Open for Innovation
KNIME

KNIME Google Cloud Integration User Guide

KNIME AG, Zurich, Switzerland
Version 4.2 (last updated on 2023-01-11)

Table of Contents

Overview

KNIME Analytics Platform includes a set of nodes to support several Google Cloud services. The supported Google Cloud services that will be covered in this guide are Google Dataproc, Google Cloud Storage, and Google BigQuery.

KNIME Analytics Platform provides further integration for Google Drive and Google Sheets.

Google Dataproc

Cluster Setup with Livy

To create a Dataproc cluster using the Google Cloud Platform web console, follow the step-by-step guide provided by Google documentation.

To setup Apache Livy in the cluster, the following additional steps are necessary:

1. Copy the file `livy.sh` from Git repository into your cloud storage bucket. This file will be used as the initialization action to install Livy on a master node within a Dataproc cluster.

   

   Please check best practices of using initialization actions.

2. During cluster creation, open the *Advanced options* at the bottom of the page

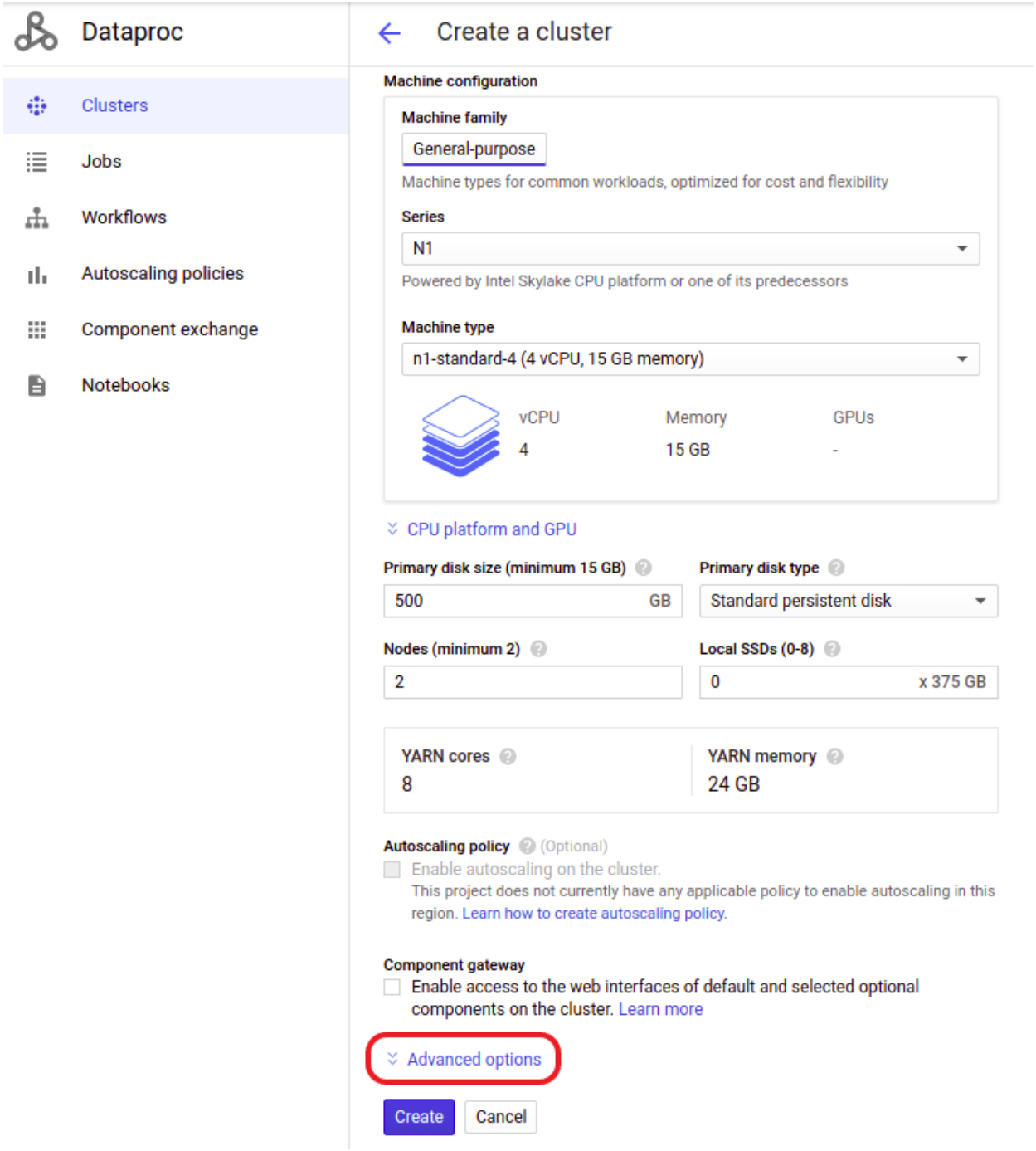

Figure 1. Advanced options in the cluster creation page

3. Select the network and subnet. Remember the network and subnet for the Access to Livy section.

Figure 2. Network and subnet

4. Select the `livy.sh` file from your cloud storage bucket in the *initialization actions* section

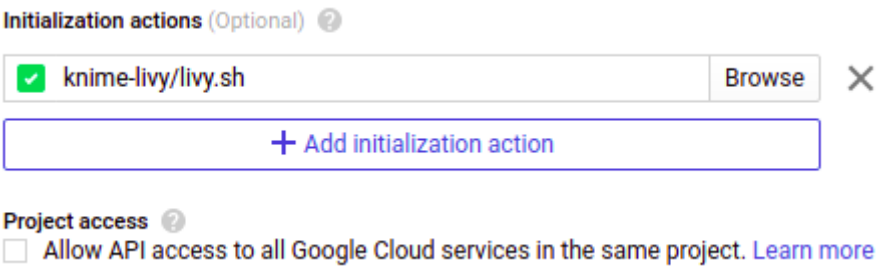

Figure 3. Set `livy.sh` as initialization action

5. Configure the rest of the cluster settings according to your needs and create the cluster.

Apache Livy is a service that interacts with a Spark cluster over a REST interface. It is the recommended service to create a Spark context in KNIME Analytics Platform.

