www.cloudskillsboost.google /course_sessions/3591643/labs/379262

Building and Executing a Pipeline Graph with Data Fusion 2.5 | Google Cloud Skills Boost

Qwiklabs : 15-19 minutes

Overview

This tutorial shows you how to use the Wrangler and Data Pipeline features in Cloud Data Fusion to clean, transform, and process taxi trip data for further analysis.

What you learn

In this lab, you will:

- Connect Cloud Data Fusion to a couple of data sources
- Apply basic transformations
- Join two data sources
- Write data to a sink

Introduction

Often times, data needs to go through a number of pre-processing steps before analysts can leverage the data to glean insights. For example, data types may need to be adjusted, anomalies removed, and vague identifiers may need to be converted to more meaningful entries. Cloud Data Fusion is a service for efficiently building ETL/ELT data pipelines. Cloud Data Fusion uses Cloud Dataproc cluster to perform all transforms in the pipeline.

The use of Cloud Data Fusion will be exemplified in this tutorial by using a subset of the NYC TLC Taxi Trips dataset on BigQuery.

Setup and requirements

For each lab, you get a new Google Cloud project and set of resources for a fixed time at no cost.

1. Sign in to Qwiklabs using an **incognito window**.
2. Note the lab's access time (for example, `1:15:00`), and make sure you can finish within that time. There is no pause feature. You can restart if needed, but you have to start at the beginning.
3. When ready, click **Start lab**.
4. Note your lab credentials (**Username** and **Password**). You will use them to sign in to the Google Cloud Console.
5. Click **Open Google Console**.

6. Click **Use another account** and copy/paste credentials for **this** lab into the prompts.
   If you use other credentials, you'll receive errors or **incur charges**.

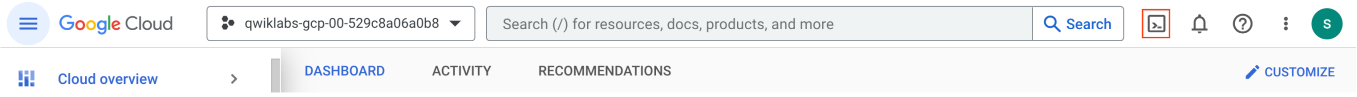
7. Accept the terms and skip the recovery resource page.

Activate Google Cloud Shell

Google Cloud Shell is a virtual machine that is loaded with development tools. It offers a persistent 5GB home directory and runs on the Google Cloud.

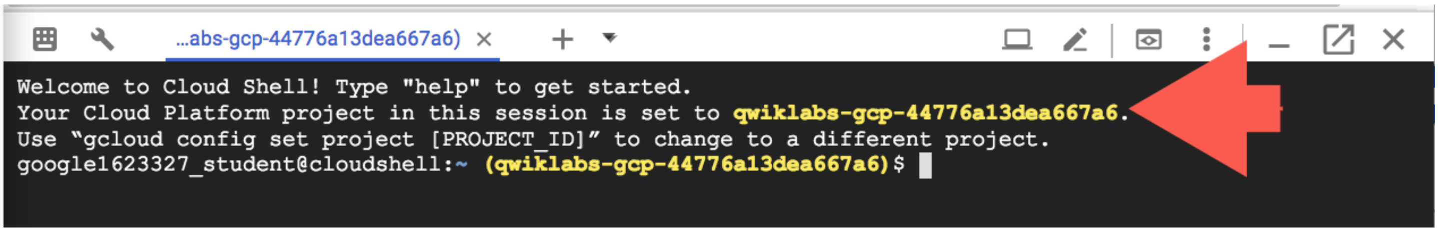
Google Cloud Shell provides command-line access to your Google Cloud resources.

1. In Cloud console, on the top right toolbar, click the Open Cloud Shell button.


Google Cloud
qwiklabs-gcp-00-529c8a06a0b8
Search (/) for resources, docs, products, and more
Search
S
Cloud overview
DASHBOARD
ACTIVITY
RECOMMENDATIONS
CUSTOMIZE

2. Click **Continue**.

It takes a few moments to provision and connect to the environment. When you are connected, you are already authenticated, and the project is set to your *PROJECT_ID*. For example:



```
Welcome to Cloud Shell! Type "help" to get started.
Your Cloud Platform project in this session is set to qwiklabs-gcp-44776a13dea667a6.
Use "gcloud config set project [PROJECT_ID]" to change to a different project.
google1623327_student@cloudshell:~ (qwiklabs-gcp-44776a13dea667a6)$
```

gcloud is the command-line tool for Google Cloud. It comes pre-installed on Cloud Shell and supports tab-completion.

- You can list the active account name with this command:

gcloud auth list

Output:

Credentialed accounts: - @.com (active)

Example output:

Credentialed accounts: - google1623327_student@qwiklabs.net

- You can list the project ID with this command:

gcloud config list project

Output:

[core] project =

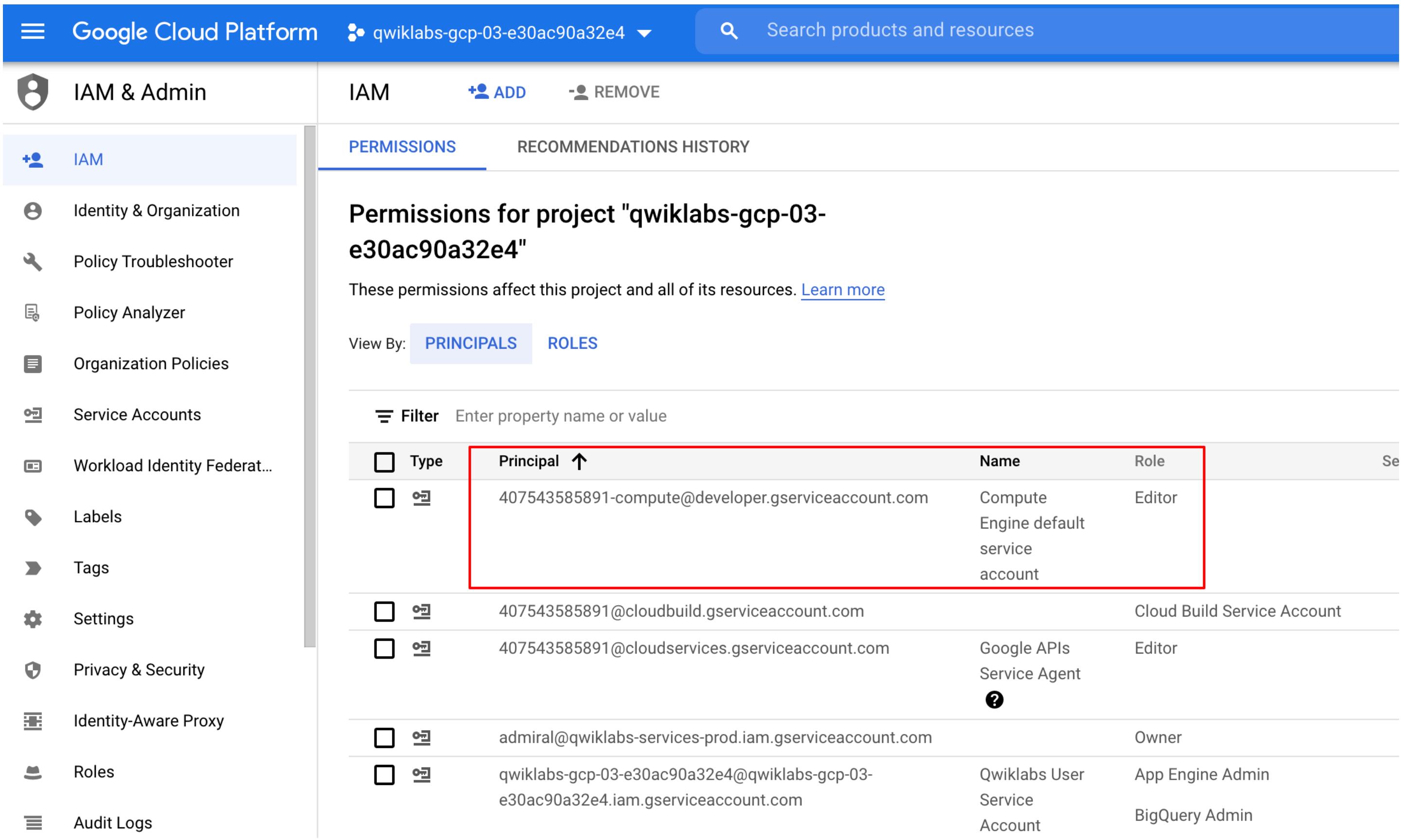
Example output:

[core] project = qwiklabs-gcp-44776a13dea667a6 **Note:** Full documentation of **gcloud** is available in the gcloud CLI overview guide .

Check project permissions

Before you begin your work on Google Cloud, you need to ensure that your project has the correct permissions within Identity and Access Management (IAM).

1. In the Google Cloud console, on the **Navigation menu** (☰), select **IAM & Admin** > **IAM**.
2. Confirm that the default compute Service Account `{project-number}-compute@developer.gserviceaccount.com` is present and has the `editor` role assigned. The account prefix is the project number, which you can find on **Navigation menu** > **Home**.



