

KNIME Google Cloud Integration

Benutzerhandbuch

KNIME AG, Zürich, Schweiz

Version 5.7 (letzte Aktualisierung auf)



Inhaltsverzeichnis

Überblick
Google Cloud Storage
 Google Authenticator
 Google Cloud Storage
Google BigQuery. . . .
 Verbinden Sie mit BigQuery
 Erstellen Sie eine BigQuery Tabl
Google Dataproc. . .
 Cluster Setup mit Lay
 Verbinden Sie mit Dataproc
 Apache Hive in Google Data

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.

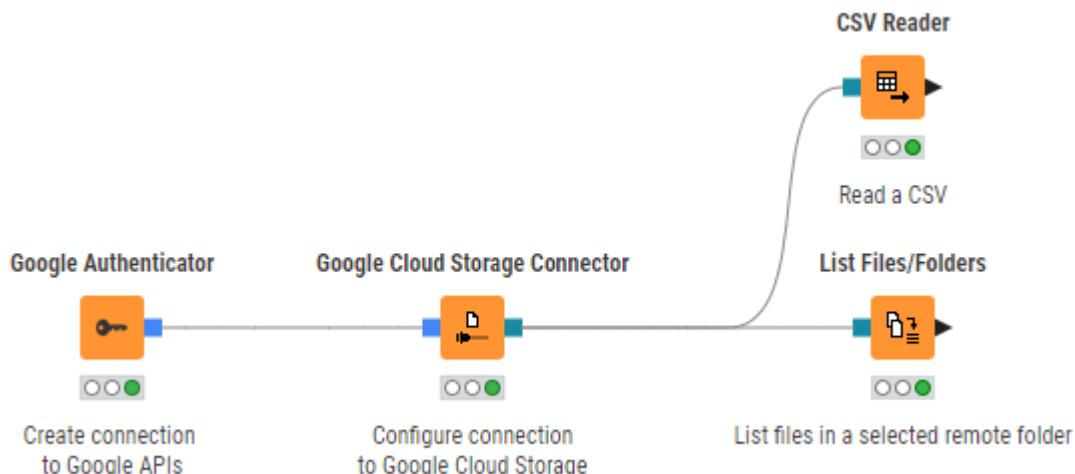


Abbildung 1. Verbinden mit und arbeiten mit Google Cloud Storage

>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.