Access to Livy

To find the external IP address of the master node where Livy is running:

1. Click on the cluster name in the cluster list page
2. Go to *VM Instances* and click on the master node

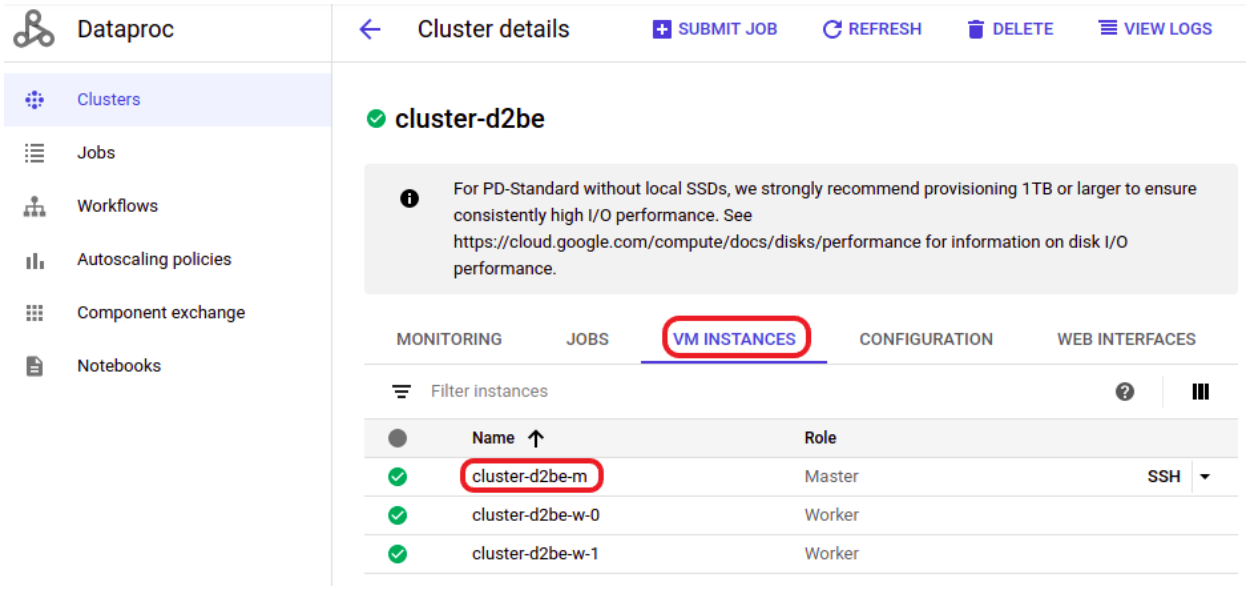

Figure 4. Select the master node in the VM instances list

3. On the *VM Instances* page, scroll down to the *Network interfaces* section. Find the network and subnet that you selected in the previous Cluster Setup with Livy section, and you will find the external IP address of the master node.

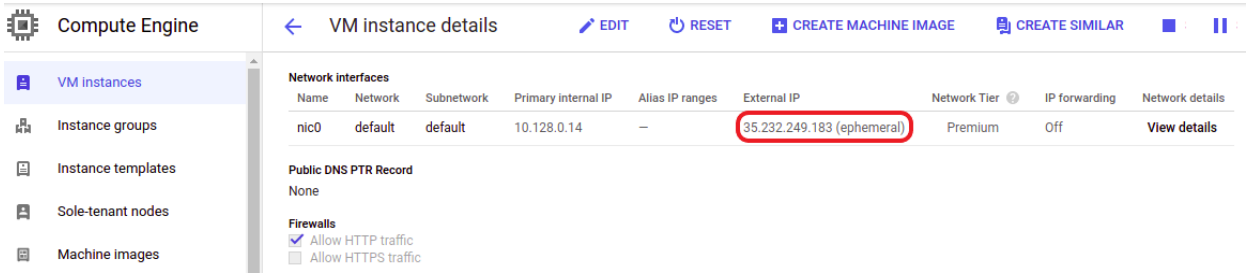

Figure 5. Find the external IP address of the master node

Livy Firewall Setup

To allow access to Livy from the outside, you have to configure the firewall:

1. Click on the cluster name in the cluster list page
2. Go to *VM Instances* and click on the master node
3. On the *VM Instances* page, scroll down to the *Firewalls* section and make sure the checkbox *Allow HTTP traffic* is enabled

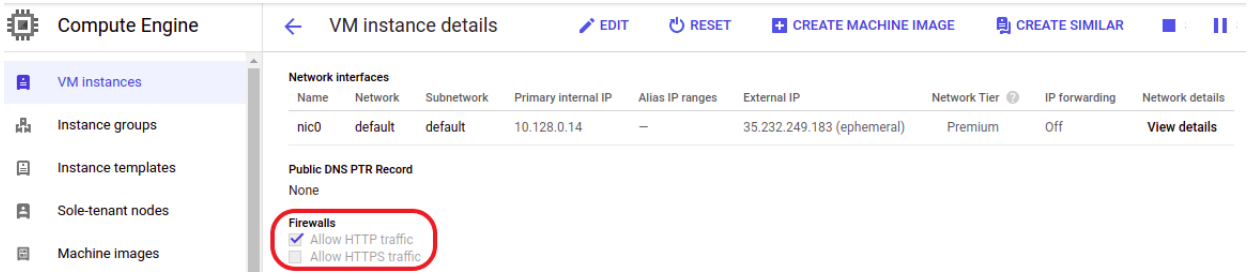

Figure 6. Check Allow HTTP traffic in the Firewalls section

4. Next, go to the *VPC network* page
5. In *Firewall* section of the *VPC network* page, select the *default-allow-http* rule