Permissions for project "qwiklabs-gcp-03-e30ac90a32e4"

These permissions affect this project and all of its resources. Learn more

View By: PRINCIPALS ROLES

Filter Enter property name or value

☐	Type	Principal ↑	Name	Role	Se
☐		407543585891-compute@developer.gserviceaccount.com	Compute Engine default service account	Editor	
☐		407543585891@cloudbuild.gserviceaccount.com		Cloud Build Service Account	
☐		407543585891@cloudservices.gserviceaccount.com	Google APIs Service Agent ❓	Editor	
☐		admiral@qwiklabs-services-prod.iam.gserviceaccount.com		Owner	
☐		qwiklabs-gcp-03-e30ac90a32e4@qwiklabs-gcp-03-e30ac90a32e4.iam.gserviceaccount.com	Qwiklabs User Service Account	App Engine Admin BigQuery Admin	

Note: If the account is not present in IAM or does not have the `editor` role, follow the steps below to assign the required role.

1. In the Google Cloud console, on the **Navigation menu**, click **Home**.
2. Copy the project number (e.g. `729328892908`).
3. On the **Navigation menu**, select **IAM & Admin** > **IAM**.
4. At the top of the **IAM** page, click **Add**.
5. For **New principals**, type:

{project-number}-compute@developer.gserviceaccount.com

6. Replace `{project-number}` with your project number.

7. For **Role**, select **Project** (or Basic) > **Editor**.
8. Click **Save**.

Task 1. Creating a Cloud Data Fusion instance

Thorough directions for creating a Cloud Data Fusion instance can be found in the Creating a Cloud Data Fusion instance Guide. The essential steps are as follows:

1. To ensure the training environment is properly configured you must first stop and restart the Cloud Data Fusion API. Run the command below in the Cloud Shell. It will take a few minutes to complete.

```
gcloud services disable datafusion.googleapis.com
```

Your output says that the operation finished successfully.

Next, restart the connection to the Cloud Data Fusion API.

2. In the Google Cloud Console, enter **Cloud Data Fusion API** in the top search bar. Click on the result for Cloud Data Fusion API.

3. On the page that loads click **Enable**.

4. When the API has been enabled again, the page will refresh and show the option to disable the API along with other details on the API usage and performance.

5. On the **Navigation menu**, select **Data Fusion**.

6. To create a Cloud Data Fusion instance, click **Create an Instance**.

7. Enter a name for your instance.

8. Select **Basic** for the Edition type.

9. Under **Authorization** section, click **Grant Permission**.

10. Leave all other fields as their defaults and click **Create**.

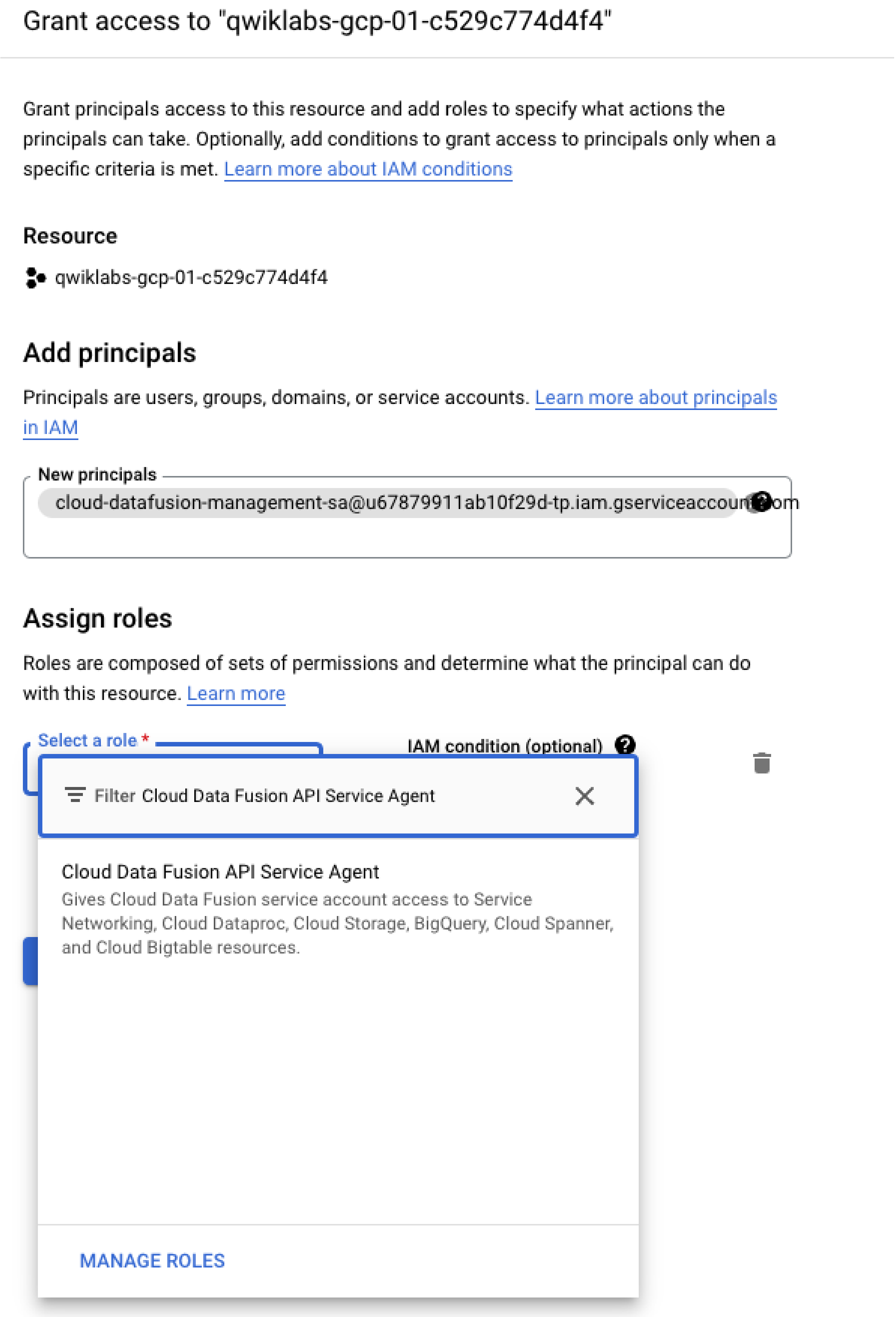
Note: Creation of the instance can take around 15 minutes.

11. Once the instance is created, you need one additional step to grant the service account associated with the instance permissions on your project. Navigate to the instance details page by clicking the instance name.

12. Copy the service account to your clipboard.

13. In the GCP Console navigate to the **IAM & Admin > IAM**.

14. On the IAM Permissions page, click **+Grant Access** add the service account you copied earlier as a new principals and grant the **Cloud Data Fusion API Service Agent** role.



Grant access to "qwiklabs-gcp-01-c529c774d4f4"

Grant principals access to this resource and add roles to specify what actions the principals can take. Optionally, add conditions to grant access to principals only when a specific criteria is met. Learn more about IAM conditions

Resource

qwiklabs-gcp-01-c529c774d4f4

Add principals

Principals are users, groups, domains, or service accounts. Learn more about principals in IAM

New principals: cloud-datafusion-management-sa@u67879911ab10f29d-tp.iam.gserviceaccount.com

Assign roles

Roles are composed of sets of permissions and determine what the principal can do with this resource. Learn more

Select a role *

IAM condition (optional)

Filter Cloud Data Fusion API Service Agent

Cloud Data Fusion API Service Agent
Gives Cloud Data Fusion service account access to Service Networking, Cloud Dataproc, Cloud Storage, BigQuery, Cloud Spanner, and Cloud Bigtable resources.

15. Click **Save**.

Task 2. Loading the data

Once the Cloud Data Fusion instance is up and running, you can start using Cloud Data Fusion. However, before Cloud Data Fusion can start ingesting data you have to take some preliminary steps.

1. In this example, Cloud Data Fusion will read data out of a storage bucket. In the cloud shell console execute the following commands to create a new bucket and copy the relevant data into it:

export BUCKET=$GOOGLE_CLOUD_PROJECT gcloud storage buckets create gs://$BUCKET gcloud storage cp gs://cloud-training/OCBL017/ny-taxi-2018-sample.csv gs://$BUCKET **Note:** The created bucket name is your project id.

2. In the command line, execute the following command to create a bucket for temporary storage items that Cloud data Fusion will create:

gcloud storage buckets create gs://$BUCKET-temp **Note:** The created bucket name is your project id followed by "-temp".

3. Click the **View Instance** link on the Data Fusion instances page, or the details page of an instance. Click **username**. If prompted to take a tour of the service click on **No, Thanks**. You should now be in the Cloud Data Fusion UI.

Note: You may need to reload or refresh the Cloud Fusion UI pages to allow prompt loading of the page.

4. **Wrangler** is an interactive, visual tool that lets you see the effects of transformations on a small subset of your data before dispatching large, parallel-processing jobs on the entire dataset. On the Cloud Data Fusion UI, choose **Wrangler**. On the left side, there is a panel with the pre-configured connections to your data, including the Cloud Storage connection.