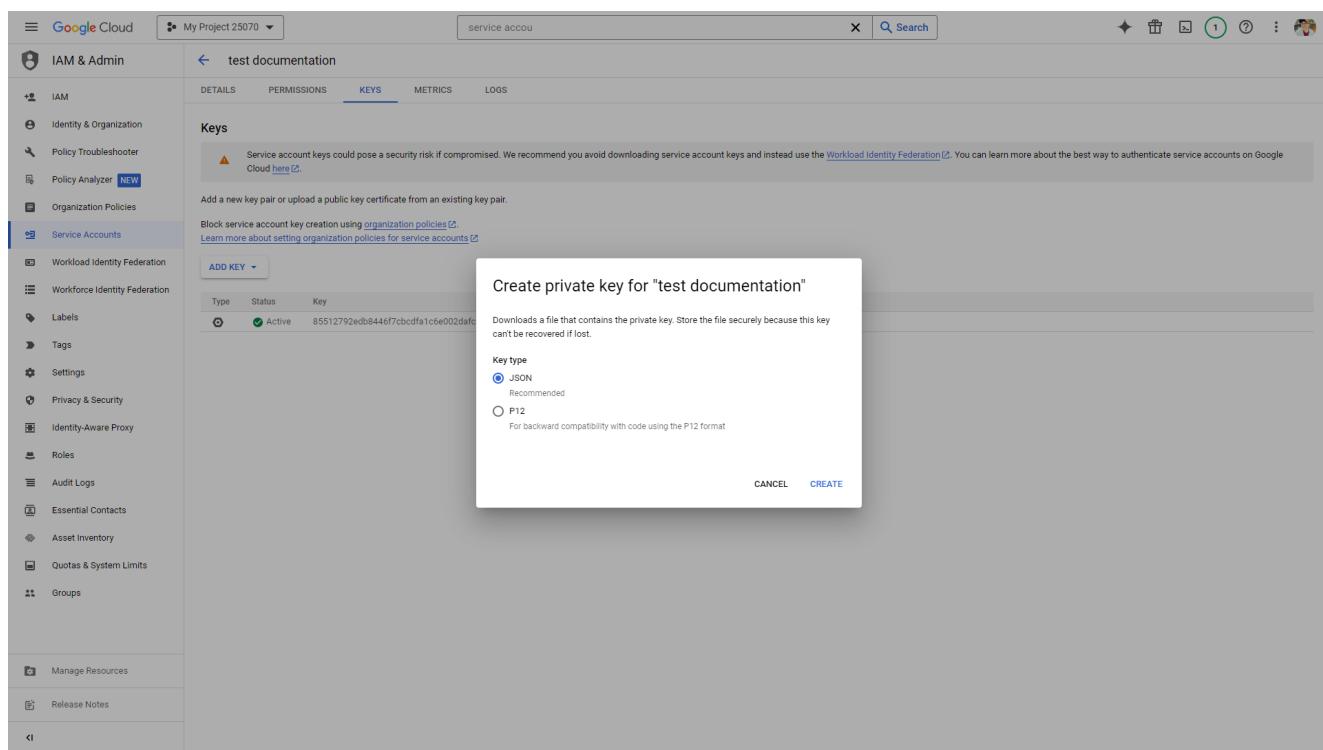


Abbildung 2. Wählen Sie die JSON- oder P12-Datei als Dienstkontoschlüssel

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:

