Open for Innovation
KNIME

KNIME Google Cloud Integration

Benutzerhandbuch

KNIME AG, Zürich, Schweiz

Version 5.7 (letzte Aktualisierung auf)

Inhaltsverzeichnis

Überblick

KNIME Analytics Platform enthält eine Reihe von Knoten, um mehrere Google Cloud-Dienste zu unterstützen.

Die unterstützten Google Cloud-Dienste, die in diesem Leitfaden erfasst werden, sind Google Dataproc , Google Cloud-Speicher , und Google BigQuery .

KNIME Analytics Platform bietet weitere Integration Google Drive .

Google Cloud-Speicher

KNIME Google Cloud Storage Verbindung Erweiterung bietet Knoten zur Verbindung mit Google Cloud Storage.

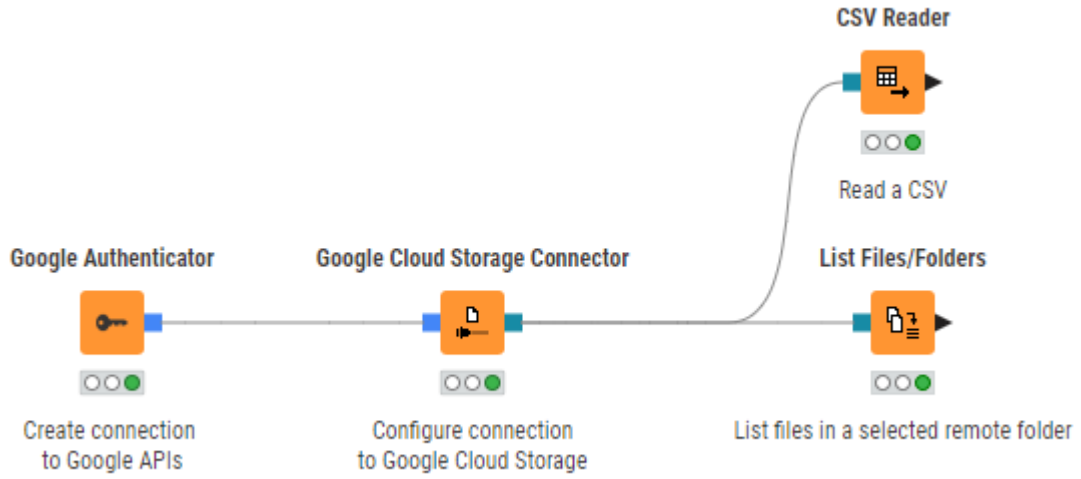

Abbildung 1. Verbinden mit und arbeiten mit Google Cloud Storage

Abbildung 1 zeigt ein Beispiel für die Verbindung zu Google Cloud Storage und die Zusammenarbeit mit dem Remote-Dateien.

Google Authentication

Die Google Authentication node ermöglicht die Authentifizierung mit den verschiedenen Google APIs Verwendung einer API-Schlüsseldatei. Um diesen Knoten nutzen zu können, müssen Sie ein Projekt im Google Cloud-Konsole . Weitere Informationen zur Erstellung eines Projekts auf der Google Cloud Console, Bitte folgen Sie Google-Dokumentation .

Dann müssen Sie ein Service-Konto und einen API-Schlüssel erstellen. Sie können entweder auswählen JSON oder P12 als API-Schlüsselformat (siehe [fig_select_p12]) Die Service-Account-E-Mail hat das Format sa-name@project-id.iam.gserviceaccount.com wenn Name eine eindeutige Kennung ist und

Projektid ist die ID des Projekts.

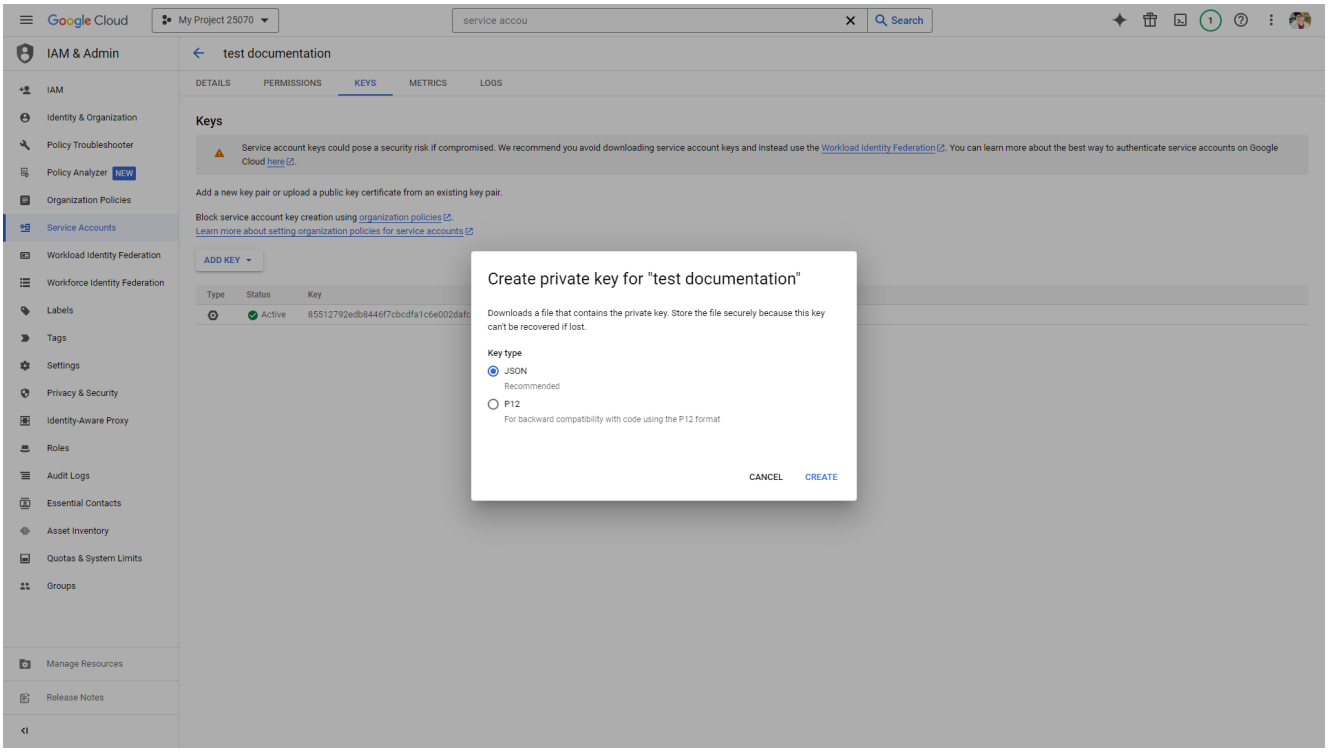

Abbildung 2. Wählen Sie die JSON- oder P12-Datei als Dienstkontoschlüssel

Die API-Schlüsseldatei wird automatisch auf Ihre lokale Maschine heruntergeladen. Beachten Sie, dass Sie sollte die Datei an einem sicheren Ort auf Ihrem lokalen System speichern.