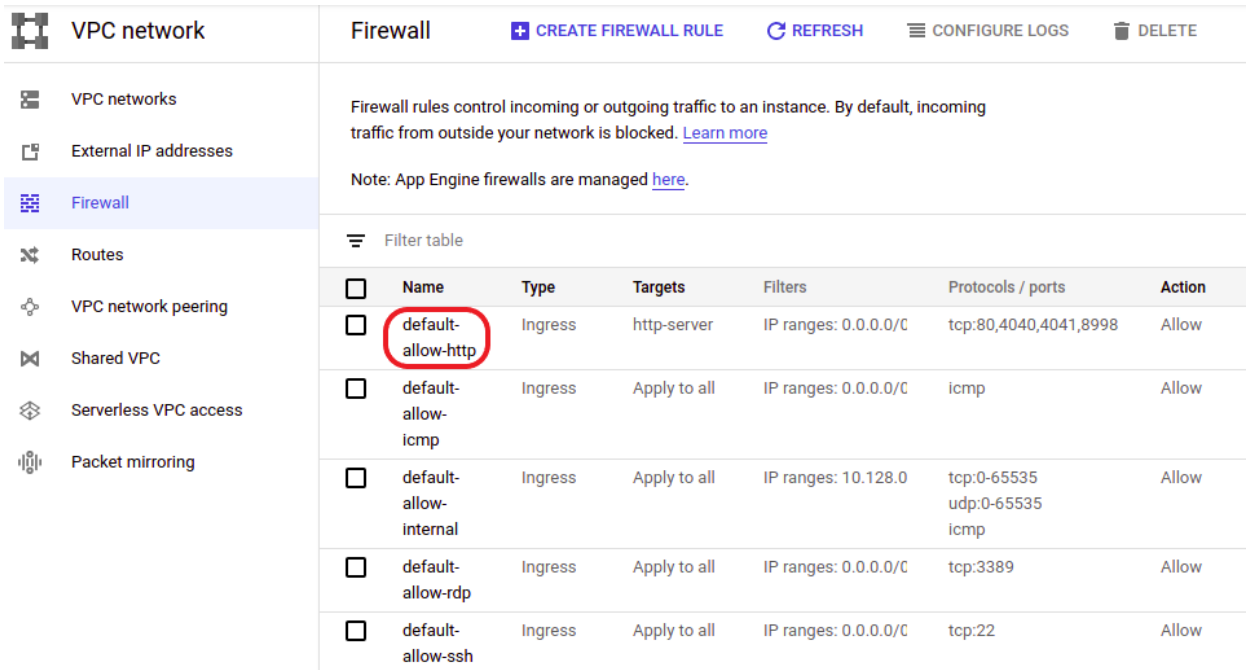

Figure 7. Open the default-allow-http firewall rule

6. Make sure that `tcp:8998` is included in the allowed protocol and ports list, and that your IP address is included in the allowed IP addresses list.