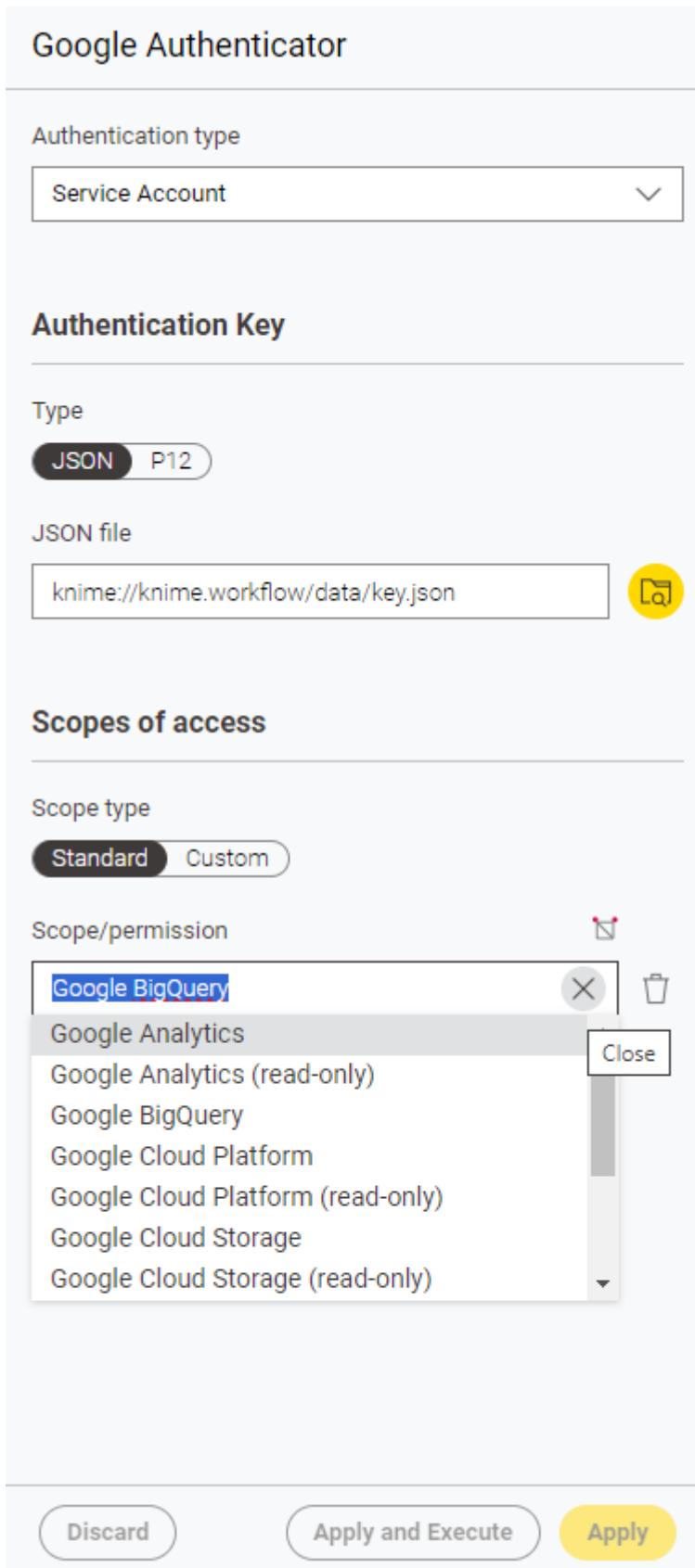


Abbildung 3. Node Konfiguration Dialog von Google Authenticator Knoten

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.

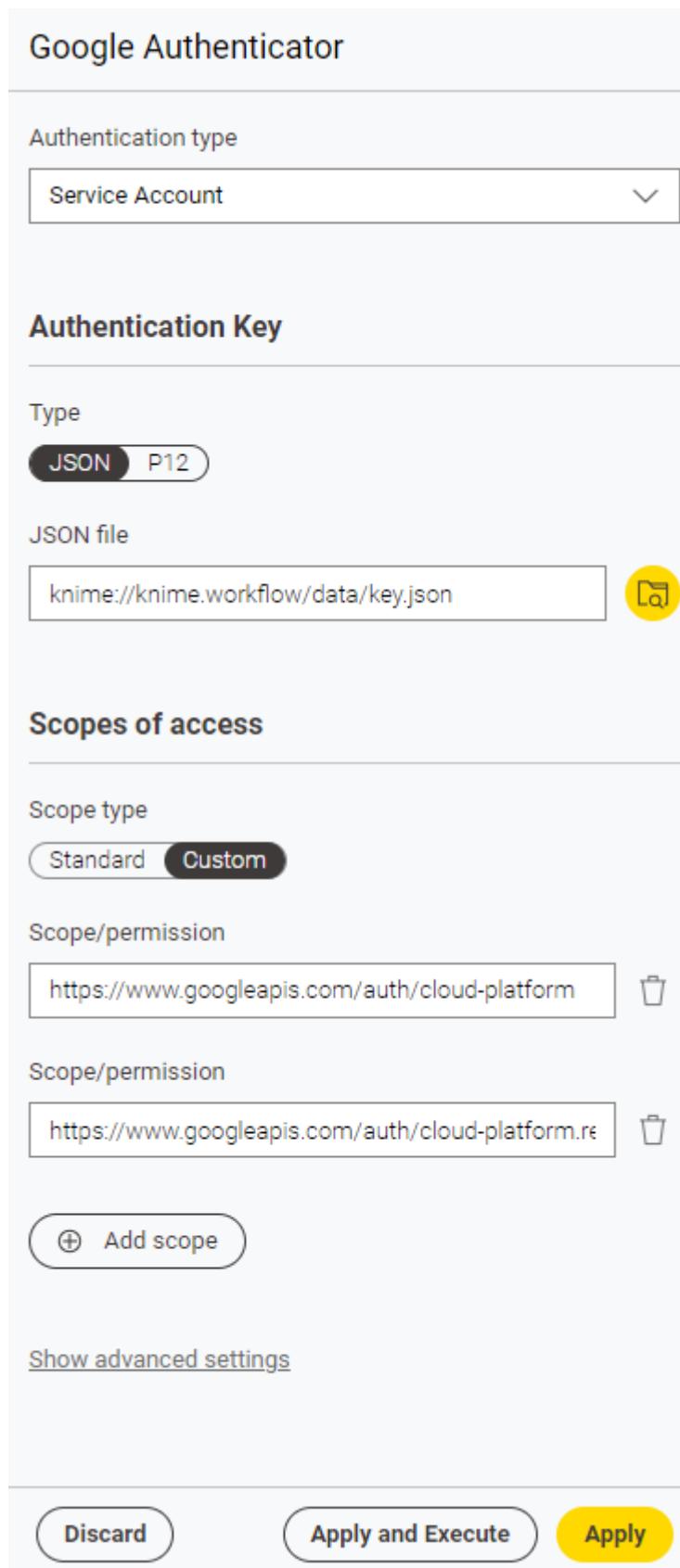


Abbildung 4. Node Konfigurationsdialog von Google Authenticator node - benutzerdefinierte Scopes