5. Under **GCS**, select **Cloud Storage Default**.

6. Click on the bucket corresponding to your project name.

7. Select **ny-taxi-2018-sample.csv**. The data is loaded into the Wrangler screen in row/column form.

8. In the **Parsing Options** window, set **Use First Row as Header** as `True`. The data splits into multiple columns.

9. Click **Confirm**.

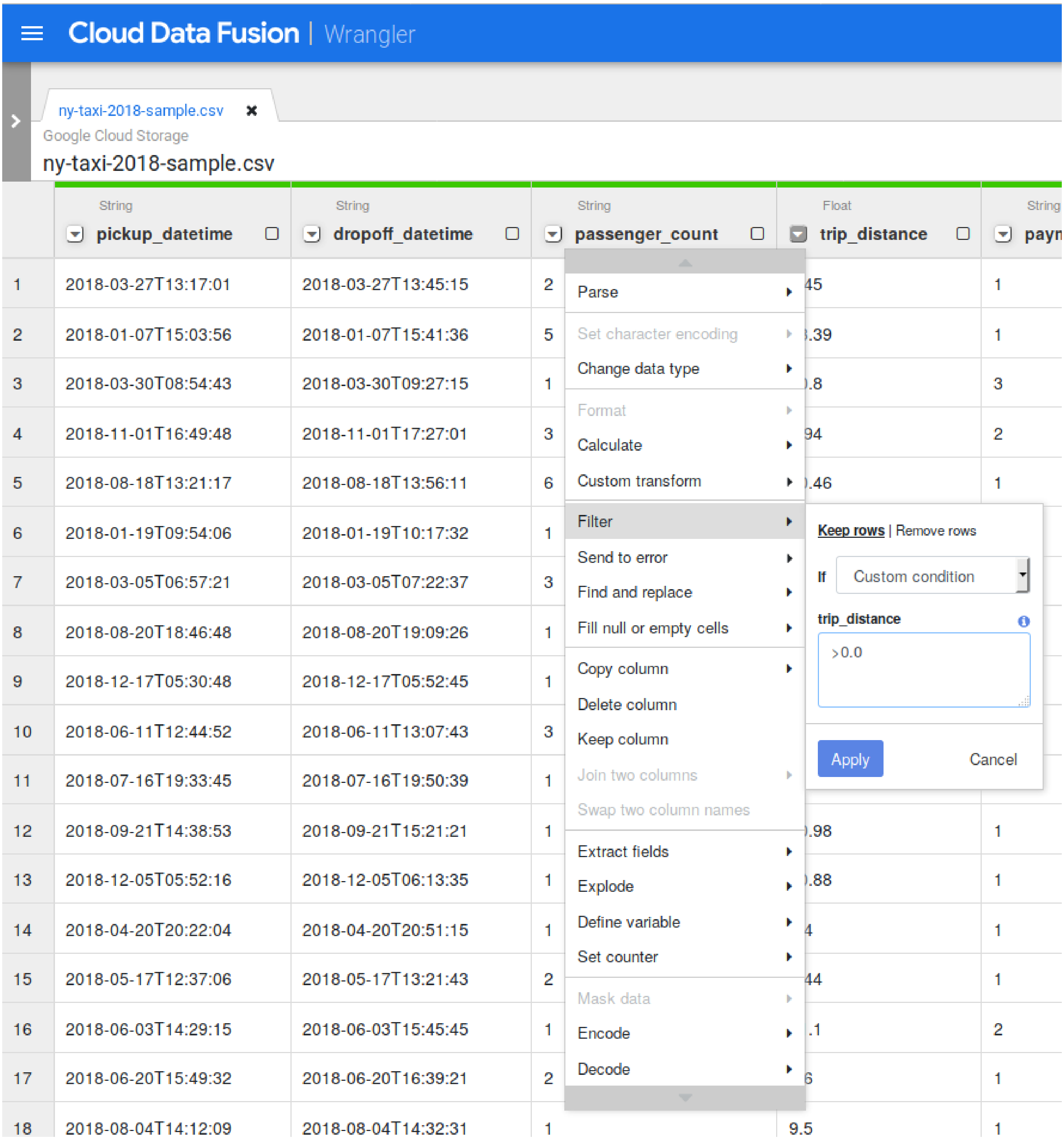
Task 3. Cleaning the data

Now, you will perform some transformations to parse and clean the taxi data.

1. Click the **Down** arrow next to the `trip_distance` column, select **Change data type** and then click on **Float**. Repeat for the `total_amount` column.

2. Click the **Down** arrow next to the `pickup_location_id` column, select **Change data type** and then click on **String**.

3. If you look at the data closely, you may find some anomalies, such as negative trip distances. You can avoid those negative values by filtering out in **Wrangler**. Click the **Down** arrow next to the `trip_distance` column and select **Filter**. Click if **Custom condition** and input `>0.0`



If: Custom condition

trip_distance: >0.0

	String pickup_datetime	String dropoff_datetime	String passenger_count	Float trip_distance	String payn
1	2018-03-27T13:17:01	2018-03-27T13:45:15	2	45	1
2	2018-01-07T15:03:56	2018-01-07T15:41:36	5	.39	1
3	2018-03-30T08:54:43	2018-03-30T09:27:15	1	.8	3
4	2018-11-01T16:49:48	2018-11-01T17:27:01	3	94	2
5	2018-08-18T13:21:17	2018-08-18T13:56:11	6	.46	1
6	2018-01-19T09:54:06	2018-01-19T10:17:32	1		
7	2018-03-05T06:57:21	2018-03-05T07:22:37	3		
8	2018-08-20T18:46:48	2018-08-20T19:09:26	1		
9	2018-12-17T05:30:48	2018-12-17T05:52:45	1		
10	2018-06-11T12:44:52	2018-06-11T13:07:43	3		
11	2018-07-16T19:33:45	2018-07-16T19:50:39	1		
12	2018-09-21T14:38:53	2018-09-21T15:21:21	1	.98	1
13	2018-12-05T05:52:16	2018-12-05T06:13:35	1	.88	1
14	2018-04-20T20:22:04	2018-04-20T20:51:15	1	4	1
15	2018-05-17T12:37:06	2018-05-17T13:21:43	2	44	1
16	2018-06-03T14:29:15	2018-06-03T15:45:45	1	.1	2
17	2018-06-20T15:49:32	2018-06-20T16:39:21	2	6	1
18	2018-08-04T14:12:09	2018-08-04T14:32:31	1	9.5	1

4. Click on **Apply**.

Task 4. Creating the pipeline

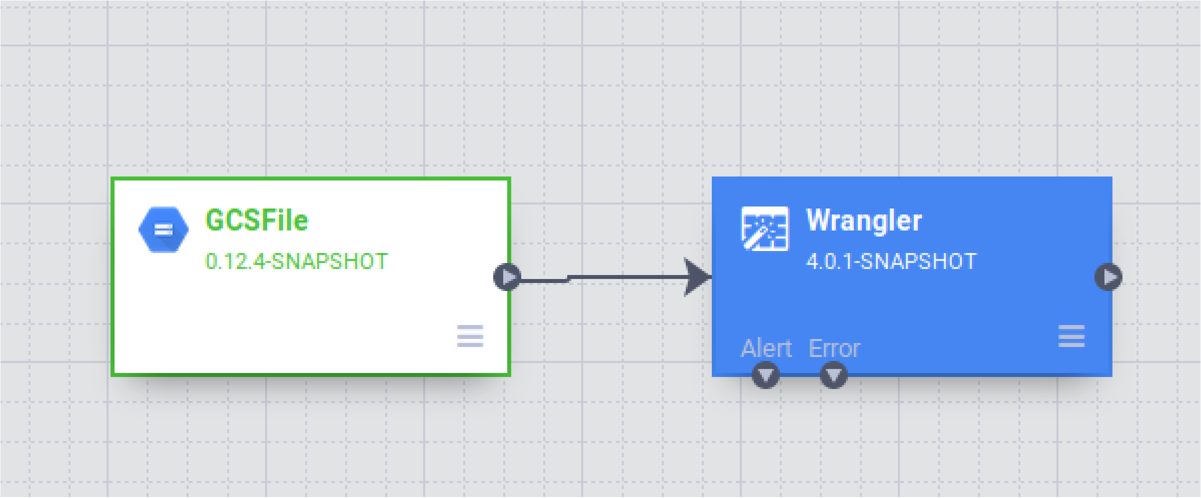
Basic data cleansing is now complete and you've run transformations on a subset of your data. You can now create a batch pipeline to run transformations on all your data.

Cloud Data Fusion translates your visually built pipeline into an Apache Spark or MapReduce program that executes transformations on an ephemeral Cloud Dataproc cluster in parallel. This enables you to easily execute complex transformations over vast quantities of data in a scalable, reliable manner, without having to wrestle with infrastructure and technology.

1. On the upper-right side of the Google Cloud Fusion UI, click **Create a Pipeline**.

2. In the dialog that appears, select **Batch pipeline**.

3. In the Data Pipelines UI, you will see a GCSFile source node connected to a Wrangler node. The Wrangler node contains all the transformations you applied in the Wrangler view captured as directive grammar. Hover over the Wrangler node and select **Properties**.



.