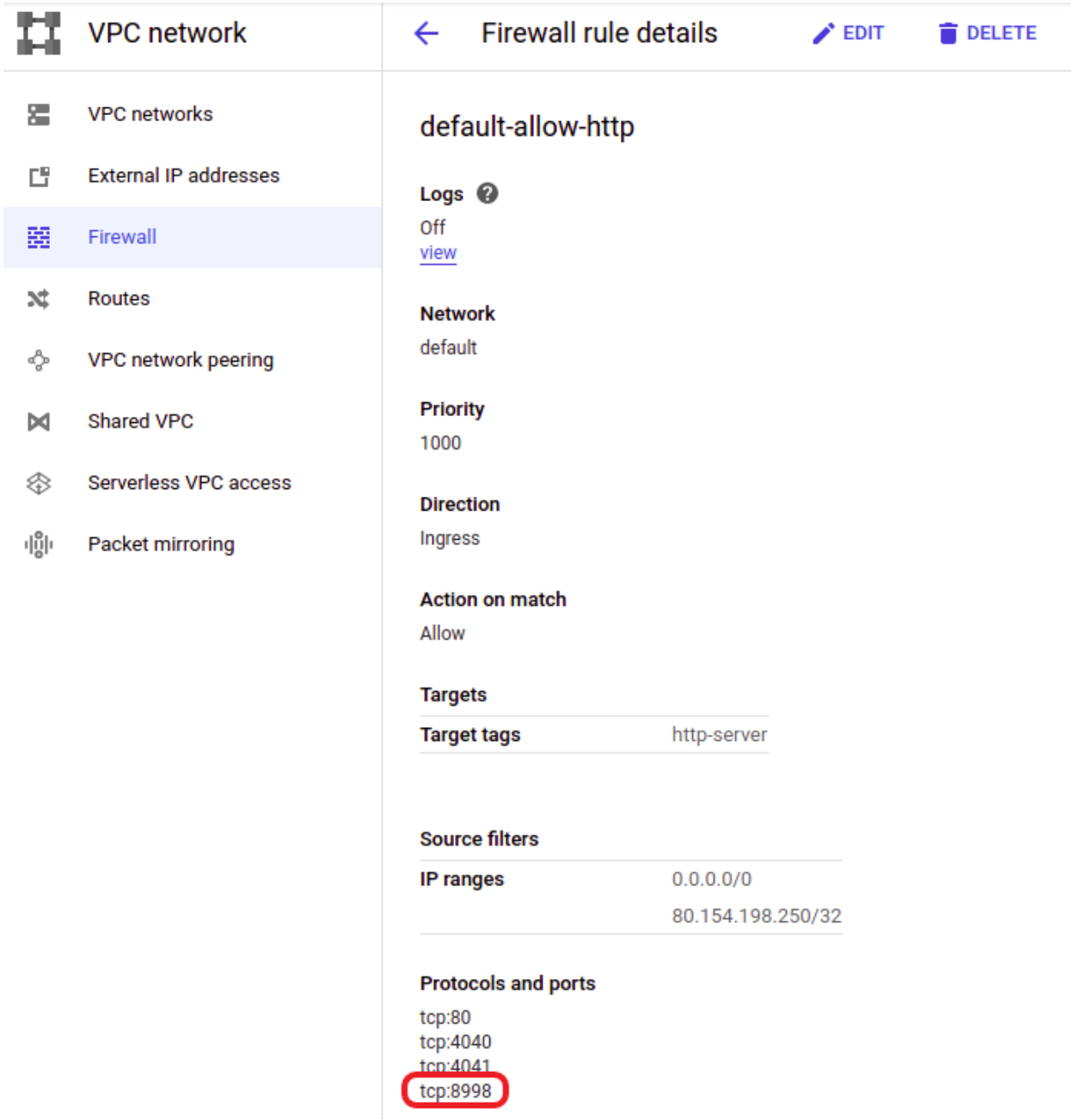

Figure 8. Make sure to allow access to certain ports and IP addresses

Once you have followed these steps, you will be able to access the Dataproc cluster via KNIME Analytics Platform using Apache Livy.

Connect to Dataproc cluster

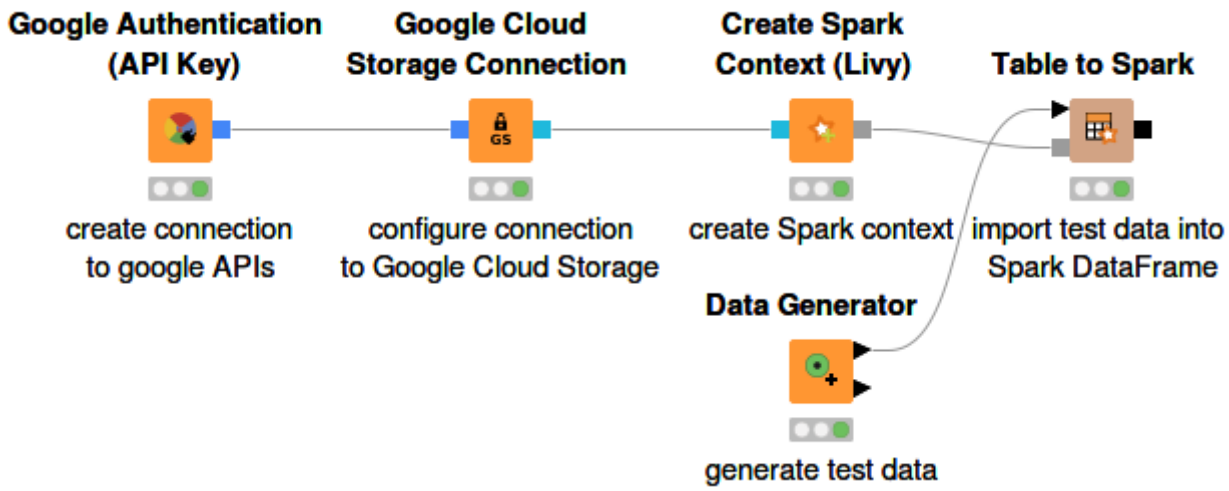

Figure 9. Connecting to Dataproc cluster

Figure 9 shows how to establish a connection to a running Dataproc cluster via KNIME Analytics Platform. The Google Authentication (API Key) node and Google Cloud Storage Connection node are used to create a connection to google APIs and to Google Cloud Storage respectively. For more information on both nodes, please check out the Google Cloud Storage section of this guide.

The Create Spark Context (Livy) node creates a Spark context via Apache Livy. Inside the node configuration dialog, the most important settings are:

- The Livy URL. It has the format `http://<IP-ADDRESS>:8998` where `<IP-ADDRESS>` is the external IP address of the master node of the Dataproc cluster. To find the external IP address of your Dataproc cluster, check out the Access to Livy section.
- Under *Advanced* tab, it is mandatory to set the *staging area for Spark jobs*. The staging area, which is located in the connected Google Cloud Storage system, will be used to exchange temporary files between KNIME and the Spark context.

The rest of settings can be configured according to your needs. For more information on the Create Spark Context (Livy) node, please check out our Amazon EMR documentation.

Once the Spark context is created, you can use any number of the KNIME Spark nodes from the KNIME Extension for Apache Spark to visually assemble your Spark analysis flow to be executed on the cluster.

Apache Hive in Google Dataproc

This section describes how to establish a connection to Apache Hive™ on Dataproc in KNIME Analytics Platform.