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 Angebot](#) Die 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

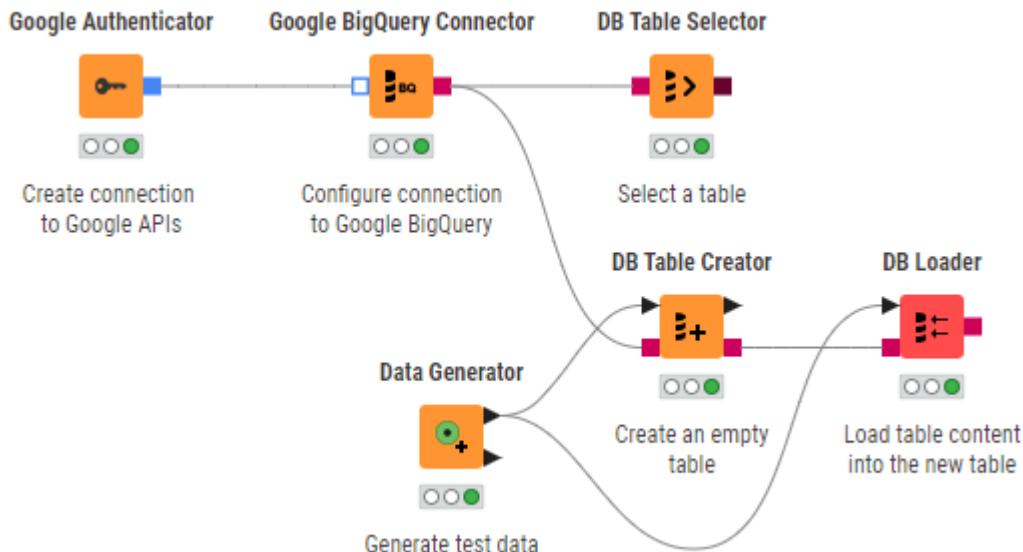


Abbildung 5. Verbinden mit und arbeiten mit Google BigQuery

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 [Abschnitt.](#page2)
Google Authenticator Node, bitte auf die



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

Für weitere Informationen über die 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

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. 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 Einrichtung [Apokalypse](#) im Cluster sind folgende zusätzliche Schritte erforderlich:

[ANHANG Kopieren der Datei ivy.!](#) 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

The screenshot shows the 'Create a cluster' page in the Google Cloud Dataproc interface. On the left, there's a sidebar with icons for Clusters, Jobs, Workflows, Autoscaling policies, Component exchange, and Notebooks. The 'Clusters' option is selected. The main area is titled 'Create a cluster' and contains several configuration sections:

- Machine configuration** section:
 - Machine family:** General-purpose (selected)
 - Series:** N1
 - Machine type:** n1-standard-4 (4 vCPU, 15 GB memory)
 - Summary:** 4 vCPU, 15 GB Memory, - GPUs
- CPU platform and GPU** section:
 - Primary disk size (minimum 15 GB):** 500 GB
 - Primary disk type:** Standard persistent disk
 - Nodes (minimum 2):** 2
 - Local SSDs (0-8):** 0 x 375 GB
- YARN cores:** 8
- YARN memory:** 24 GB
- Autoscaling policy (Optional):**
 - Enable autoscaling on the cluster.
 - This project does not currently have any applicable policy to enable autoscaling in this region. [Learn how to create autoscaling policy.](#)
- Component gateway:**
 - Enable access to the web interfaces of default and selected optional components on the cluster. [Learn more](#)
- Advanced options** (button highlighted with a red oval)
- Create** and **Cancel** buttons

