server-Arbeitslast ist anders. Deshalb basieren Optimierungsveruche auf Ihrer jeweiligen Bereitstellungsarbeitslast. Die folgenden Eigenschaften und Tipps zur Optimierung sind Schlüsseländerungen, die sich auf die Ergebnisse von Benchmarking- und Skalierungstests von Analytic Server ausgewirkt haben.

Bei der Ausführung des ersten Jobs in Analytic Server startet der Server eine persistente Spark-Anwendung, die aktiv bleibt, bis Analytic Server beendet wird. Die persistente Spark-Anwendung ordnet Clusterressourcen zu und behält die Zuordnung aller Clusterressourcen bei, solange Analytic Server aktiv ist, sogar wenn ein Analytic Server-Job nicht aktiv ausgeführt wird. Planen Sie sorgfältig, wie viele Ressourcen der Spark-Anwendung von Analytic Server zugeordnet werden sollen. Wenn alle Clusterressourcen der Spark-Anwendung von Analytic Server zugeordnet werden, ist es möglich, dass andere Jobs verzögert oder nicht ausgeführt werden. Diese Jobs könnten in eine Warteschlange gestellt werden und auf genügend freie Ressourcen warten, aber diese Ressourcen würden von der Spark-Anwendung von Analytic Server belegt.

Wenn mehrere Analytic Server-Services konfiguriert und bereitgestellt werden, kann jede Serviceinstanz potenziell ihre eigene persistente Spark-Anwendung zuordnen. Beispiel: Wenn zwei Analytic Server-Services für die Unterstützung von Hochverfügbarkeits-Failover bereitgestellt werden, sehen Sie zwei aktive persistente Spark-Anwendungen, von denen jede Clusterressourcen zuordnet.

Eine zusätzliche Komplexität ist, dass Analytic Server in bestimmten Situationen einen MapReduce-Job startet, der Clusterressourcen erfordert. Diese MapReduce-Jobs erfordern Ressourcen, die nicht der Spark-Anwendung zugeordnet sind. Die spezifischen Komponenten, die MapReduce-Jobs erfordern, sind PSM-Modellerstellungen.

Die folgenden Eigenschaften können so konfiguriert werden, dass sie einer Spark-Anwendung Ressourcen zuordnen. Wenn sie in der Datei `spark-defaults.conf` der Spark-Installation festgelegt sind, werden sie für alle Spark-Jobs zugeordnet, die in der Umgebung ausgeführt werden. Wenn sie in der Analytic Server-Konfiguration als angepasste Eigenschaften unter dem Abschnitt **Custom analytic.cfg** festgelegt sind, werden sie nur der Spark-Anwendung von Analytic Server zugeordnet.

spark.executor.memory

Zu verwendende Speichermenge pro Executorprozess.

spark.executor.instances

Die Anzahl der zu startenden Executorprozesse.

spark.executor.cores

Die Anzahl der Executor-Worker-Threads pro Executorprozess. Dieser Wert sollte zwischen 1 und 5 liegen.

Beispiel zum Festlegen der drei Spark-Schlüsseleigenschaften. In einem HDFS-Cluster gibt es 10 Datenknoten und jeder Datenknoten hat 24 logische Kerne und 48 GB Speicher und führt nur HDFS-Prozesse aus. Nachfolgend wird eine Möglichkeit zum Konfigurieren der Eigenschaften für diese Umgebung unter der Annahme beschrieben, dass Sie nur Analytic Server-Jobs für diese Umgebung ausführen und die maximale Zuordnung für eine einzelne Spark-Anwendung von Analytic Server wünschen.

- Legen Sie spark.executor.instances=20 fest. Dadurch wird versucht, zwei Spark-Executorprozesse pro Datenknoten auszuführen.
- Legen Sie spark.executor.memory=22G fest. Dadurch wird die maximale Größe des Heapspeichers für jeden Spark-Executorprozess auf 22 GB gesetzt, wodurch auf jedem Datenknoten 44 GB zugeordnet werden. Andere JVMs und das Betriebssystem benötigen den zusätzlichen Speicher.
- Legen Sie spark.executor.cores=5 fest. Dadurch werden 5 Worker-Threads für jeden Spark-Executor bereitgestellt, insgesamt 10 Worker-Threads pro Datenknoten.

Überwachung der Spark-Benutzerschnittstelle auf aktive Jobs

Ein Überlauf auf die Festplatte kann sich auf die Leistung auswirken. Es folgen einige Beispiele für mögliche Lösungen:

- Vergrößern Sie den Speicher und ordnen Sie ihn Spark-Executorn über **spark.executor.memory** zu.
- Verringern Sie die Anzahl für **spark.executor.cores**. Dadurch wird die Anzahl gleichzeitig ablaufender Arbeitsthreads, die Speicher reservieren, aber auch der Grad der Parallelität für die Jobs verringert.
- Ändern Sie die Spark-Speichereigenschaften. **spark.shuffle.memoryFraction** und **spark.storage.memoryFraction** steuern den Prozentsatz des zugeordneten Spark-Executor-Heapspeichers für Spark.

Sicherstellen, dass der Namensknoten genügend Speicher hat

Wenn die Anzahl der Blöcke in HDFS groß ist und zunimmt, stellen Sie sicher, dass Sie den Heapspeicher des Namensknotens vergrößern, damit diese Zunahme verarbeitet werden kann. Dies ist eine übliche Optimierungsempfehlung für HDFS.