Öffnen Sie den Google Authenticator-Knotenkonfigurationsdialog:

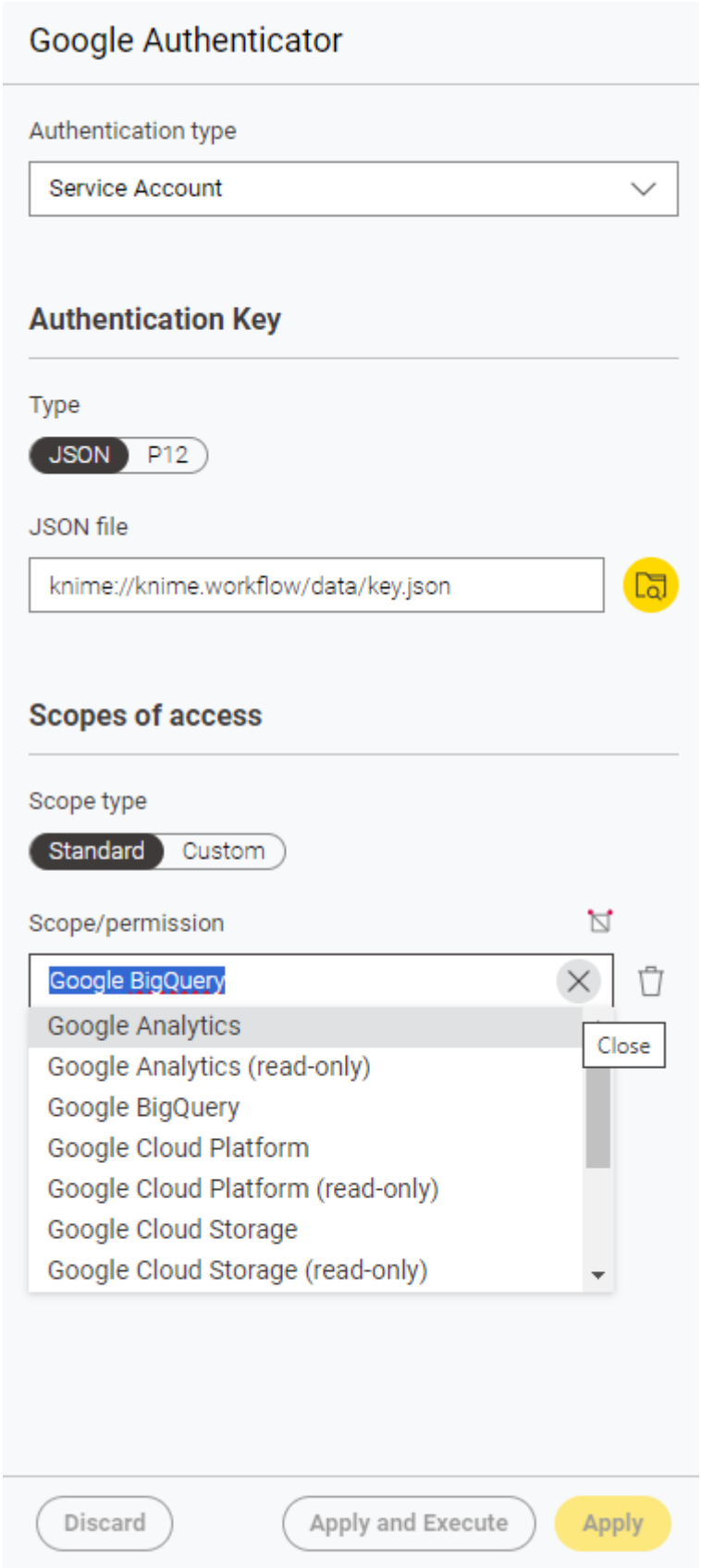

Abbildung 3. Node Konfiguration Dialog von Google Authenticator Knoten

Unter Authentication Typ Wählen Service Account . Nun, innerhalb des Knoten-Dialogs, müssen Sie die folgenden Einstellungen konfigurieren:

• Wählen Sie den Schlüsseltyp Authentication aus. Sie können entweder einen JSON oder einen P12 API-Schlüsseltyp auswählen.

- Wenn Sie P12 als API-Schlüsseltyp fügen Sie Ihre Servicekonto E-Mail. Wenn Sie nicht schon einen haben, bitte folgen Sie der Google-Dokumentation wie man erstellt ein Servicekonto.

• Klicken Sie im Node Dialog auf die Schaltfläche durchsuchen und wählen Sie die Schlüsseldatei.

• Fügen Sie die OAuth 2.0-Bereiche die für diese Verbindung gewährt wird. Sie sollten wählen die Reichweiten abhängig von der Ebene des Zugriffs, die Sie benötigen. So können Sie wählen Standard , click Anwendungsbereich und wählen Sie einen Bereich unter den verfügbaren im Menü.

> Um das entsprechende zu sehen Anwendungsbereich unter der Standard Liste der Bereiche, die Sie muss zuerst die Erweiterung installiert haben. Zum Beispiel KNIME Google Cloud-Speicheranschluss oder KNIME Große Abfrage Erweiterung.

- Alternativ wählen Sie Zoll und die gewünschte hinzufügen Anwendungsbereich auf das Feld. Sie können mehrere benutzerdefinierte Bereiche hinzufügen, indem Sie klicken Anwendungsbereich und Sie können löschen die Bereiche, die Sie durch Klicken auf die Icon.