Abbildung 6. Erweiterte Optionen in der Cluster-Erstellungsseite

3. Wählen Sie das Netzwerk und das Subnetz aus. Denken Sie an das Netzwerk und Subnetz für die Abschnitt.
[Libyen](#page13)

Network ?
default

Subnetwork ?
default (10.128.0.0/20)

Network tags ? (Optional)
[Empty input field]

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

Initialization actions (Optional) ?

knime-livy/livy.sh Browse X

+ Add initialization action

Project access ?
 Allow API access to all Google Cloud services in the same project. [Learn more](#)

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

The screenshot shows the 'Cluster details' page for a cluster named 'cluster-d2be'. On the left sidebar, 'Clusters' is selected. At the top right, there are buttons for 'SUBMIT JOB', 'REFRESH', 'DELETE', and 'VIEW LOGS'. Below the title, there is a note about provisioning 1TB or larger for PD-Standard without local SSDs. The 'VM INSTANCES' tab is highlighted with a red box. Under this tab, a table lists three instances: 'cluster-d2be-m' (Master), 'cluster-d2be-w-0' (Worker), and 'cluster-d2be-w-1' (Worker). The 'cluster-d2be-m' row is circled in red.

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

3. Auf der VM-Gerichte Seite, scrollen bis zum

Netzwerkschnittstellen Abschnitt. Finden Sie die
[Abschnitt,](#page10)

Netzwerk und Subnetz, das Sie im vorherigen

und Sie finden die externe IP-Adresse des Stammknotens.

The screenshot shows the 'VM instance details' page for the 'cluster-d2be-m' instance. On the left sidebar, 'VM instances' is selected. At the top right, there are buttons for 'EDIT', 'RESET', 'CREATE MACHINE IMAGE', 'CREATE SIMILAR', and more. The 'Network interfaces' section shows one interface: 'nic0' with 'default' network and 'default' subnetwork, having an 'External IP' of '35.232.249.183 (ephemeral)'. This IP address is circled in red.

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

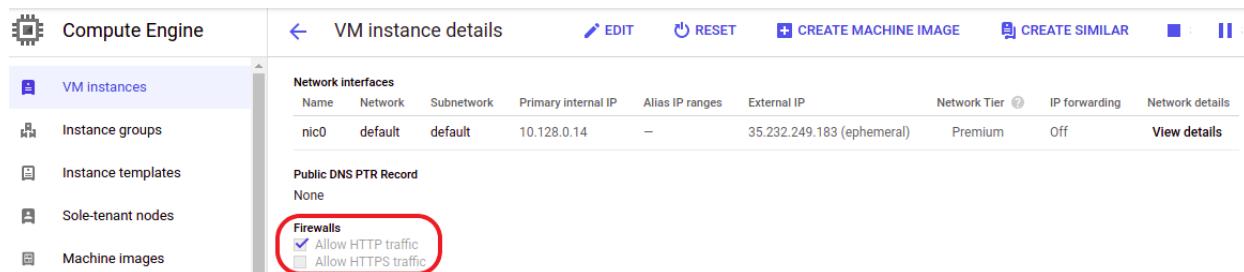


Abbildung 11. Überprüfen Erlauben Sie HTTP-Verkehr im Bereich Firewalls

L 347 vom 20.12.2013, S. 1). Weiter geht's zur [VPC-Netz Seite](#)