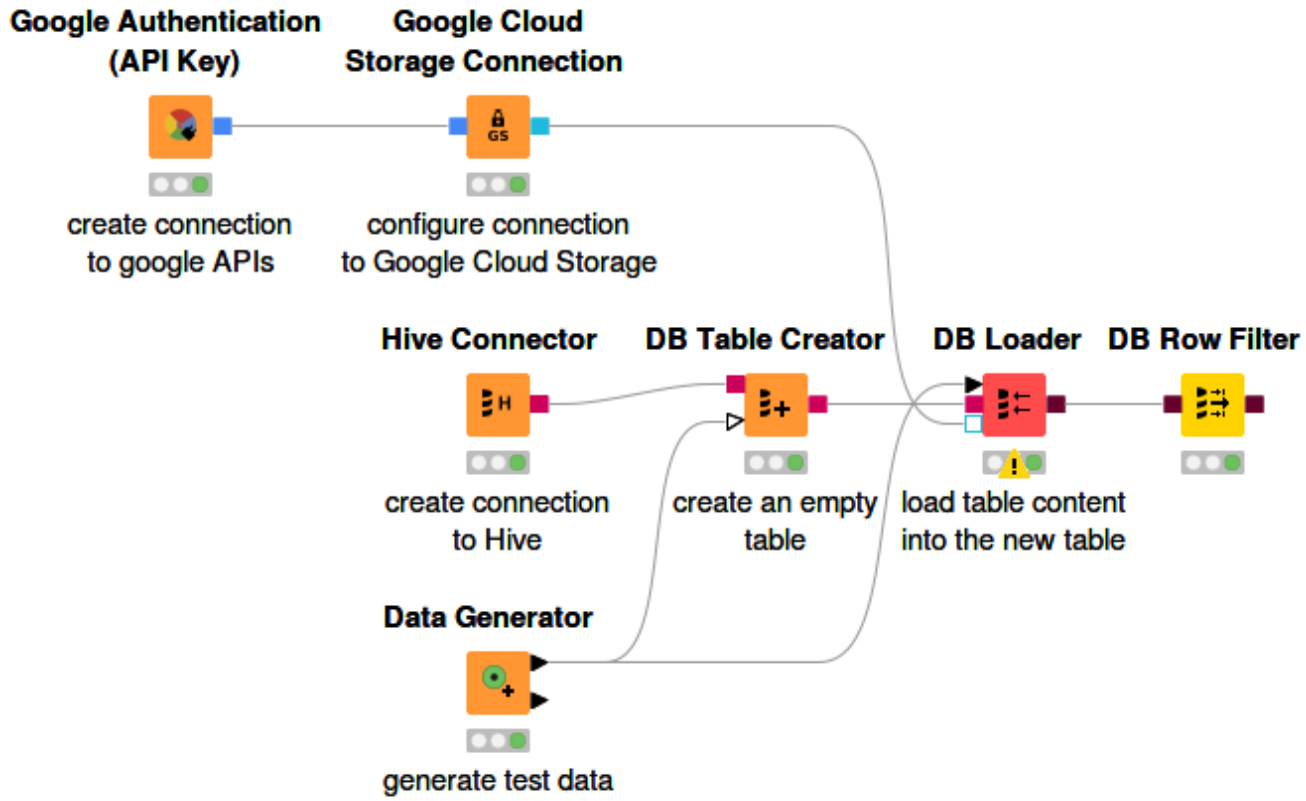

Figure 10. Connect to Hive and create a Hive table

Figure 10 shows how to connect to Hive running on a Dataproc cluster and how to create a Hive table.

The Hive Connector node is bundled by default with the open-source Apache Hive JDBC driver. Proprietary drivers are also supported, but need to be registered first. Follow the guide on how to register a Hive JDBC driver in KNIME documentation.

Once the Hive JDBC driver is registered, you can configure the Hive Connector node. For more information on how to configure the settings in the node configuration dialog, please refer to the KNIME documentation. Executing the node will create a connection to Apache Hive and you can use any KNIME database nodes to visually assemble your SQL statements.

To enable access to Hive from KNIME Analytics Platform, make sure that the Hive port (10000 by default) is opened in the firewall rules. To configure this, check out the Livy Firewall Setup section and change the firewall rule accordingly.

Google Cloud Storage

KNIME Google Cloud Storage Connection extension provides nodes to connect to Google Cloud Storage. Additionally, there is a new extension KNIME Google Cloud Storage Connector (Labs) that provide even more nodes for connecting and working in Google Cloud Storage, and it also supports the new KNIME file handling framework.