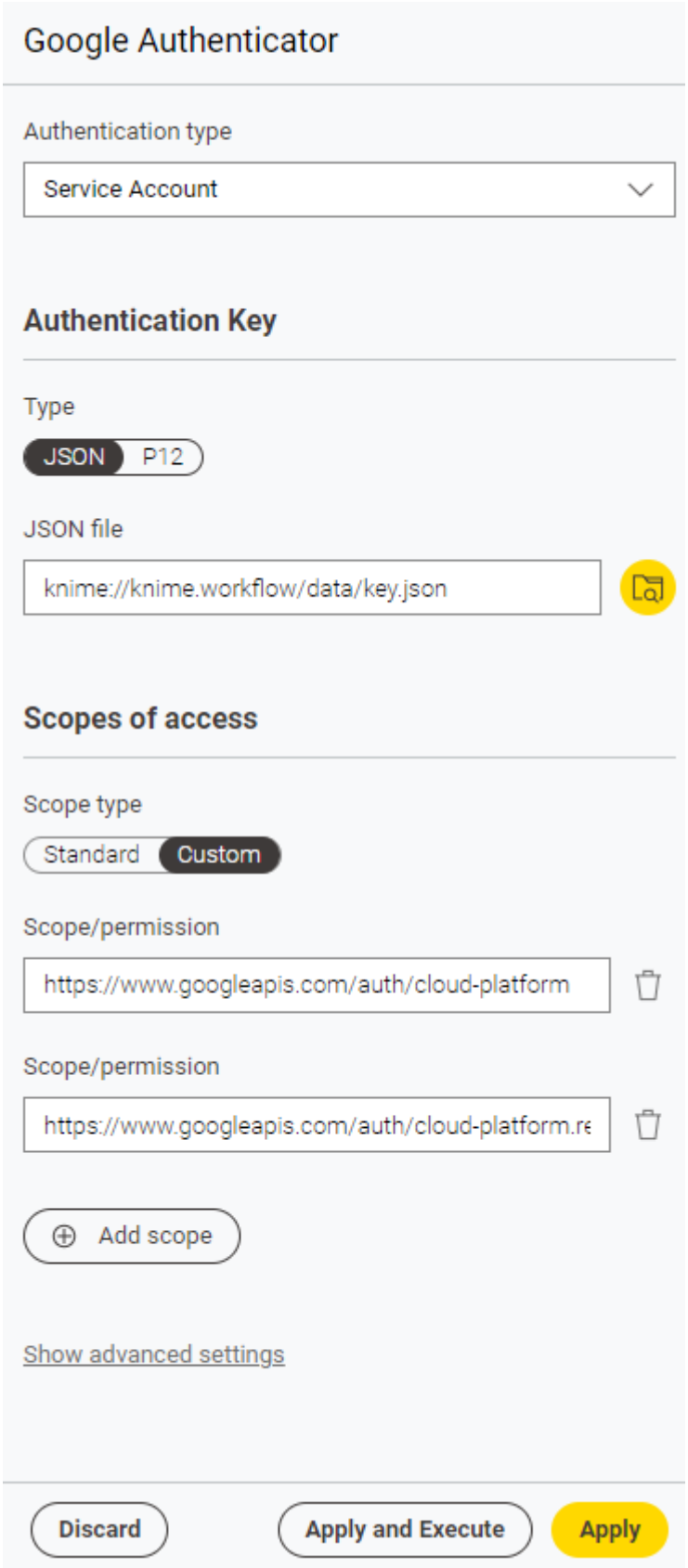

Abbildung 4. Node Konfigurationsdialog von Google Authenticator node - benutzerdefinierte Scopes

Anstelle des Google Authenticator-Knotens verwenden Sie auch den Secrets Retriever das Google-Geheimnis von der Gefällt mir . Für weitere Details wie Sie ein Google-Geheimnis im Secret Store einrichten, lesen Sie bitte die KNIME Secrets Benutzerhandbuch .

Google Cloud Storage Connector

Die Google Cloud Storage Connector Knoten verbindet sich mit Google Cloud Storage und ermöglicht nachgeschaltete Knoten, um auf Google Cloud Storage innerhalb eines bestimmten Projekts mit dem neuen KNIME-Dateihandling-Knoten.

Der Knotenkonfigurationsdialog des Google Cloud Storage Connector-Knotens enthält:

- Projekt-ID. Dies ist die Google Cloud-Projekt-ID. Für weitere Informationen zum Finden Ihr Projekt ID, bitte check out the Google-Dokumentation .
- Arbeitsverzeichnis. Das Arbeitsverzeichnis muss als absoluter Pfad angegeben werden und es ermöglicht nachgeschalteten Knoten den Zugriff auf Dateien/Ordner mittels relativer Pfade, d.h. Pfade, die nicht einen führenden Slash. Wenn nicht angegeben, ist das Standard-Arbeitsverzeichnis / .

  Path syntax: Pfade für Google Cloud Storage werden mit einer UNIX-ähnlichen Syntax, z. /mybucket/myfolder/myfile . Der Pfad besteht üblicherweise aus:

  - Ein führender Slash (/)
  - Nach dem Namen eines Eimers (Mybucket im obigen Beispiel, gefolgt von Slash
  - Gefolgt durch den Namen eines Objekts im Eimer (myfolder/myfile in der Beispiel).
- Wege normalisieren. Die Pathnormalisierung eliminiert redundante Komponenten eines Pfades, z. /a/../b/c kann normalisiert werden /b/c . Wenn diese redundanten Komponenten mögen ./ oder . sind Teil eines vorhandenen Objekts, dann muss die Normalisierung deaktiviert werden, um Zugang zu ihnen richtig.
- Unter der Erweiterte Tab, es ist möglich, die Verbindung einzustellen und Zeitauslesen.

Dieser Knoten unterstützt derzeit nur den Google Authenticator Knoten für Authentifizierung.

Google BigQuery

KNIME Analytics Platform enthält eine Reihe von Knoten zur Unterstützung Google BigQuery . Die KNIME Großes AngebotDie Erweiterung erfolgt über die KNIME Analytics Platform Version 4.1.

Die Einrichtung der KNIME Analytics Platform für Google BigQuery hat folgende Voraussetzungen:

ANHANG Erstellen Sie ein Projekt in der Google Cloud Console. Weitere Informationen zum Erstellen einer Projekt auf Google Cloud Console, bitte folgen Sie der Google-Dokumentation .

2. Erstellen Sie ein Service-Konto. Wenn Sie noch keinen haben, folgen Sie bitte dem Google Dokumentation wie man ein Service-Konto erstellt.

3. Laden Sie die JDBC Treiber für Google BigQuery , unzip, und speichern Sie es in Ihrem lokalen Maschine. Registrieren Sie den JDBC-Treiber auf der KNIME Analytics Platform, indem Sie Tutorial in der KNIME Dokumentation .