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

VPC network		Firewall	+ CREATE FIREWALL RULE	REFRESH	CONFIGURE LOGS	DELETE																																				
VPC networks		Firewall rules control incoming or outgoing traffic to an instance. By default, incoming traffic from outside your network is blocked. Learn more																																								
External IP addresses		Note: App Engine firewalls are managed here .																																								
Firewall		Filter table																																								
Routes		<table border="1"> <thead> <tr> <th>Name</th> <th>Type</th> <th>Targets</th> <th>Filters</th> <th>Protocols / ports</th> <th>Action</th> </tr> </thead> <tbody> <tr> <td>default-allow-http</td> <td>Ingress</td> <td>http-server</td> <td>IP ranges: 0.0.0.0/0</td> <td>tcp:80,4040,4041,8998</td> <td>Allow</td> </tr> <tr> <td>default-allow-icmp</td> <td>Ingress</td> <td>Apply to all</td> <td>IP ranges: 0.0.0.0/0</td> <td>icmp</td> <td>Allow</td> </tr> <tr> <td>default-allow-internal</td> <td>Ingress</td> <td>Apply to all</td> <td>IP ranges: 10.128.0</td> <td>tcp:0-65535 udp:0-65535 icmp</td> <td>Allow</td> </tr> <tr> <td>default-allow-rdp</td> <td>Ingress</td> <td>Apply to all</td> <td>IP ranges: 0.0.0.0/0</td> <td>tcp:3389</td> <td>Allow</td> </tr> <tr> <td>default-allow-ssh</td> <td>Ingress</td> <td>Apply to all</td> <td>IP ranges: 0.0.0.0/0</td> <td>tcp:22</td> <td>Allow</td> </tr> </tbody> </table>					Name	Type	Targets	Filters	Protocols / ports	Action	default-allow-http	Ingress	http-server	IP ranges: 0.0.0.0/0	tcp:80,4040,4041,8998	Allow	default-allow-icmp	Ingress	Apply to all	IP ranges: 0.0.0.0/0	icmp	Allow	default-allow-internal	Ingress	Apply to all	IP ranges: 10.128.0	tcp:0-65535 udp:0-65535 icmp	Allow	default-allow-rdp	Ingress	Apply to all	IP ranges: 0.0.0.0/0	tcp:3389	Allow	default-allow-ssh	Ingress	Apply to all	IP ranges: 0.0.0.0/0	tcp:22	Allow
Name	Type	Targets	Filters	Protocols / ports	Action																																					
default-allow-http	Ingress	http-server	IP ranges: 0.0.0.0/0	tcp:80,4040,4041,8998	Allow																																					
default-allow-icmp	Ingress	Apply to all	IP ranges: 0.0.0.0/0	icmp	Allow																																					
default-allow-internal	Ingress	Apply to all	IP ranges: 10.128.0	tcp:0-65535 udp:0-65535 icmp	Allow																																					
default-allow-rdp	Ingress	Apply to all	IP ranges: 0.0.0.0/0	tcp:3389	Allow																																					
default-allow-ssh	Ingress	Apply to all	IP ranges: 0.0.0.0/0	tcp:22	Allow																																					
VPC network peering																																										
Shared VPC																																										
Serverless VPC access																																										
Packet mirroring																																										

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.

VPC network		Firewall rule details	EDIT	DELETE
VPC networks				
External IP addresses				
Firewall		default-allow-http		
Routes		Logs ?		
VPC network peering		Off		
View				
Shared VPC		Network		
Serverless VPC access		default		
Packet mirroring		Priority		
		1000		
		Direction		
		Ingress		
		Action on match		
		Allow		
		Targets		
		Target tags	http-server	
		Source filters		
		IP ranges	0.0.0.0/0	
			80.154.198.250/32	
		Protocols and ports		
		tcp:80		
		tcp:4040		
		tcp:4041		
		tcp:8998		