4. At this stage, you can apply more transformations by clicking the **Wrangle** button. Delete the `extra` column by pressing the red trashcan icon beside its name. Click **Validate** on the top right corner to check for any errors. To close the Wrangler tool click the **X** button in the top right corner.

Task 5. Adding a data source

The taxi data contains several cryptic columns such as `pickup_location_id`, that aren't immediately transparent to an analyst. You are going to add a data source to the pipeline that maps the `pickup_location_id` column to a relevant location name. The mapping information will be stored in a BigQuery table.

1. In a separate tab, open the BigQuery UI in the Cloud Console. Click **Done** on the 'Welcome to BigQuery in the Cloud Console' launch page.

2. In the Explorer section of the BigQuery UI, click the three dots beside your GCP Project ID (it will start with qwiklabs).

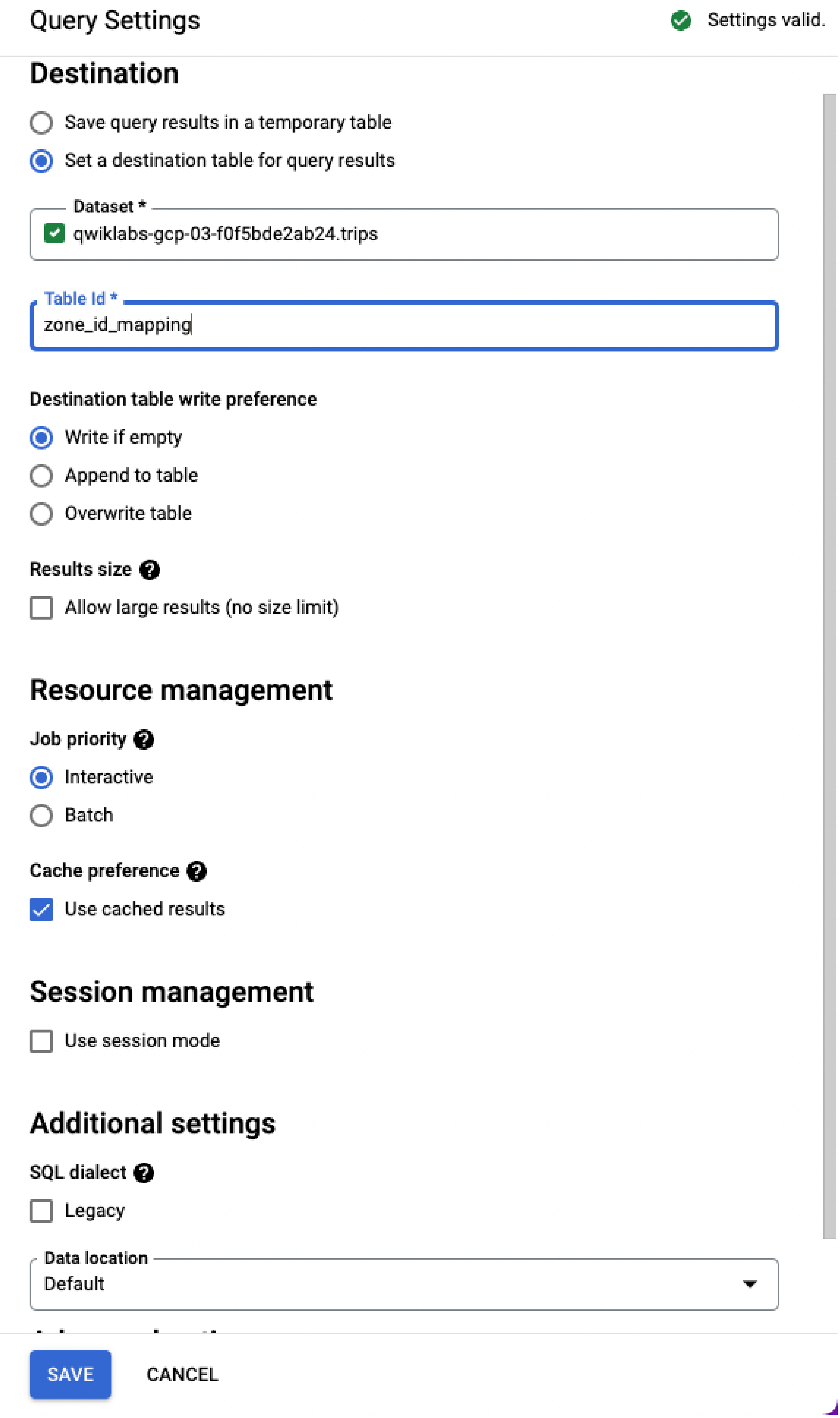
3. On the menu that appears click on **Create dataset**.

4. In the **Dataset ID** field type in `trips`.

5. Click on **Create dataset.**

6. To create the desired table in the newly created dataset, navigate to **More > Query Settings**. This process will ensure you can access your table from Cloud Data Fusion.

7. Select the item for **Set a destination table for query results**. For **Dataset** input `trips` and select from the dropdown. For **Table Id** input `zone_id_mapping`. Click **Save**.



Query Settings ✓ Settings valid.

Destination

○ Save query results in a temporary table

● Set a destination table for query results

Dataset *
☑ qwiklabs-gcp-03-f0f5bde2ab24.trips

Table Id *
zone_id_mapping

Destination table write preference

● Write if empty

○ Append to table

○ Overwrite table

Results size

□ Allow large results (no size limit)

Resource management

Job priority

● Interactive

○ Batch

Cache preference

☑ Use cached results

Session management

□ Use session mode

Additional settings

SQL dialect

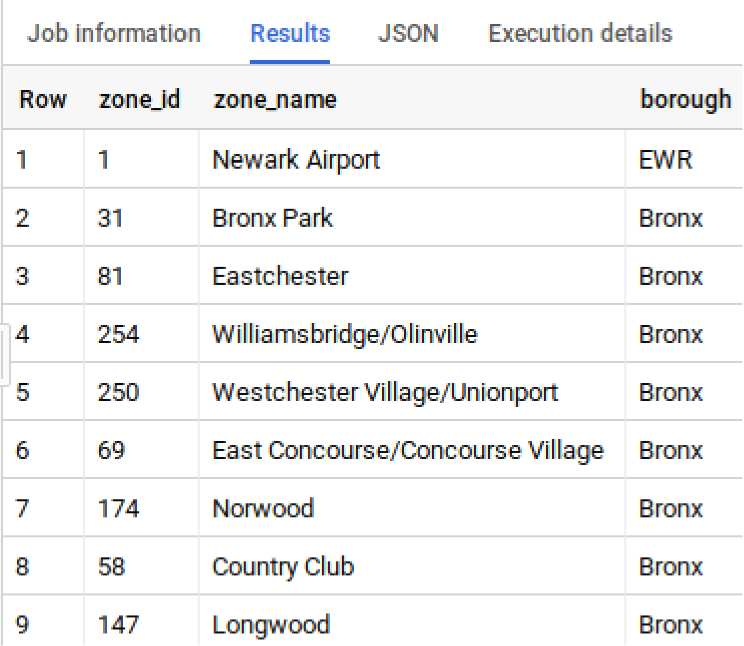
□ Legacy

Data location
Default

8. Enter the following query in the Query Editor and then click **Run**:

SELECT zone_id, zone_name, borough FROM `bigquery-public-data.new_york_taxi_trips.taxi_zone_geom`

You can see that this table contains the mapping from `zone_id` to its name and borough.