Verbinden mit BigQuery

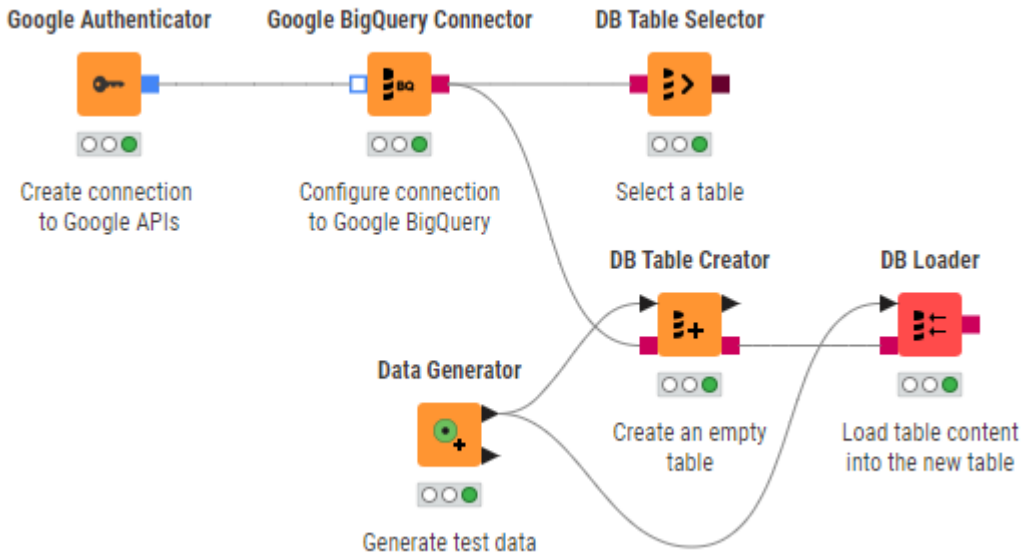

Abbildung 5. Verbinden mit und arbeiten mit Google BigQuery

Abbildung 5 zeigt, wie man mit der Google Authentication Knoten und der Google BigQuery Connector Knoten, um eine Verbindung zu BigQuery über JDBC-Treiber herzustellen. Konfigurieren Google Authenticator Node, bitte auf die <a href="#page2" style="color: #ff6600; text-decoration: Abschnitt.

Der Google BigQuery Connector-Knoten unterstützt nur die Authentifizierung über einen Service Konto oder Ihr eigener OAuth-Client (Zum Detail siehe die Google-Dokumentation)

Um den Google BigQuery Connector-Knoten zu konfigurieren, überprüfen Sie bitte, wie Sie eine Verbindung zu einem

vordefinierte Datenbank in der KNIME Dokumentation . Für den Hostname in BigQuery können Sie Angabe http://www.googleapis.com/bigquery/v2 oder Bigquery.cloud.com . Als Verwenden Sie die Projekt-ID, die Sie auf der Google Cloud Console erstellt haben.

Für weitere Informationen über die JDBC Parameter Tab oder Erweiterte Tab in der Knotenkonfiguration Dialog von Google BigQuery Connector Knoten, bitte überprüfen ! KNIME Dokumentation .

Durch die Ausführung dieses Knotens wird eine Verbindung zur BigQuery-Datenbank erstellt und Sie können jede KNIME Datenbankknoten Ihre SQL-Anweisungen visuell zusammenstellen.

Weitere Informationen zu KNIME-Datenbankknoten finden Sie in der KNIME Datenbankdokumentation .

Erstellen Sie eine BigQuery Tabelle

Um Daten von der KNIME Analytics Platform auf Google BigQuery zu exportieren (in <a href="#page8" style="c :

ANHANG Erstellen Sie das Datenbankschema/Datensatz, wo Sie die Tabelle speichern möchten, wenn es nicht existiert schon. Um einen Datensatz zu erstellen, überprüfen Sie bitte die Google-Dokumentation .

2. Erstellen Sie eine leere Tabelle mit der richtigen Spezifikation. Um dies zu tun, verwenden Sie die DB Table Creator Knoten. Geben Sie im Dialogfeld Knotenkonfiguration das Schema als Namen des Datensatz, den Sie im vorherigen Schritt erstellt haben. Weitere Informationen zur DB-Tabelle Creator node, bitte check the KNIME Dokumentation .

Hat die Tabelle Spaltennamen, die Leerzeichen enthalten, z. Spalte 1 , stellen Sie sicher, die Raumzeichen zu löschen, weil sie automatisch ersetzt durch Ja. bei der Tischschöpfung, z. Spalte 1 und das wird zu Konflikten führen, da Spaltennamen nicht mehr übereinstimmen.

3. Sobald die leere Tabelle erstellt ist, verwenden Sie die DB Loader Knoten, um den Tabelleninhalt in die neu erstellte Tabelle. Weitere Informationen zum DB Loader-Knoten finden Sie unter KNIME Dokumentation .

Google Dataproc

Cluster Setup mit Livy

Um einen Dataproc-Cluster mit der Google Cloud Platform Web-Konsole zu erstellen, folgen Sie dem Schritt-
von der Google-Dokumentation .

Zur EinrichtungApokalypse im Cluster sind folgende zusätzliche Schritte erforderlich:

ANHANG Kopieren der Datei livy. ! von Git Repository in Ihren Cloud-Speicher Eimer. Diese Datei wird als Initialisierungsaktion verwendet, um Livy auf einem Stammknoten innerhalb eines Dataproc zu installieren Cluster.

Bitte überprüfen Best Practices der Verwendung von Initialisierungsaktionen.