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

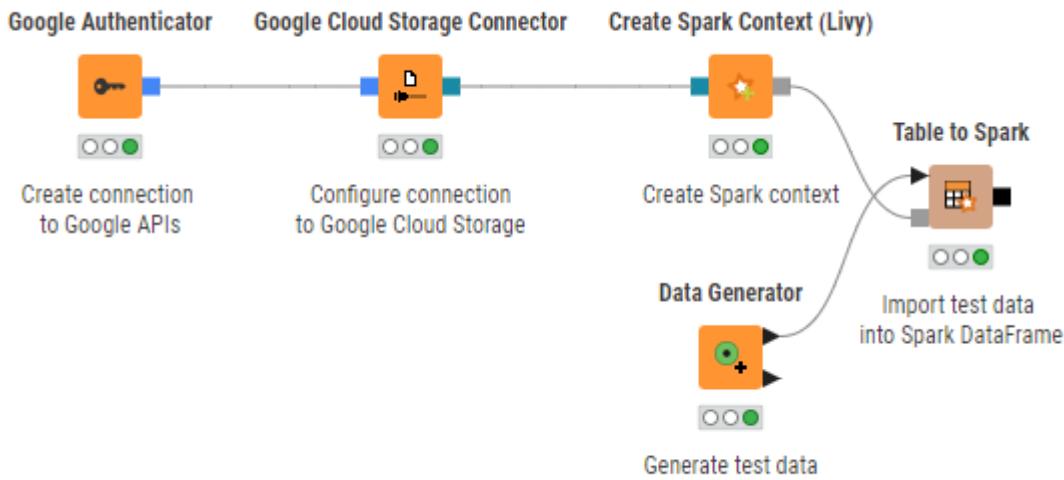


Abbildung 14. Verbindung zum Dataproc-Cluster

[Abbildung 14](#page16) 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
[Abschnitt dieser Führungen.](#page2)

[Lagerung](#page2)

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
Adressen Ihres Dataproc-Clusters, check out the

[Abschnitt](#page13)

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 Hive™ auf Dataproc in KNIME herstellen kann

Analyseplattform.

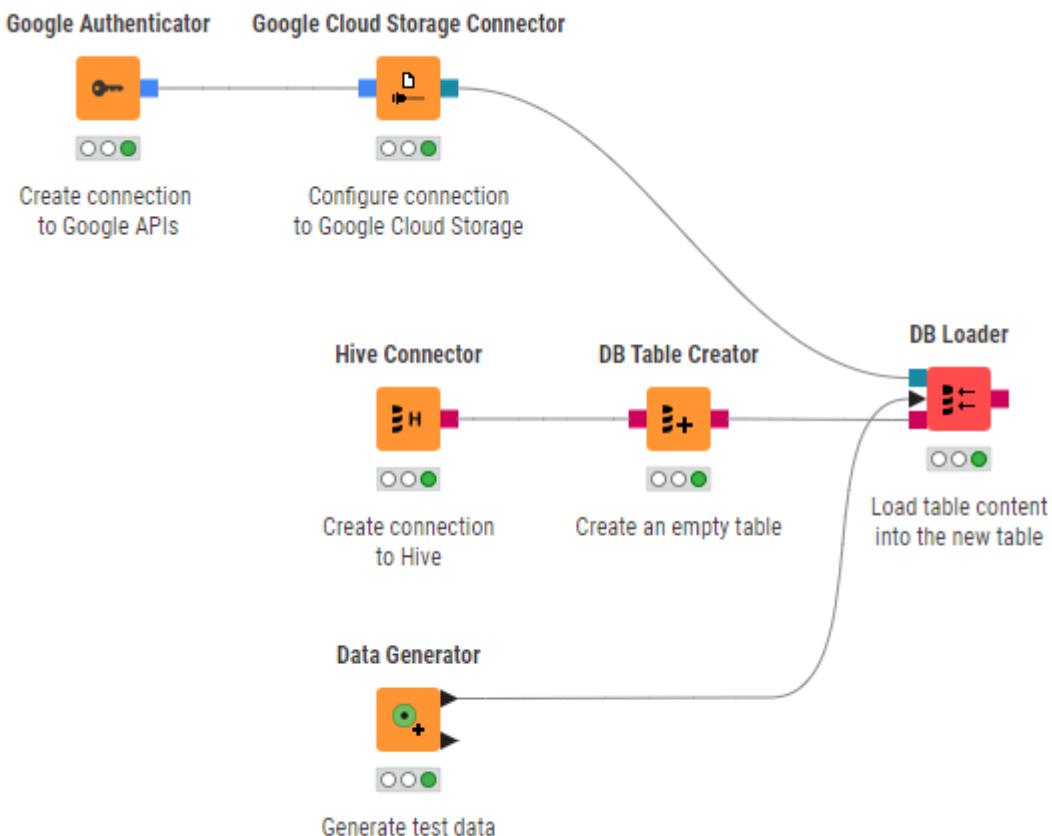


Abbildung 15. Verbinden Sie mit Hive und erstellen Sie eine Hive-Tabelle

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

Hive Tisch.