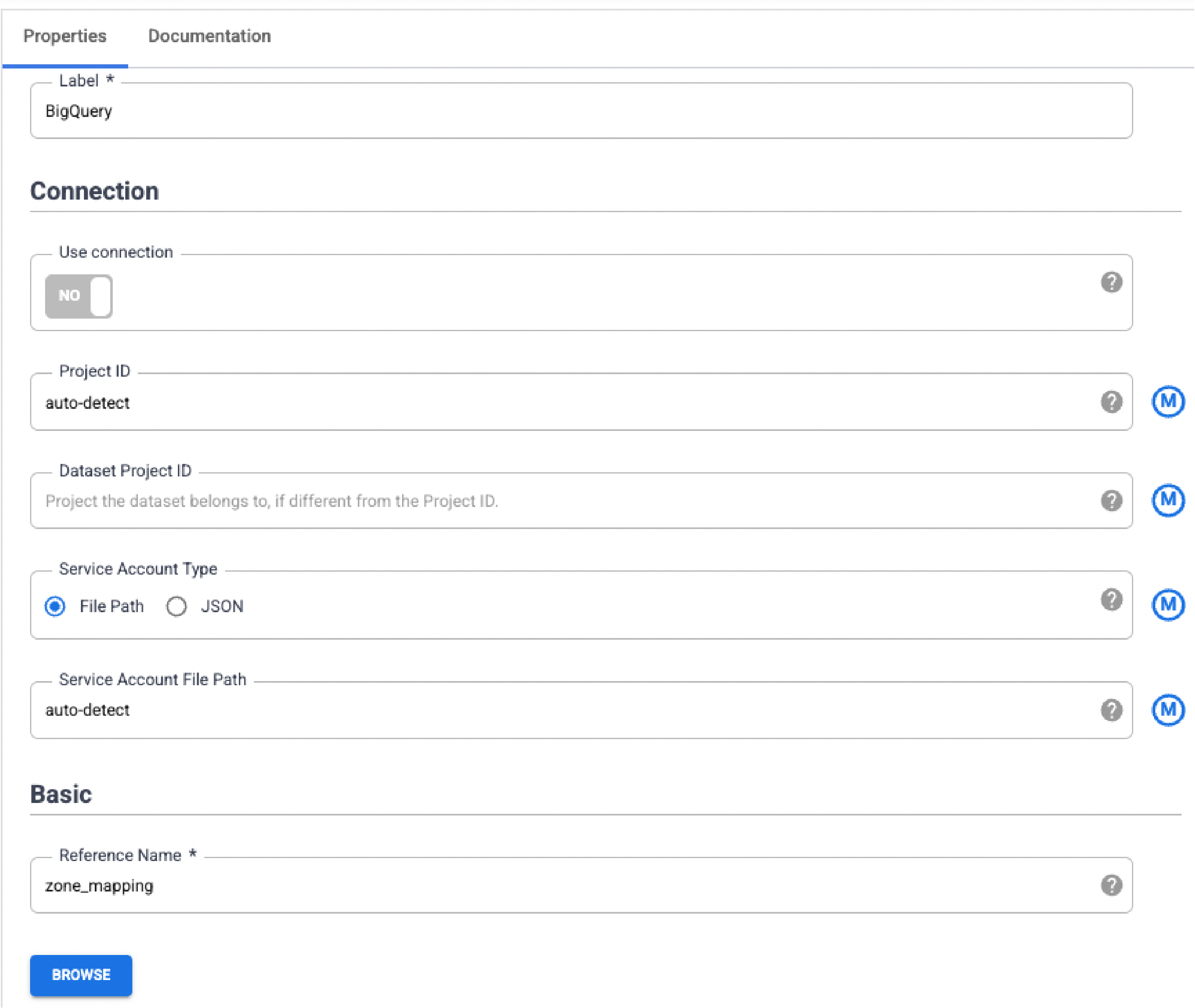


Row	zone_id	zone_name	borough
1	1	Newark Airport	EWR
2	31	Bronx Park	Bronx
3	81	Eastchester	Bronx
4	254	Williamsbridge/Olinville	Bronx
5	250	Westchester Village/Unionport	Bronx
6	69	East Concourse/Concourse Village	Bronx
7	174	Norwood	Bronx
8	58	Country Club	Bronx
9	147	Longwood	Bronx

9. Now, you will add a source in your pipeline to access this BigQuery table. Return to the tab where you have Cloud Data Fusion open, from the Plugin palette on the left, select **BigQuery** from the **Source** section. A BigQuery source node appears on the canvas with the two other nodes.

10. Hover over the new BigQuery source node and click **Properties**.

11. To configure the **Reference Name**, enter `zone_mapping`, which is used to identify this data source for lineage purposes.



Label * BigQuery

Connection

Use connection NO

Project ID auto-detect

Dataset Project ID Project the dataset belongs to, if different from the Project ID.

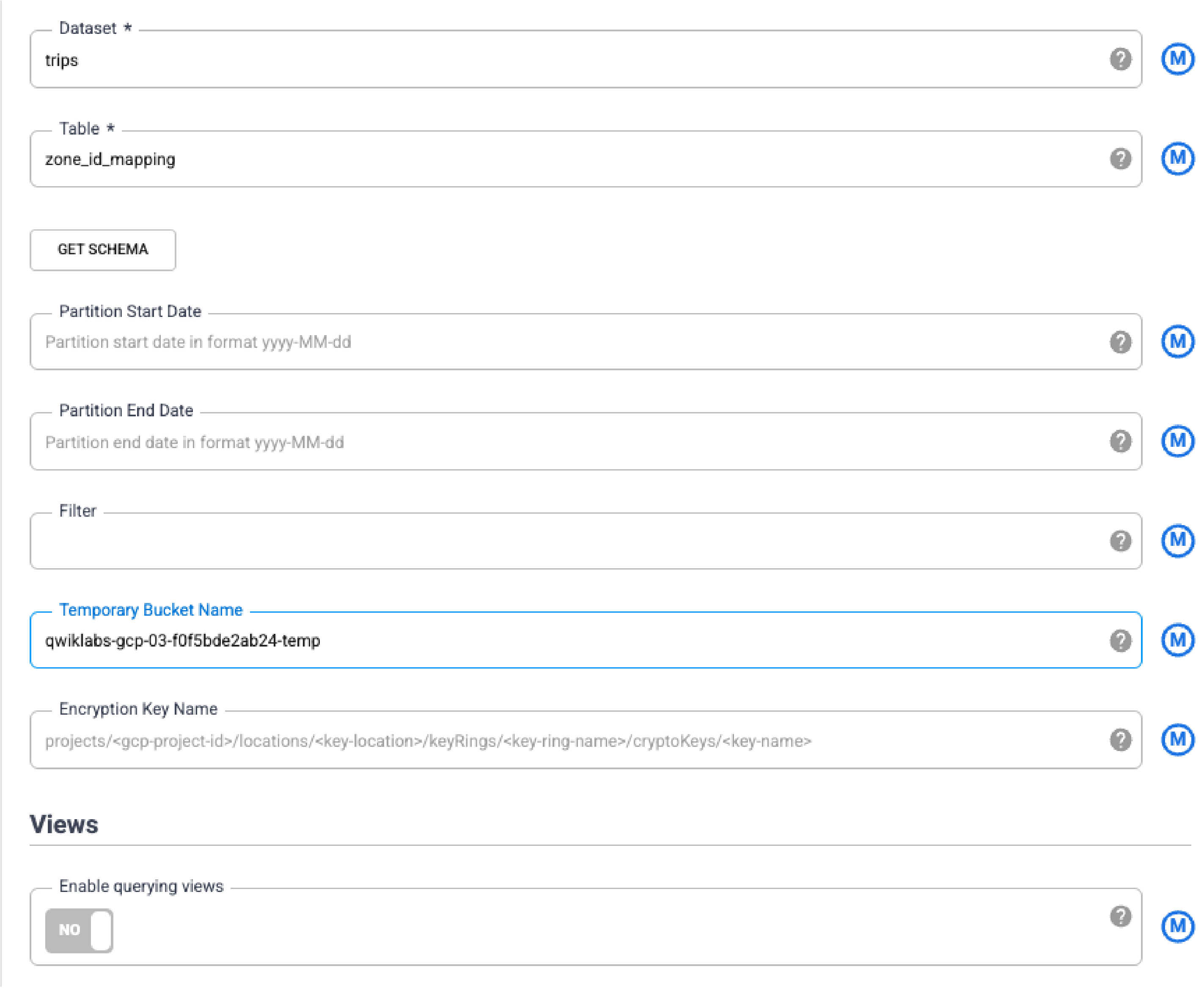
Service Account Type ● File Path ○ JSON

Service Account File Path auto-detect

Basic

Reference Name * zone_mapping

12. The BigQuery **Dataset** and **Table** configurations are the Dataset and Table you setup in BigQuery a few steps earlier: `trips` and `zone_id_mapping`. For **Temporary Bucket Name** input the name of your project followed by "-temp", which corresponds to the bucket you created in Task 2.



Dataset * trips

Table * zone_id_mapping

Partition Start Date: Partition start date in format yyyy-MM-dd

Partition End Date: Partition end date in format yyyy-MM-dd

Filter

Temporary Bucket Name: qwiklabs-gcp-03-f0f5bde2ab24-temp

Encryption Key Name: projects/<gcp-project-id>/locations/<key-location>/keyRings/<key-ring-name>/cryptoKeys/<key-name>

Views

Enable querying views: NO

13. To populate the schema of this table from BigQuery, click **Get Schema**. The fields will appear on the right side of the wizard.