Ändern der Speichermenge für Caching

Standardmäßig hat **spark.storage.memoryFraction** den Wert 0.6. Dies kann auf bis zu 0.8 vergrößert werden, falls die HDFS-Blockgröße 64 MB beträgt. Ist die HDFS-Blockgröße der Eingabedaten größer als 64 MB, kann dieser Wert nur vergrößert werden, wenn der pro Aufgabe zugeordnete Speicher größer als 2 GB ist.

Optimierung der Leistung von Modellscoring

Sie können die Leistung von Modellscoring-Jobs für große Datasets mit der Apache Spark-Engine mithilfe der folgenden Schritte verbessern. Beachten Sie, dass diese Schritte sich nicht auf den Betrieb von anderen Services als Analytic Server-Services im Cluster auswirken sollten.

1. Prüfen Sie, ob libtcmalloc_minimal.so{/Version} bereits auf jedem Knoten im Cluster installiert ist.

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

2. Wenn libtcmalloc_minimal.so nicht installiert ist, installieren Sie entweder das betriebssystemspezifische Paket mit der Bibliothek libtcmalloc_minimal auf jedem Knoten in Ihrem Cluster oder erstellen und installieren Sie libtcmalloc_minimal manuell. Beispiel:

   Ubuntu:

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

Red Hat Enterprise Linux 6.x (x64):

a. Installieren Sie das EPEL-Repository für RedHat (falls es nicht bereits installiert ist).

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

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

Manuelle Erstellung:

a. Laden Sie gperftools-2.4.tar.gz über den folgenden Link herunter: https://github.com/gperftools/gperftools/releases

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

c. `cd gperftools-2.4`

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

e. `make`

f. `sudo make install`

3. Notieren Sie einen der Speicherorte der installierten Bibliotheksdatei libtcmalloc_minimal.so{.Version}, der nach der Ausführung des folgenden Befehls auf mindestens einem der Knoten zurückgegeben wird.

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

   Wenn der Cluster Knoten enthält, auf denen verschiedene Betriebssysteme ausgeführt werden, kann sich diese Datei an mehreren Speicherorten befinden.

4. Rufen Sie in der Ambari-Konsole die Analytic Server-Konfiguration auf und konfigurieren Sie unter dem Abschnitt **Custom analytics.cfg** den Schlüssel spark.executorEnv.LD_PRELOAD mit dem Speicherort der Bibliothek als Wert. Starten Sie den Analytic Server-Service nach dieser Änderung erneut. Beispiel: Wenn die Bibliothek im Verzeichnis /usr/lib64/libtcmalloc_minimal.so.4 installiert ist, sieht die Konfiguration wie folgt aus:

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

   Wenn mehrere Speicherorte erforderlich sind, trennen Sie sie wie im folgenden Beispiel durch Leerzeichen voneinander.

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

   Wenn auf einem der Knoten die Bibliothek `libtcmalloc_minimal.so` nicht an einem der konfigurierten Speicherorte installiert ist, wird zwar kein Fehler verursacht, aber die Leistung beim Modellscoring kann auf diesen Knoten langsamer sein.

Mapseitiger Spark-Join

Die Spark-Join-Implementierung von Analytic Server unterstützt die mapseitige Joinfunktionalität nicht (der Spark-Join ist hauptsächlich eine Reduce-Seite). Die Implementierung nutzt nicht die mapseitigen Joins, um Joins zu optimieren, wenn eine Eingabe klein ist. Wenn Sie den mapseitigen Join nicht nutzen, führt dies zu einem extrem ressourcenintensiven Spark-Job, der letztendlich fehlschlägt.

Sie können Joins bei der Ausführung von mapseitigen Analytic Server-Spark-Joins (oder nativen Spark-Jobs, die auf der kleinsten RDD-Größe basieren) optimieren, indem Sie der Datei `analytics.cfg` (`SPSS Analytic Server/Configs/Custom analytics.cfg`) oder `analytics-meta` die Eigenschaft `spark.msj.maxBroadcast` hinzufügen.

Bemerkungen

Die vorliegenden Informationen wurden für Produkte und Services entwickelt, die auf dem deutschen Markt angeboten werden. IBM stellt dieses Material möglicherweise auch in anderen Sprachen zur Verfügung. Für den Zugriff auf das Material in einer anderen Sprache kann eine Kopie des Produkts oder der Produktversion in der jeweiligen Sprache erforderlich sein.

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

Diese Informationen können technische Ungenauigkeiten oder typografische Fehler enthalten. 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 Director of Licensing
IBM Europe, Middle East & Africa
Tour Descartes
2, avenue Gambetta
92066 Paris La Defense
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.

Die angeführten Leistungsdaten und Kundenbeispiele dienen nur zur Illustration. Die tatsächlichen Ergebnisse beim Leistungsverhalten sind abhängig von der jeweiligen Konfiguration und den Betriebsbedingungen.

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.

COPYRIGHTLIZENZ:

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:

Marken

IBM, das IBM Logo und ibm.com sind Marken oder eingetragene Marken der IBM Corporation 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 (The Minister for the Cabinet Office) und eine eingetragene Marke, die beim U.S. Patent and Trademark Office eingetragen ist.

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

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, der IBM Corporation und von Quantum in den USA und/oder anderen Ländern.

Gedruckt in Deutschland