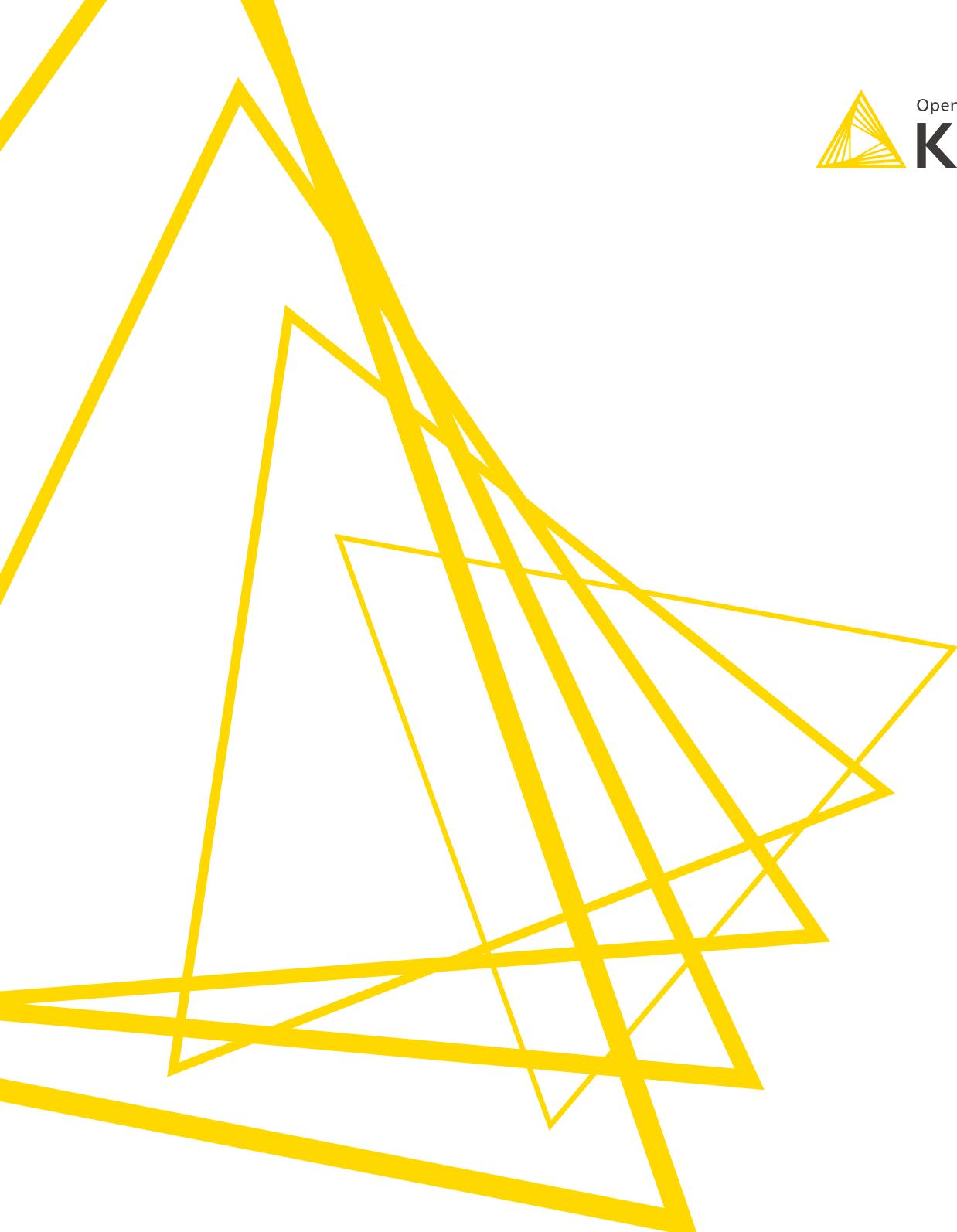
Die [Hive Connector](#) node wird standardmäßig mit dem open-source Apache Hive JDBC gebündelt Fahrer. Auch proprietäre Treiber werden unterstützt, müssen aber zuerst registriert werden. Folgen Sie der Führung über die [Registrierung eines Hive JDBC Treibers in KNIME Dokumentation](#).

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,
[~~~~](#) Sieden Sie sich hier
Abschnitt und Änderung der Firewall-Regel

entsprechend.



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