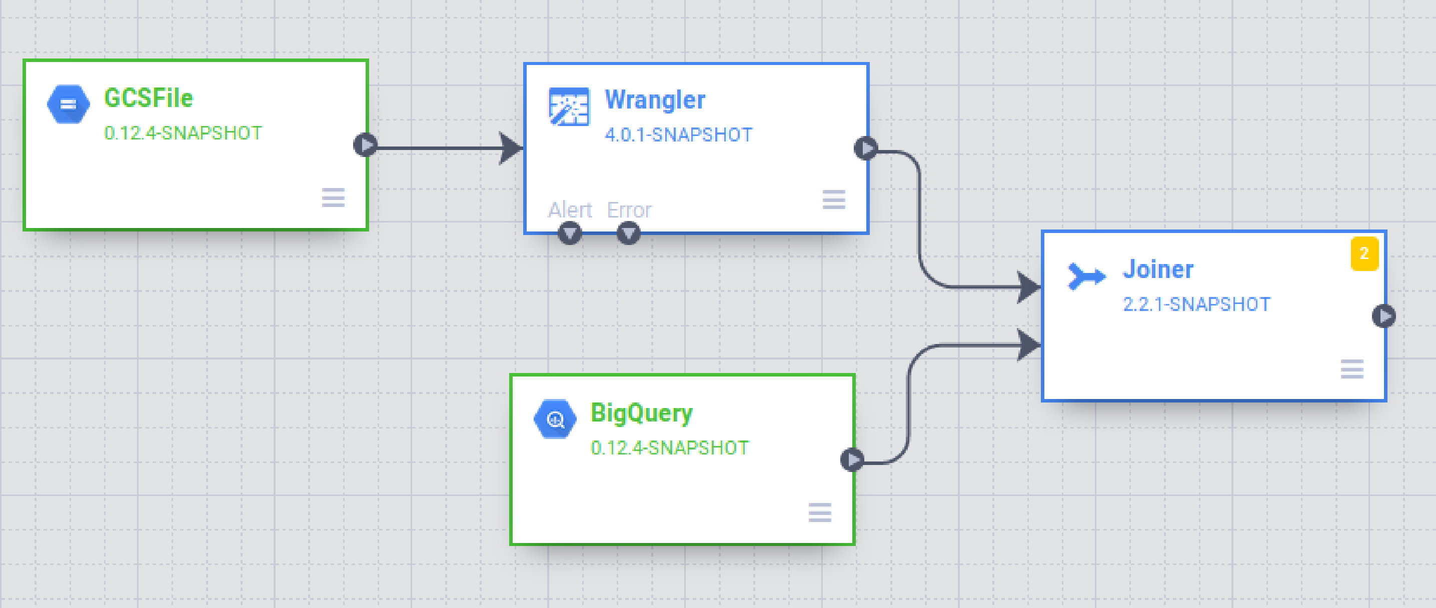
14. Click **Validate** on the top right corner to check for any errors. To close the BigQuery Properties window click the **X** button in the top right corner.

Task 6. Joining two sources

Now you can join the two data sources—taxi trip data and zone names—to generate more meaningful output.

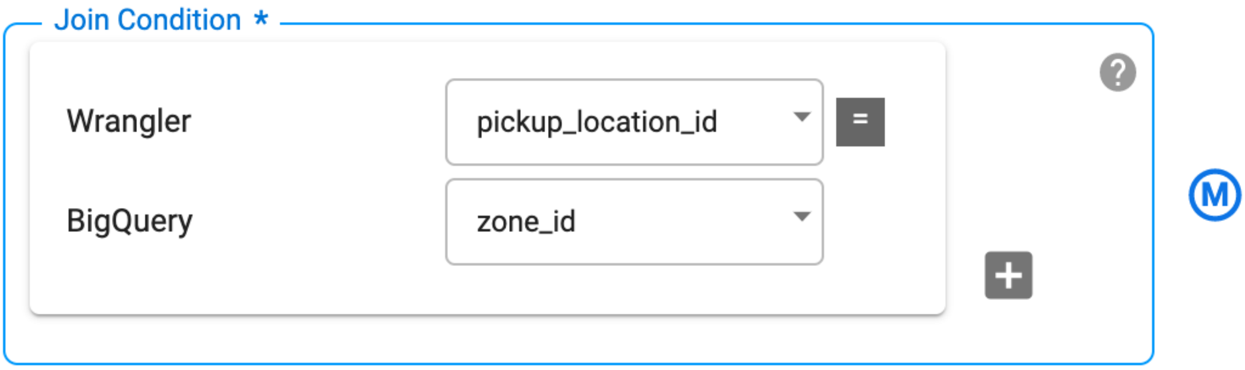
1. Under the **Analytics** section in the Plugin Palette, choose **Joiner**. A **Joiner** node appears on the canvas.

2. To connect the Wrangler node and the BigQuery node to the Joiner node: Drag a connection arrow **>** on the right edge of the source node and drop on the destination node.



3. To configure the **Joiner** node, which is similar to a SQL JOIN syntax:

- Click **Properties** of **Joiner**.
- Leave the label as **Joiner**.
- Change the **Join Type** to **Inner**
- Set the **Join Condition** to join the `pickup_location_id` column in the Wrangler node to the `zone_id` column in the BigQuery node.

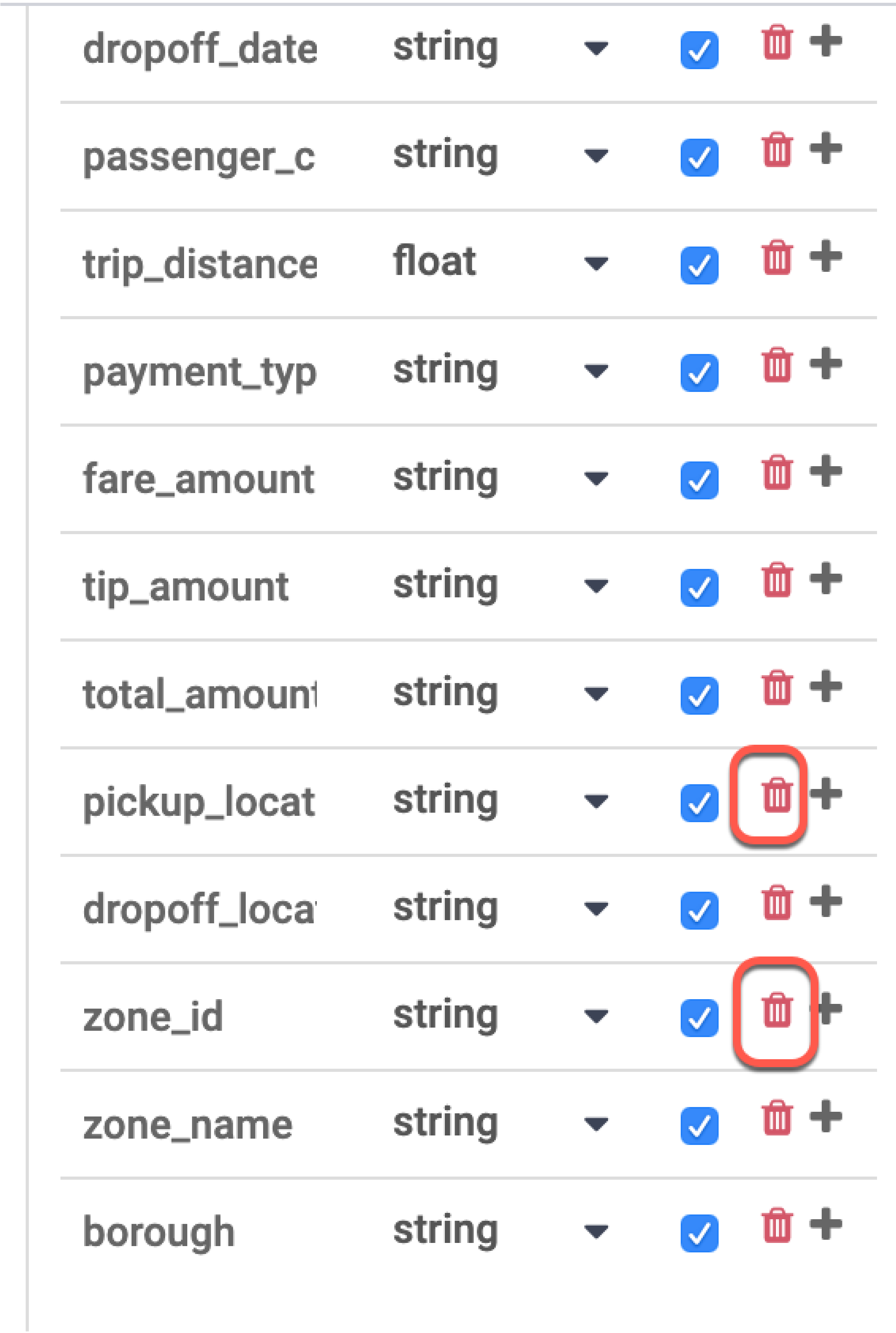


Join Condition *

Wrangler pickup_location_id =

BigQuery zone_id

- To generate the schema of the resultant join, click **Get Schema**.
- In the **Output Schema** table on the right, remove the `zone_id` and `pickup_location_id` fields by hitting the red garbage can icon.