2. Während der Cluster-Kreation öffnen Sie die Erweiterte Optionen am Ende der Seite

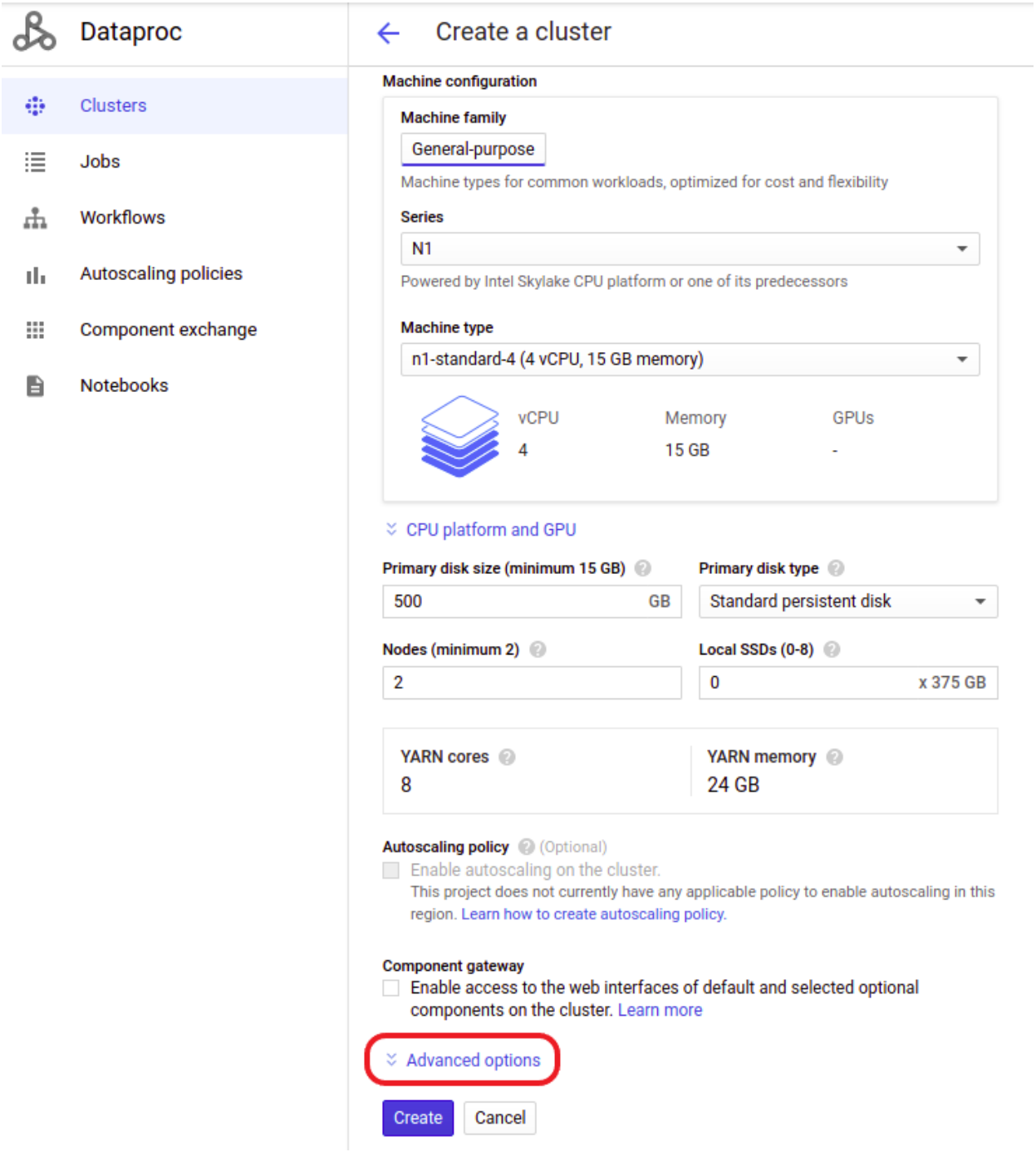

Abbildung 6. Erweiterte Optionen in der Cluster-Erstellungsseite

<a href="#page13" styl

3. Wählen Sie das Netzwerk und das Subnetz aus. Denken Sie an das Netzwerk und Subnetz für die Libyen Abschnitt.

Abbildung 7. Netzwerk und Subnetz

L 347 vom 20.12.2013, S. 1). Wählen Sie die livy. ! Datei aus Ihrem Cloud-Speicher Eimer in der Initialisierungsaktionen Abschnitt

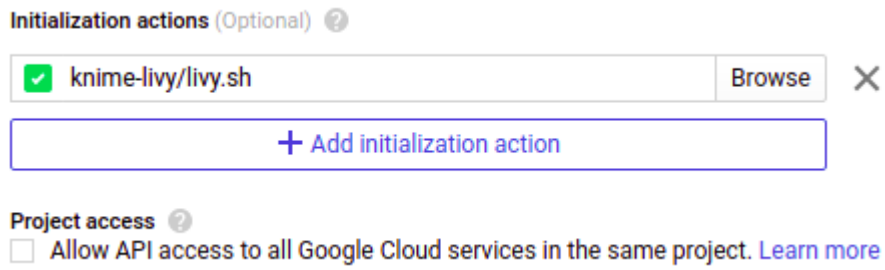

Abbildung 8. Set livy. ! als Initialisierungsaktion

5. Konfigurieren Sie den Rest der Clustereinstellungen nach Ihren Bedürfnissen und erstellen Sie die Cluster.

> Apache Livy ist ein Service, der mit einem Spark-Cluster über einem REST interagiert Schnittstelle. Es ist der empfohlene Dienst, einen Spark Kontext in KNIME zu erstellen Analyseplattform.

Zugang zu Livy

Um die externe IP-Adresse des Stammknotens zu finden, wo Livy läuft:

ANHANG Klicken Sie auf den Clusternamen in der Clusterliste Seite

2. Gehen Sie. VM-Gerichte und klicken Sie auf den Hauptknoten

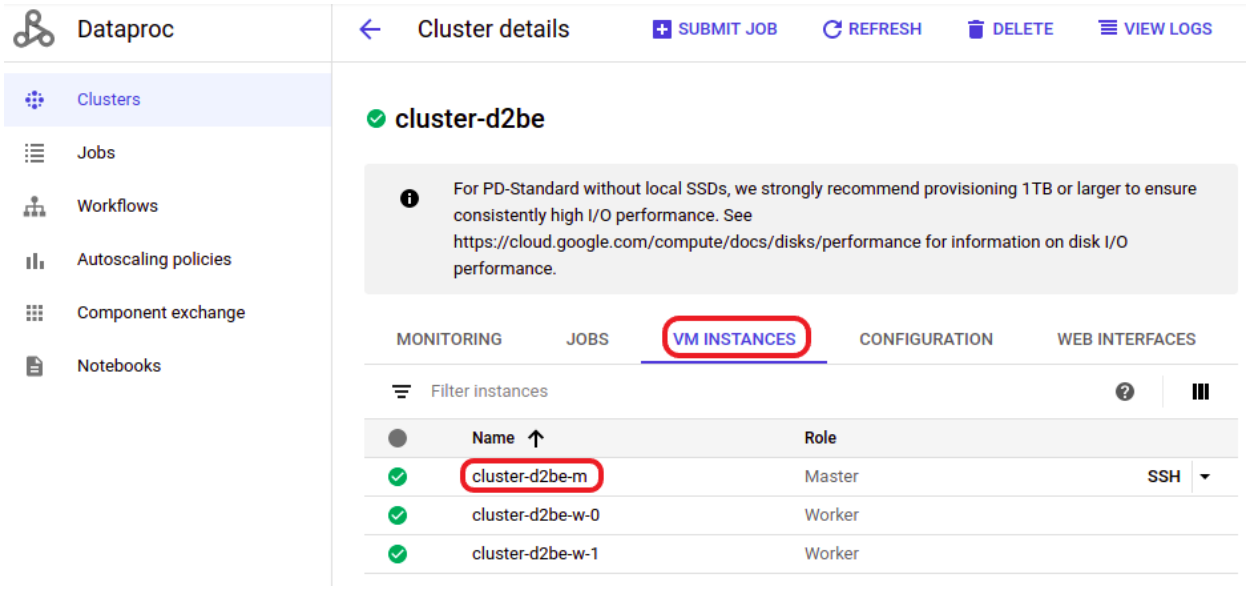

Abbildung 9. Wählen Sie den Stammknoten in der VM-Instanzliste

3. Auf der VM-Gerichte Seite, scrollen bis zum Netzwerkschnittstellen Abschnitt. Finden Sie die Netzwerk und Subnetz, das Sie im vorherigen Abschnitt,