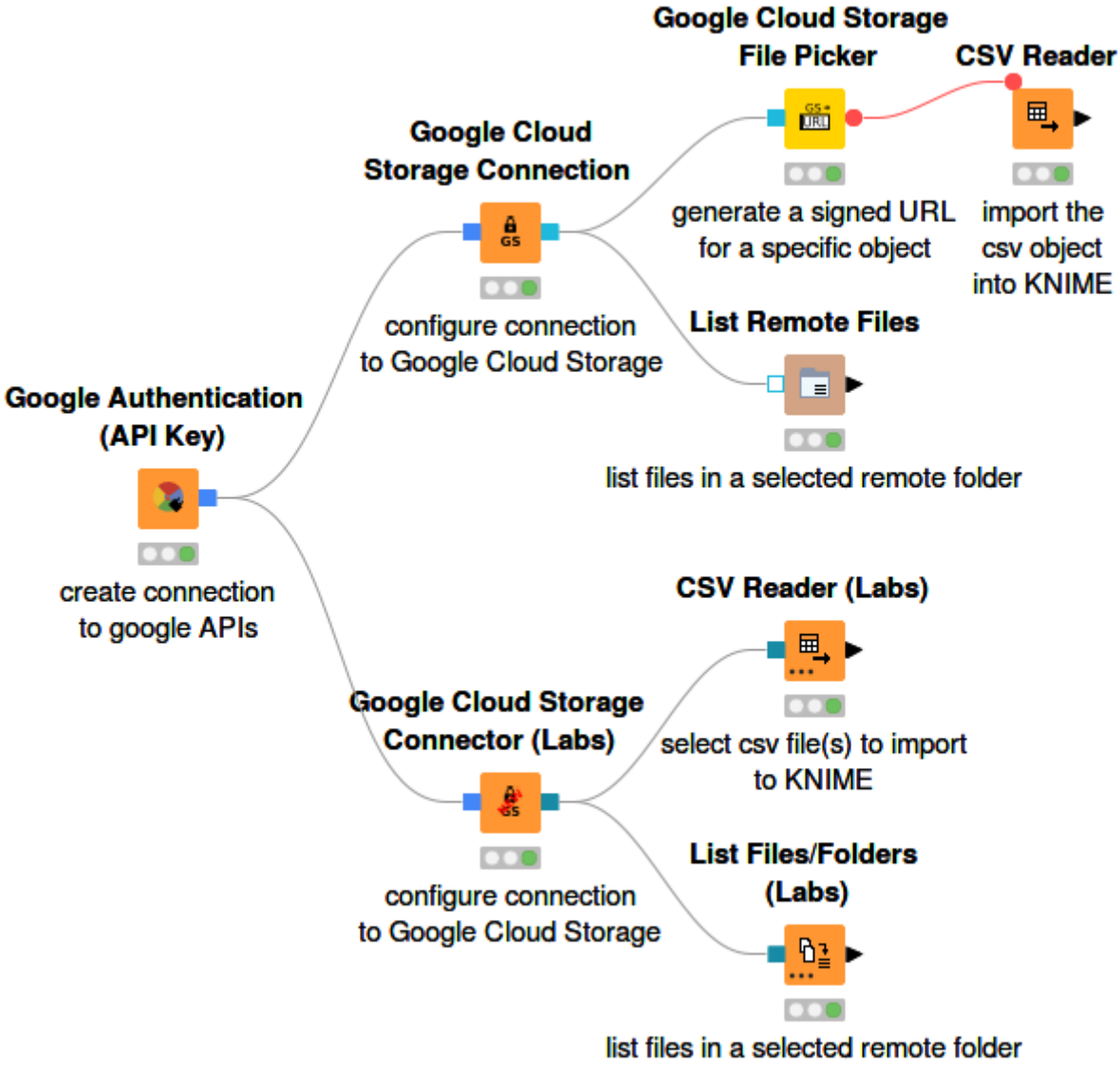

Figure 11. Connecting to and working with Google Cloud Storage

Figure 11 shows an example on how to connect to Google Cloud Storage and work with the remote files using both extensions.

Google Authentication (API Key)

The Google Authentication (API Key) node allows you to authenticate with the various Google APIs using a P12 key file. To be able to use this node, you have to create a project at the

Google Cloud Console. For more information on how to create a project on Google Cloud Console, please follow the Google documentation.

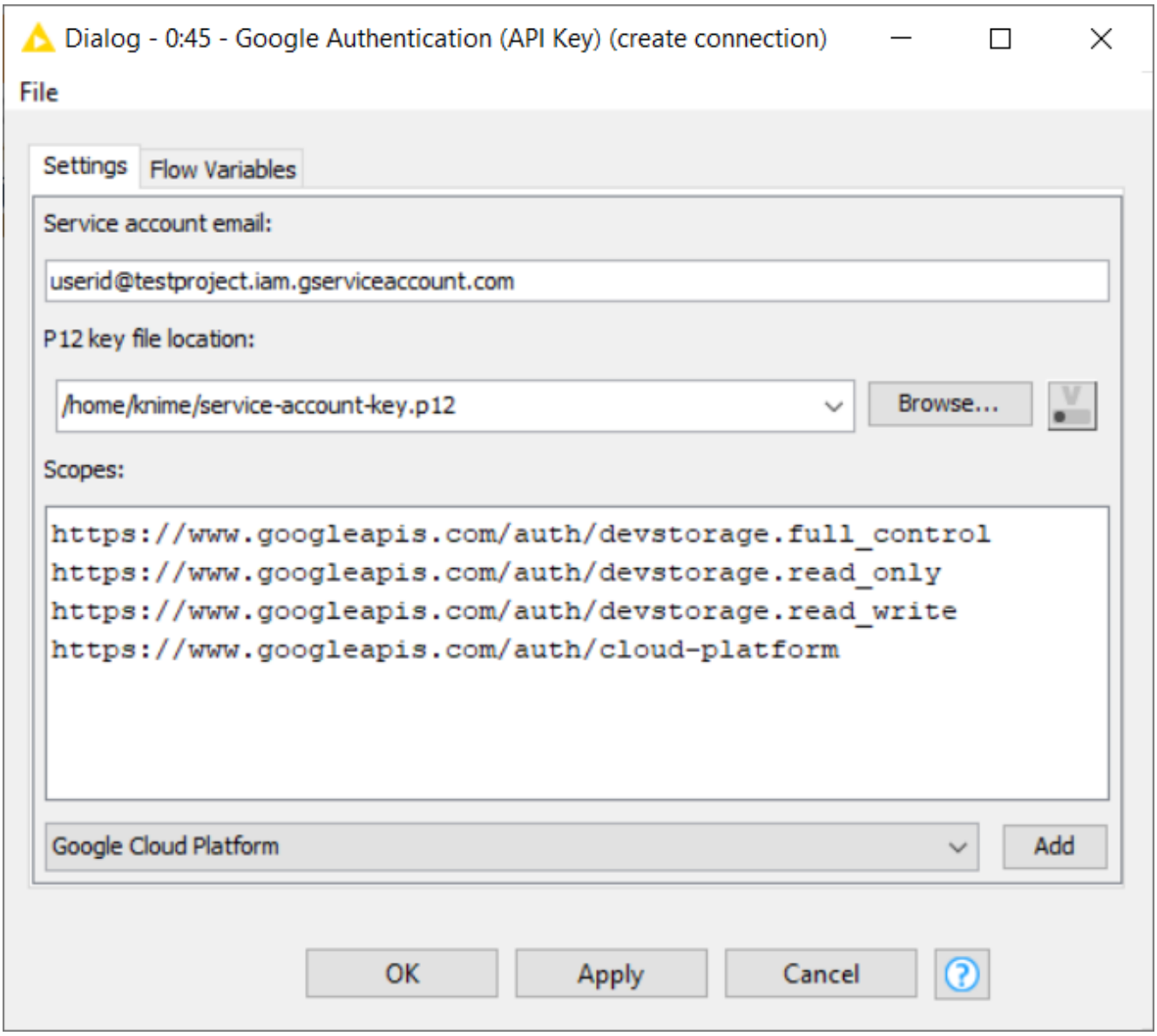

Figure 12. Node configuration dialog of Google Authentication (API Key) node

Figure 12 shows the node configuration dialog of the Google Authentication (API Key). Inside the node dialog, you have to configure the following settings:

- Service account email. If you don't have one already, please follow the Google documentation on how to create a service account. After creating the service account, it is essential to select *P12* as the service account key (see Figure 13). The service acccount email has the format of `sa-name@project-id.iam.gserviceaccount.com` where `sa-name` is a unique identifier, and `project-id` is the ID of the project.