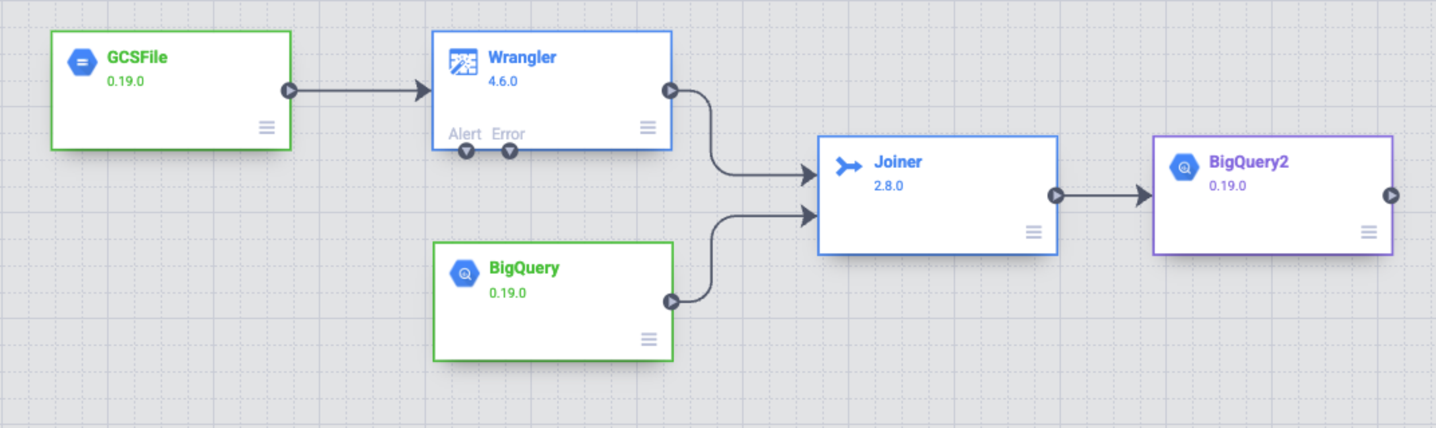
dropoff_date	string	☑
passenger_c	string	☑
trip_distance	float	☑
payment_typ	string	☑
fare_amount	string	☑
tip_amount	string	☑
total_amount	string	☑
pickup_locat	string	☑
dropoff_loca	string	☑
zone_id	string	☑
zone_name	string	☑
borough	string	☑

- Click **Validate** on the top right corner to check for any errors. Close the window by clicking the **X** button in the top right corner.

Task 7. Storing the output to BigQuery

You will store the result of the pipeline into a BigQuery table. Where you store your data is called a sink.

1. In the **Sink** section of the Plugin Palette, choose **BigQuery**.
2. Connect the **Joiner** node to the **BigQuery** node. Drag a connection arrow > on the right edge of the source node and drop on the destination node.

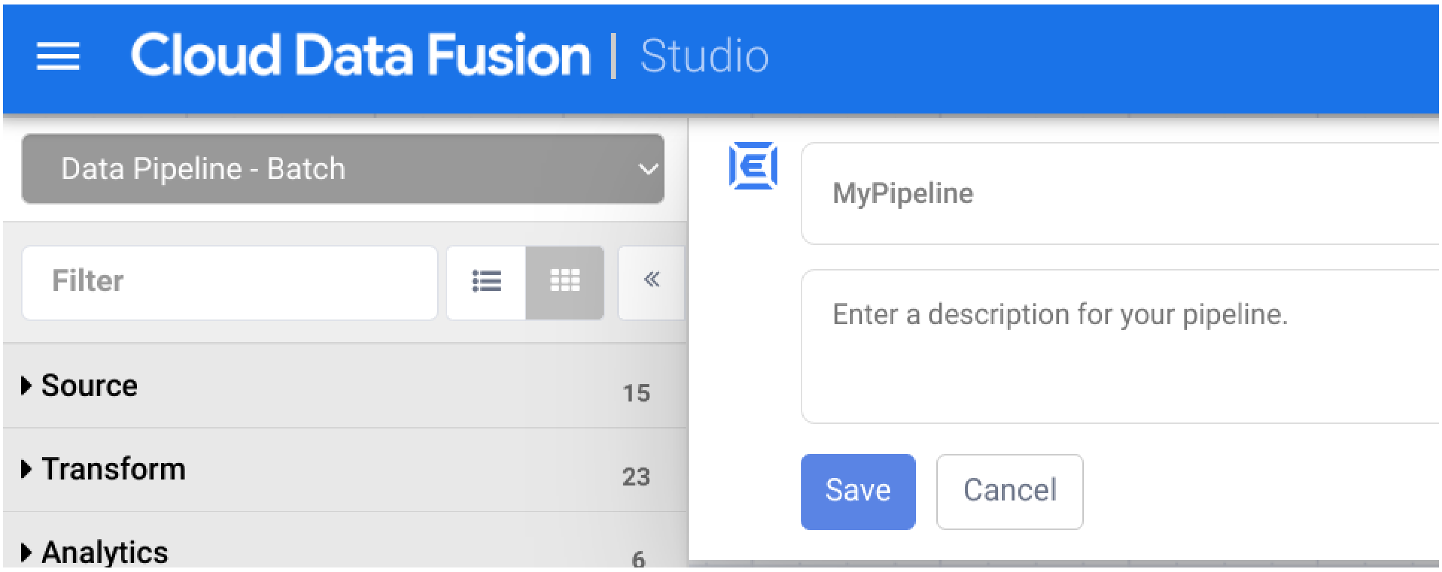


3. Open the **BigQuery2** node by hovering on it and then clicking **Properties**. You will next configure the node as shown below. You will use a configuration that's similar to the existing BigQuery source. Provide `bq_insert` for the **Reference Name** field and then use `trips` for the **Dataset** and the name of your project followed by "-temp" as **Temporary Bucket Name**. You will write to a new table that will be created for this pipeline execution. In **Table** field, enter `trips_pickup_name`.
4. Click **Validate** on the top right corner to check for any errors. Close the window by clicking the **X** button in the top right corner.

Task 8. Deploying and running the pipeline

At this point you have created your first pipeline and can deploy and run the pipeline.

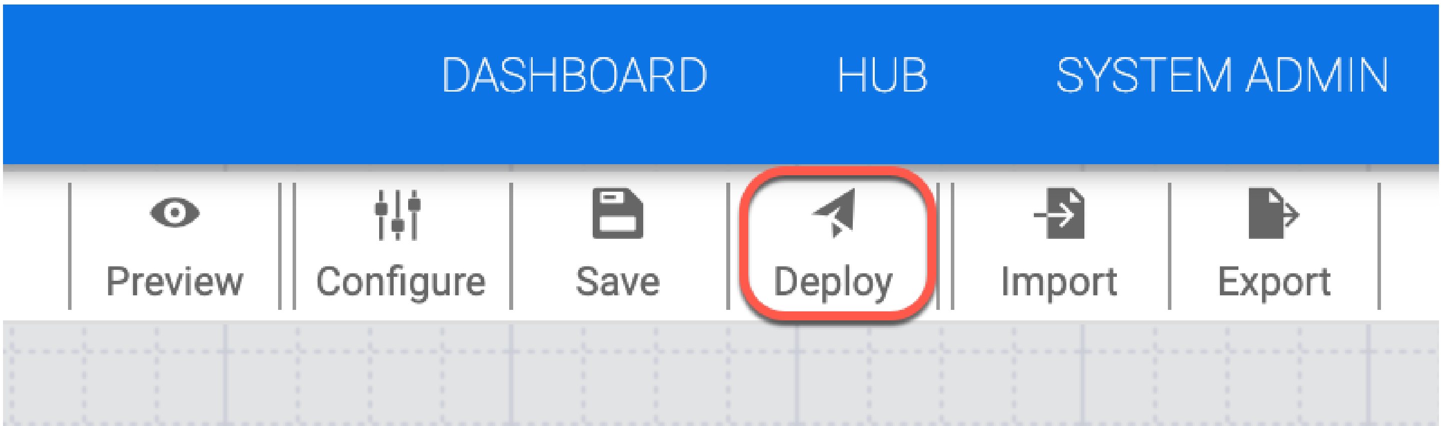
1. Name your pipeline in the upper left corner of the Data Fusion UI and click **Save**.



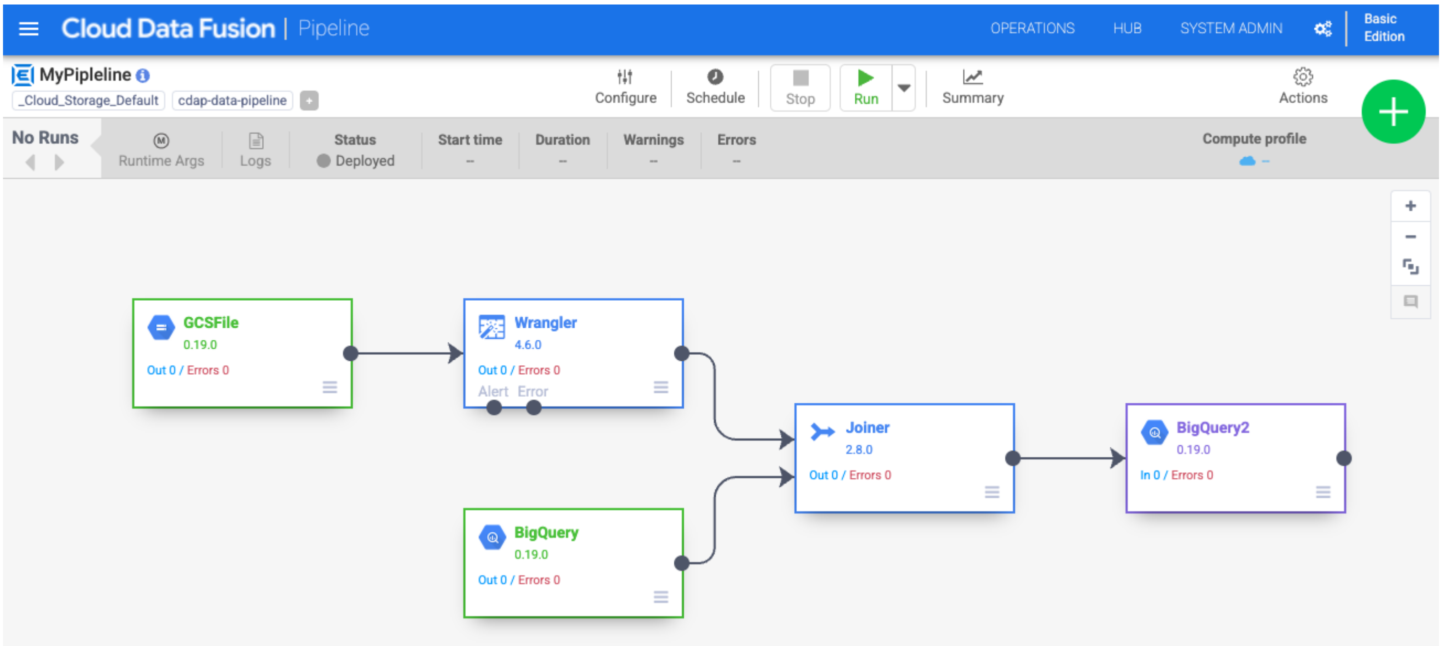
MyPipeline

Enter a description for your pipeline.