<a href="#page10" style="color: #ff6600; tex

und Sie finden die externe IP-Adresse des Stammknotens.

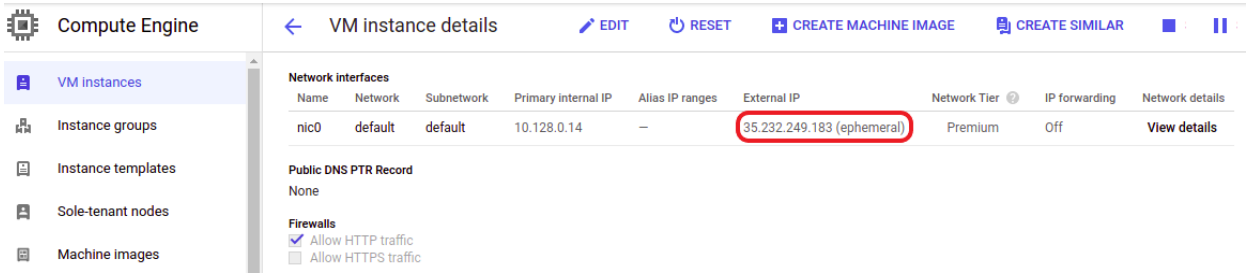

Abbildung 10. Finden Sie die externe IP-Adresse des Stammknotens

Livy Firewall Setup

Um den Zugriff auf Livy von außen zu ermöglichen, müssen Sie die Firewall konfigurieren:

ANHANG Klicken Sie auf den Clusternamen in der Clusterliste Seite

2. Gehen Sie. VM-Gerichte und klicken Sie auf den Hauptknoten

3. Auf der VM-Gerichte Seite, scrollen bis zum Firewalls Abschnitt und stellen Sie sicher, dass Versenden von HTTP-Verkehr zulassen wird aktiviert

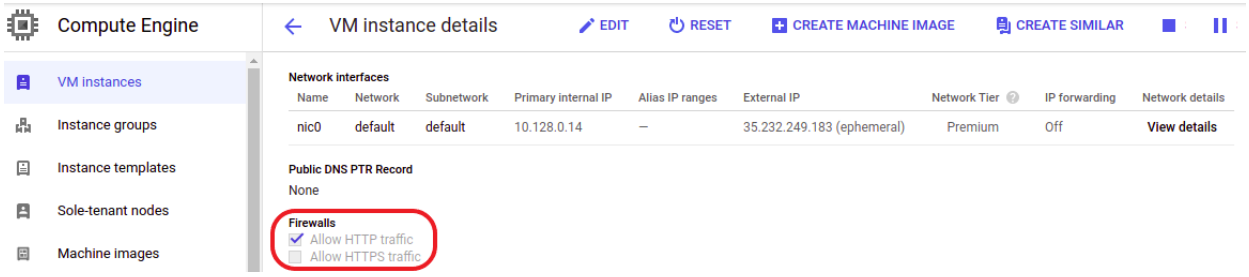

Abbildung 11. Überprüfen Erlauben Sie HTTP-Verkehr im Bereich Firewalls

L 347 vom 20.12.2013, S. 1). Weiter geht's zum VPC-Netz Seite

5. In Firewall Abschnitt der VPC-Netz Seite, wählen Sie die Standard-allow-http Regel

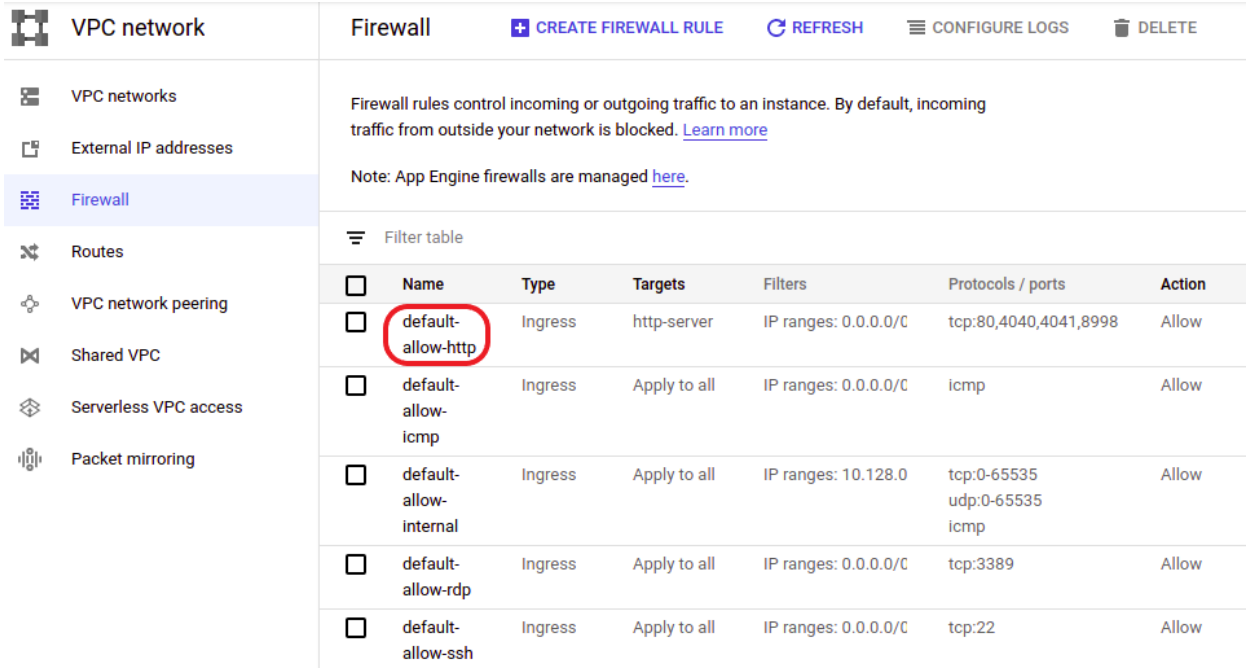

Abbildung 12. Öffnen Sie die Standard-allow-http Firewall-Regel

6. Stellen Sie sicher, dass `tcp:8998` ist in der zulässigen Protokoll- und Portliste enthalten, und dass Sie Die IP-Adresse ist in der erlaubten IP-Adressenliste enthalten.