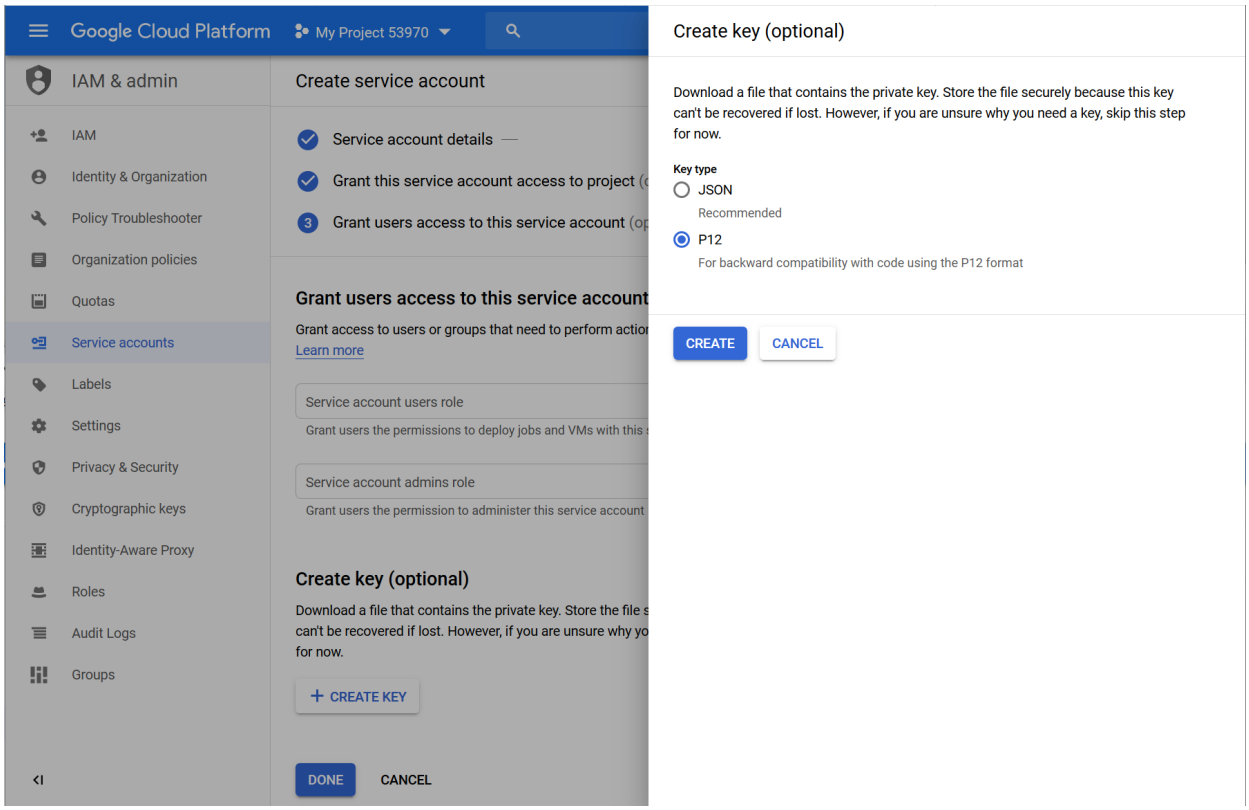

Figure 13. Select P12 file as the service account key

- P12 key file location. After creating the service account in the previous step, select *P12* as the service account key (see Figure 13). The P12 file will be downloaded automatically to your local machine. Note that you should store the P12 file in a secure place on your local system.
- The OAuth 2.0 scopes that will be granted for this connection. You should select the scopes depending on the level of access that you need.

Google Cloud Storage Connection

The Google Cloud Storage Connection node connects KNIME Analytics Platform with Google Cloud Storage. The output of this node is a Google Could Storage connection information and it allows you to work with your files inside a certain project using the remote file handling nodes. For example, you can create directory, list, delete, download and upload files from and to Google Cloud Storage.

The node configuration dialog of the Google Cloud Storage Connection node is fairly simple. You need to enter the *Project identifier*, which is the Google Cloud project ID. For more information on how to find your project ID, please check out the Google documentation.

Google Cloud Storage File Picker

The Google Cloud Storage File Picker node creates a pre-signed URL for a specific file. This URL can then be used by any of the reader nodes in KNIME Analytics Platform to read the selected file directly from Google Cloud Storage, or share it with other users without the need for authentication. Please note that the generated URL is only valid for a specific period of time. Upon expiration, the URL will no longer remain active and an attempt to access the URL will generate an error.

In the node configuration dialog of the Google Cloud Storage File Picker node, you have to select:

- The remote file, for which the signed URL should be created
- The expiration time. This defines the duration or UTC time after the signed URL becomes invalid. The maximum expiration time on Google Cloud Storage is 7 days.

The output of this node is a flow variable containing the signed URL pointing to the selected file. You can connect this flow variable to any KNIME file handling node, e.g the CSV Reader node (as shown in Figure 11). Inside the node configuration dialog of the CSV Reader node, simply set the input location using the value from the flow variable.

Google Cloud Storage Connector (Labs)

The Google Cloud Storage Connector (Labs) node also connects to Google Cloud Storage and allows downstream nodes to access Google Cloud Storage inside a certain project using the new KNIME file handling nodes.

The new KNIME file handling nodes are a part of KNIME Analytics Platform installation and can be found under *Node Repository > KNIME Labs > File Handling (Labs)*.