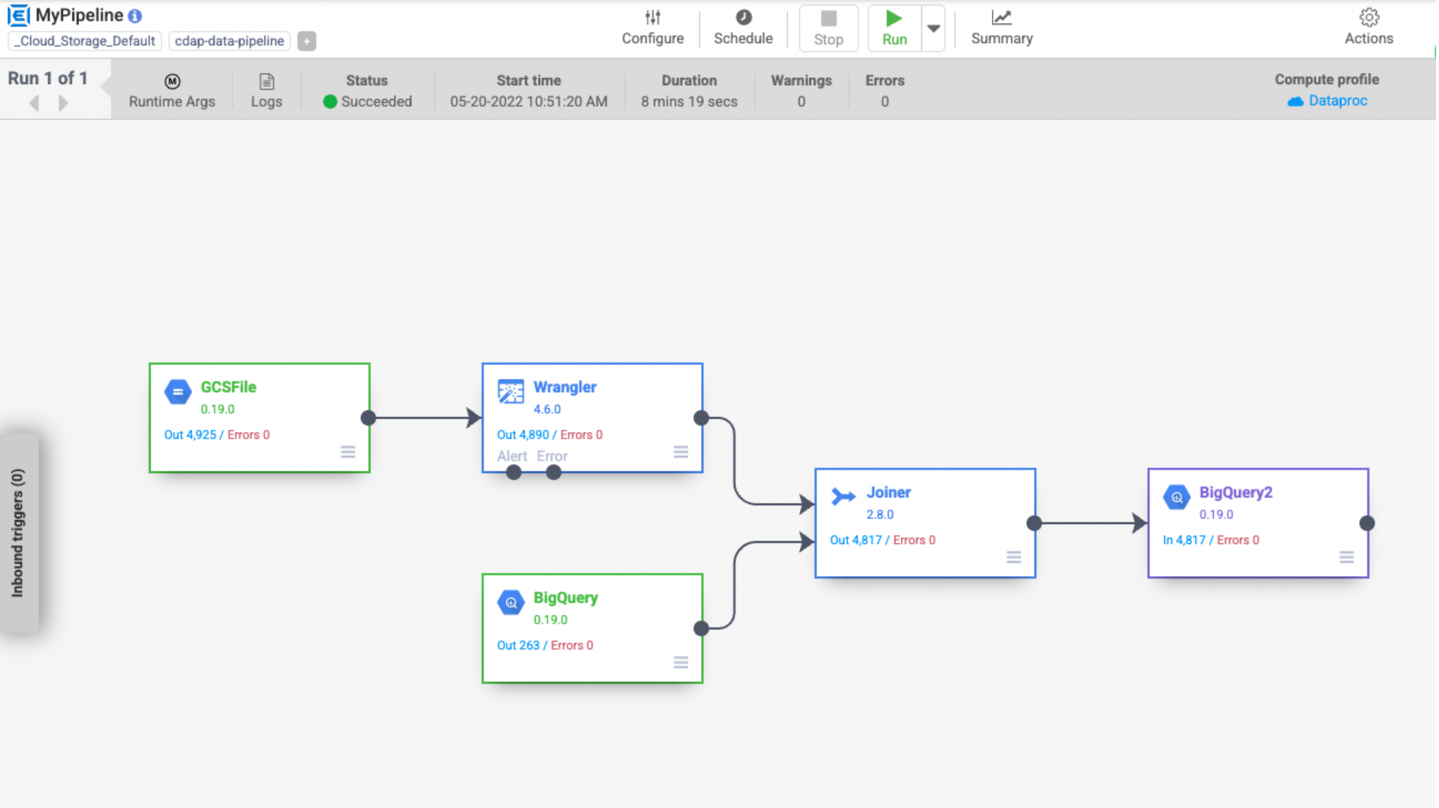
2. Now you will deploy the pipeline. In the upper-right corner of the page, click **Deploy**.



3. On the next screen click **Run** to start processing data.



When you run a pipeline, Cloud Data Fusion provisions an ephemeral Cloud Dataproc cluster, runs the pipeline, and then tears down the cluster. This could take a few minutes. You can observe the status of the pipeline transition from *Provisioning* to *Starting* and from *Starting* to *Running* to *Succeeded* during this time.



Status: Succeeded
Start time: 05-20-2022 10:51:20 AM
Duration: 8 mins 19 secs
Warnings: 0
Errors: 0
Compute profile: Dataproc

Note: The pipeline transition may take 10-15 minutes to succeeded.

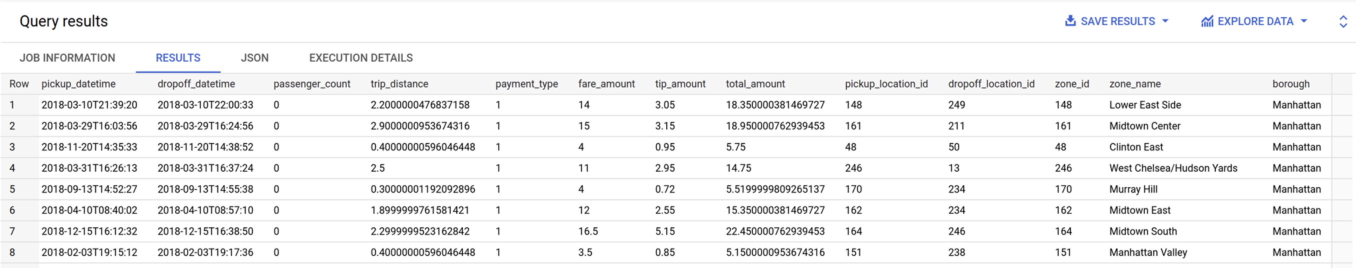
Task 9. Viewing the results

To view the results after the pipeline runs:

- Return to the tab where you have BigQuery open. Run the query below to see the values in the `trips_pickup_name` table:

  SELECT * FROM `trips.trips_pickup_name`

  BQ RESULTS



Row	pickup_datetime	dropoff_datetime	passenger_count	trip_distance	payment_type	fare_amount	tip_amount	total_amount	pickup_location_id	dropoff_location_id	zone_id	zone_name	borough
1	2018-03-10T21:39:20	2018-03-10T22:00:33	0	2.2000000476837158	1	14	3.05	18.350000381469727	148	249	148	Lower East Side	Manhattan
2	2018-03-29T16:03:56	2018-03-29T16:24:56	0	2.9000000953674316	1	15	3.15	18.950000762939453	161	211	161	Midtown Center	Manhattan
3	2018-11-20T14:35:33	2018-11-20T14:38:52	0	0.40000000596046448	1	4	0.95	5.75	48	50	48	Clinton East	Manhattan
4	2018-03-31T16:26:13	2018-03-31T16:37:24	0	2.5	1	11	2.95	14.75	246	13	246	West Chelsea/Hudson Yards	Manhattan
5	2018-09-13T14:52:27	2018-09-13T14:55:38	0	0.30000001192092896	1	4	0.72	5.5199999809265137	170	234	170	Murray Hill	Manhattan
6	2018-04-10T08:40:02	2018-04-10T08:57:10	0	1.8999999761581421	1	12	2.55	15.350000381469727	162	234	162	Midtown East	Manhattan
7	2018-12-15T16:12:32	2018-12-15T16:38:50	0	2.2999999523162842	1	16.5	5.15	22.450000762939453	164	246	164	Midtown South	Manhattan
8	2018-02-03T19:15:12	2018-02-03T19:17:36	0	0.40000000596046448	1	3.5	0.85	5.1500000953674316	151	238	151	Manhattan Valley	Manhattan

End your lab

When you have completed your lab, click **End Lab**. Google Cloud Skills Boost removes the resources you've used and cleans the account for you.

You will be given an opportunity to rate the lab experience. Select the applicable number of stars, type a comment, and then click **Submit**.

The number of stars indicates the following:

- 1 star = Very dissatisfied
- 2 stars = Dissatisfied
- 3 stars = Neutral
- 4 stars = Satisfied
- 5 stars = Very satisfied

You can close the dialog box if you don't want to provide feedback.

For feedback, suggestions, or corrections, please use the **Support** tab.