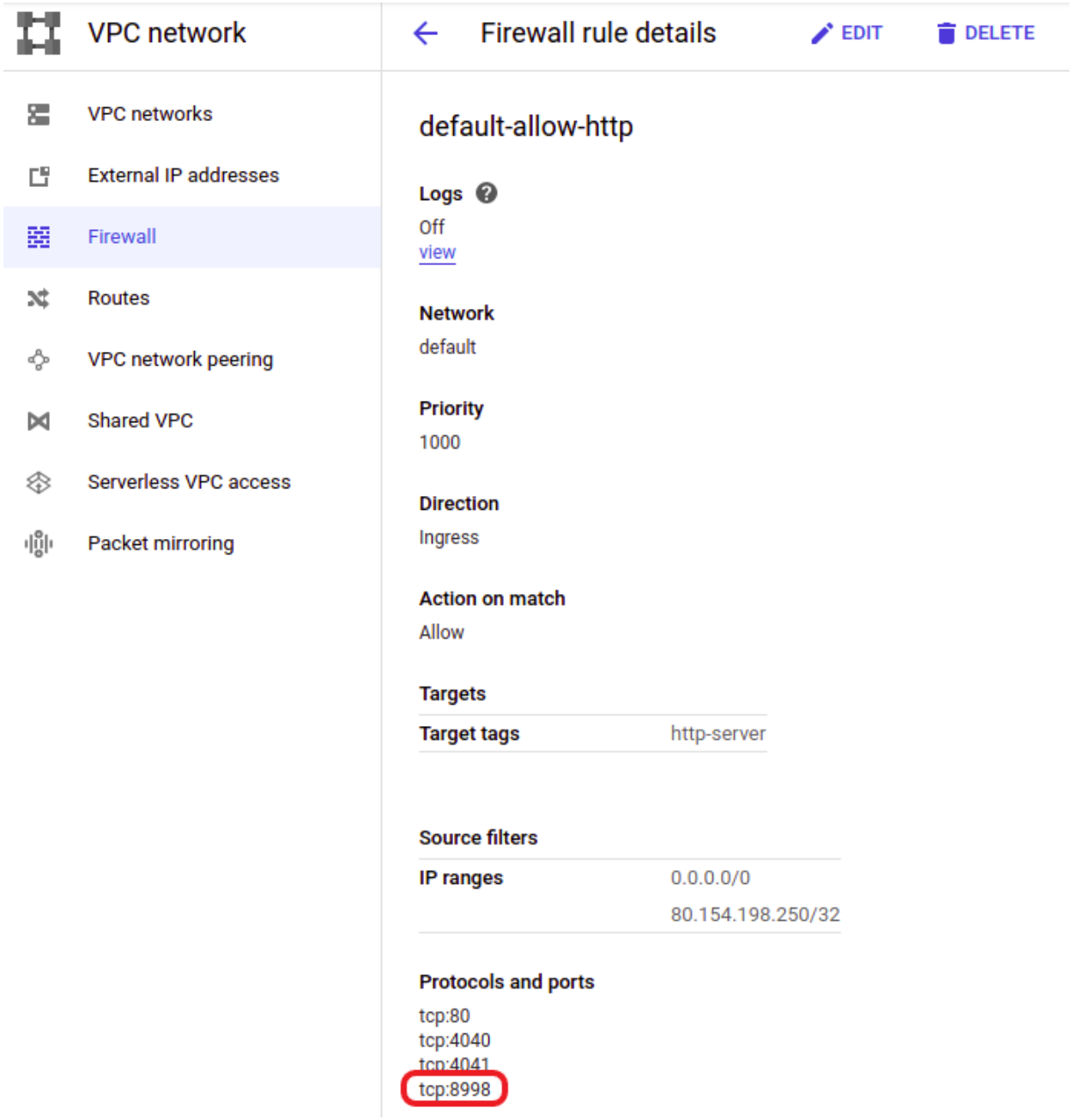

Abbildung 13. Stellen Sie sicher, dass der Zugriff auf bestimmte Ports und IP-Adressen möglich ist

Sobald Sie diese Schritte verfolgt haben, können Sie über den Dataproc-Cluster zugreifen KNIME Analytics Platform mit Apache Livy.

Verbinden Sie mit Dataproc Cluster

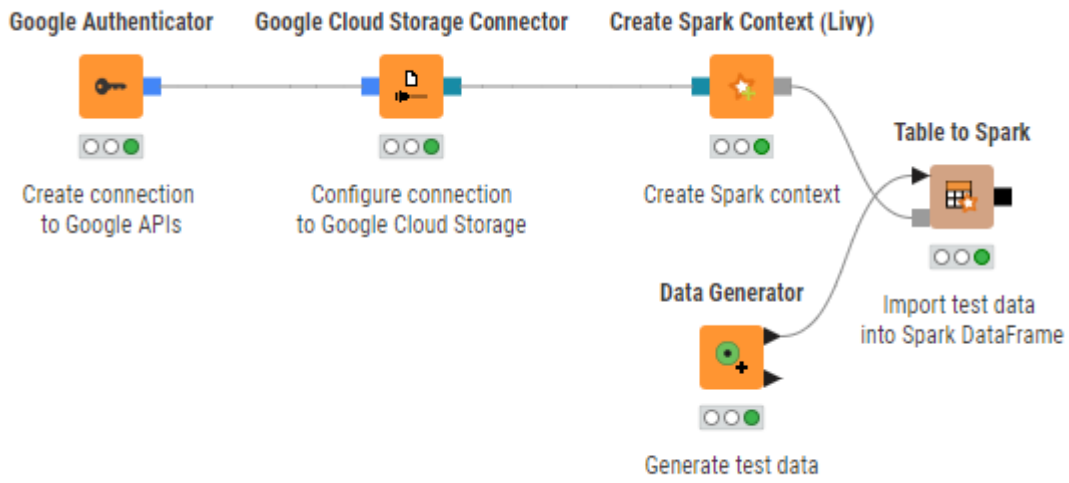

Abbildung 14. Verbindung zum Dataproc-Cluster

Abbildung 14 zeigt, wie eine Verbindung zu einem laufenden Dataproc-Cluster über KNIME aufgebaut wird Analyseplattform. Die Google Authentication Knoten und Google Cloud Storage Connector Knoten werden verwendet, um eine Verbindung zu Google APIs und zu Google Cloud Storage zu erstellen jeweils. Weitere Informationen zu beiden Knoten finden Sie in der Lagerung Abschnitt dieser Führung.