The node configuration dialog of the Google Cloud Storage Connector (Labs) node contains:

- Project ID. This is the Google Cloud project ID. For more information on how to find your project ID, please check out the Google documentation.
- Working directory. The working directory must be specified as an absolute path and it allows downstream nodes to access files/folders using relative paths, i.e. paths that do not have a leading slash. If not specified, the default working directory is `/`.

  Path syntax: Paths for Google Cloud Storage are specified with a UNIX-like syntax, e.g. `/mybucket/myfolder/myfile`. The path usually consists of:

    - A leading slash (`/`)
    - Followed by the name of a bucket (`mybucket` in the above example), followed by a slash
    - Followed by the name of an object within the bucket (`myfolder/myfile` in the above example).
- Normalize paths. Path normalization eliminates redundant components of a path, e.g. `/a/../b/./c` can be normalized to `/b/c`. When these redundant components like `../` or `.` are part of an existing object, then normalization must be deactivated in order to access them properly.
- Under the *Advanced* tab, it is possible to set the connection and read timeout.

This node currently only supports the Google Authentication (API key) node for authentication.

Google BigQuery

KNIME Analytics Platform includes a set of nodes to support Google BigQuery. The KNIME BigQuery extension is available from KNIME Analytics Platform version 4.1.

Setting up KNIME Analytics Platform for Google BigQuery has the following prerequisites:

1. Create a project in the Google Cloud Console. For more information on how to create a project on Google Cloud Console, please follow the Google documentation.
2. Create a service account. If you don't have one already, please follow the Google documentation on how to create a service account. It is essential to select *P12* as the service account key.
3. Download the JDBC driver for Google BigQuery, unzip, and store it in your local machine. Register the JDBC driver on KNIME Analytics Platform by following the tutorial in the KNIME documentation.

Connect to BigQuery

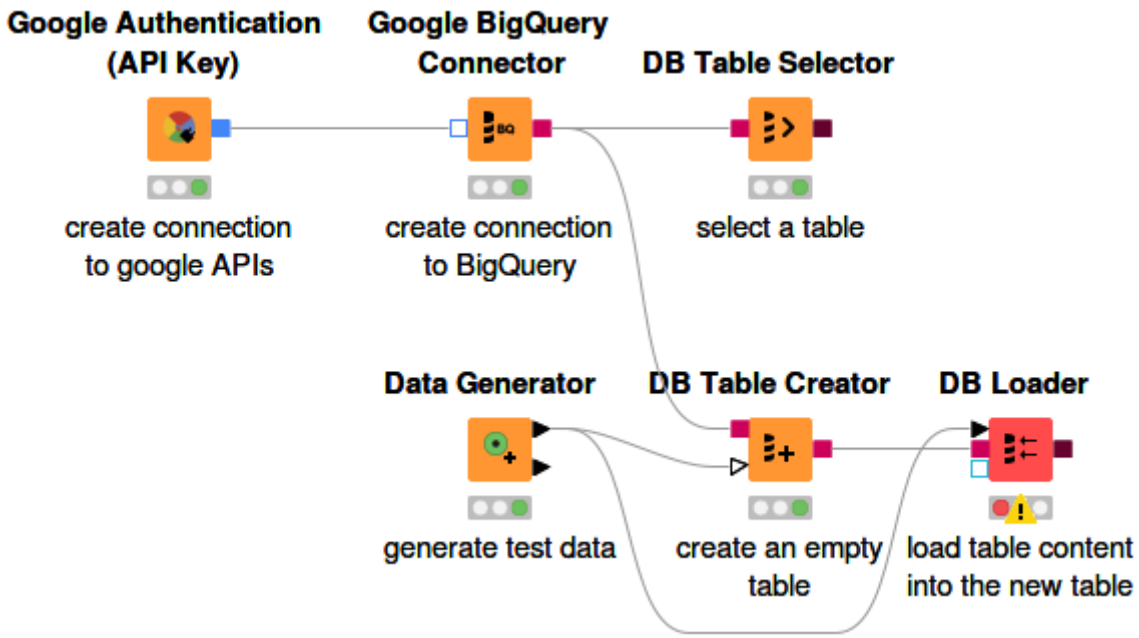

Figure 14. Connecting to and working with Google BigQuery

Figure 14 shows how to authenticate using the Google Authentication (API Key) node and the Google BigQuery Connector node to establish a connection to BigQuery via JDBC driver. To configure Google Authentication (API Key) node, please refer to the Google Authentication (API Key) section.

To configure the Google BigQuery Connector node, please check out how to connect to a

predefined database in the KNIME documentation. For the hostname in BigQuery, you can specify *https://www.googleapis.com/bigquery/v2* or *bigquery.cloud.google.com*. As the database name, use the project name you created on the Google Cloud Console.

For more information on the *JDBC parameters* tab or the *Advanced* tab in the node configuration dialog of Google BigQuery Connector node, please check out the KNIME documentation.

Executing this node will create a connection to the BigQuery database and you can use any KNIME database nodes to visually assemble your SQL statements.

For more information on KNIME database nodes, please check out the KNIME Database documentation.

Create a BigQuery table

To export data from KNIME Analytics Platform to Google BigQuery (shown in Figure 14):

1. Create the database schema/dataset where you want to store the table, if it doesn't exist already. To create a dataset, please check the Google documentation.
2. Create an empty table with the right specification. To do this, use the DB Table Creator node. Inside the node configuration dialog, specify the schema as the name of the dataset that you created in the previous step. For more information on the DB Table Creator node, please check the KNIME documentation.

   

   If the table has column names that contain space characters, e.g. `column 1`, make sure to delete the space characters because they would be automatically replaced with _ during table creation, e.g. `column_1` and this will lead to conflict, since column names will no longer match.

3. Once the empty table is created, use the DB Loader node to load the table content into the newly created table. For more information on the DB Loader node, please check the KNIME documentation.

KNIME AG
Talacker 50
8001 Zurich, Switzerland
www.knime.com
info@knime.com