Die Spark Context (Livy) erstellen node erstellt einen Spark-Kontext Apokalypse . Im Inneren der Knotenkonfiguration Dialog, die wichtigsten Einstellungen sind:

- Die Livy URL. Es hat das Format http://:8998 wenn ist die externe IP-Adresse des Stammknotens des Dataproc-Clusters. Um die externe IP zu finden Adresse Ihres Dataproc-Clusters, check out the <a href="#page13" style="color: #ff6600; text-dec Abschnitt.
- Unter Erweiterte Tab, es ist obligatorisch, die Inszenierungsbereich für Spark Jobs . Die Inszenierung Bereich, der sich im angeschlossenen Google Cloud-Speichersystem befindet, wird verwendet, um temporäre Dateien zwischen KNIME und dem Spark-Kontext austauschen.

Die restlichen Einstellungen können nach Ihren Bedürfnissen konfiguriert werden. Für weitere Informationen über die Erstellen Sie Spark Context (Livy) Node, bitte überprüfen Sie unsere Amazon Web Services Dokumentation.

Sobald der Spark-Kontext erstellt ist, können Sie eine beliebige Anzahl der KNIME Spark-Knoten von die KNIME Erweiterung für Apache Spark Ihre visuelle Montage Funkanalysestrom auf dem Cluster ausgeführt.

Apache Hive in Google Dataproc

Dieser Abschnitt beschreibt, wie man eine Verbindung zu Apache HiveTM auf Dataproc in KNIME herstellen kann
Analyseplattform.

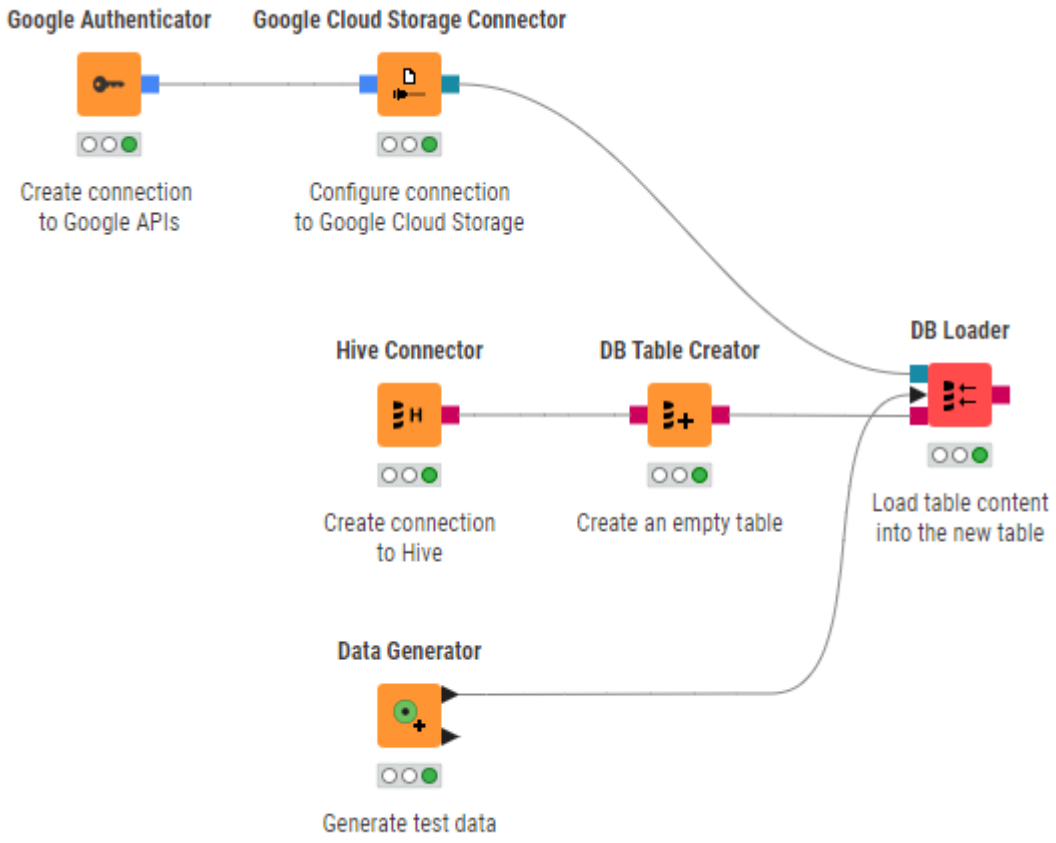

Abbildung 15. Verbinden Sie mit Hive und erstellen Sie eine Hive-Tabelle

Abbildung 15
zeigt, wie sich Hive auf einem Dataproc-Cluster anschließt und wie ein
Hive Tisch.

Die Hive Connector node wird standardmäßig mit dem open-source Apache Hive JDBC gebündelt
Fahrer. Auch proprietäre Treiber werden unterstützt, müssen aber zuerst registriert werden. Folgen Sie der Führung
über die Registrierung eines Hive JDBC Treibers in KNIME Dokumentation .

Sobald der Hive JDBC-Treiber registriert ist, können Sie den Hive Connector-Knoten konfigurieren. Für
mehr Informationen zur Konfiguration der Einstellungen im Node-Konfigurationsdialog, bitte
auf die KNIME Dokumentation . Durch Ausführen des Knotens wird eine Verbindung zu Apache erstellt
Hive und Sie können jede KNIME Datenbankknoten Ihre SQL-Anweisungen visuell zusammenstellen.

Um den Zugang zu Hive von der KNIME Analytics Platform zu ermöglichen, stellen Sie sicher, dass

Hive Port (10000 standardmäßig) wird in den Firewall-Regeln geöffnet. Um dies zu konfigurieren,

Schauen Sie sich

Abschnitt und Änderung der Firewall-Regel

entsprechend.

KNIME AG
Talacker 50
8001 Zürich, Schweiz
www.knime.com
Info@knime.com