End Lab 01:43:22

Caution: When you are in the console, do not deviate from the lab instructions. Doing so may cause your account to be blocked. Learn more.

Open Google Console

Username
student-02-5a7887f0834d@

Password
6ho9kG8jtn9z

GCP Project ID
qwiklabs-gcp-02-550a80f6

Building and Executing a Pipeline Graph with Data Fusion 2.5

2 hours 30 minutes Free ★★★★½

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 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 GCP project and set of resources for a fixed time at no cost.

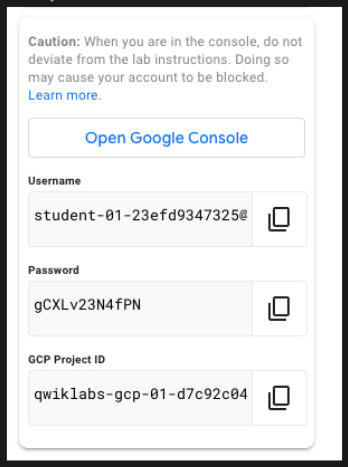
1. Make sure you signed into Qwiklabs using an **incognito window**.

2. Note the lab's access time (for example, 02:00:00 and make sure you can finish in that time block.

> 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. You will use them to sign in to Cloud Platform 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 get errors or **incur charges**.

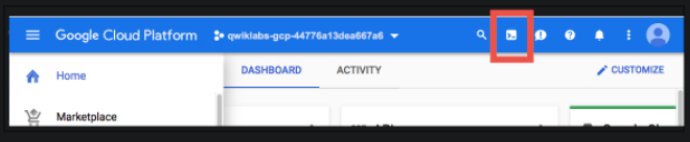
7. Accept the terms and skip the recovery resource page.

> Do not click **End Lab** unless you are finished with the lab or want to restart it. This clears your work and removes the project.

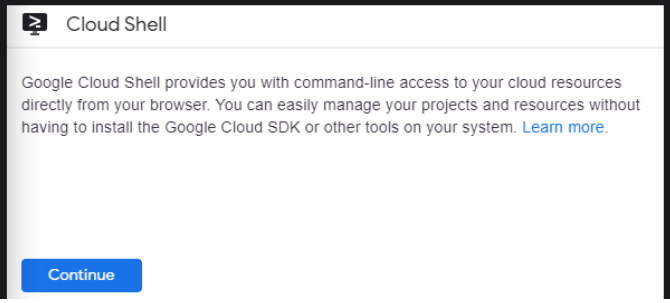
Activate Google Cloud Shell

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 GCP resources.

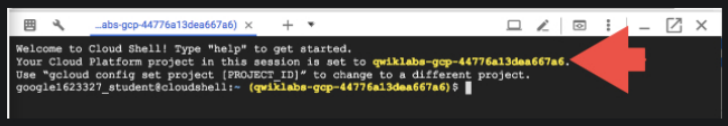
1. In GCP console, on the top right toolbar, click the Open Cloud Shell button.



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:



gcloud is the command-line tool for Google Cloud Platform. 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:
 - <myaccount>@<mydomain>.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 = <project_ID>
```

Example output:

```
[core]
project = qwiklabs-gcp-44776a13dea667a6
```

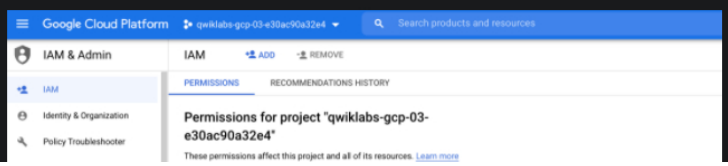
Full documentation of **gcloud** is available on Google Cloud gcloud Overview .

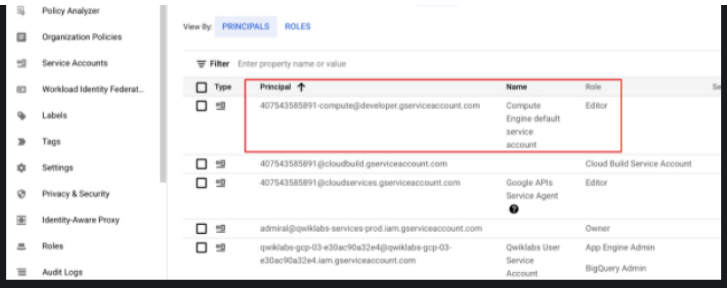
Check project permissions

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** (☰), click **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**.





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

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

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

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

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

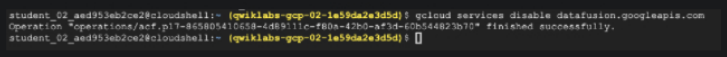
Task 1: Creating a Cloud Data Fusion instance

Thorough directions for creating a Cloud Data Fusion instance can be found here. 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
```

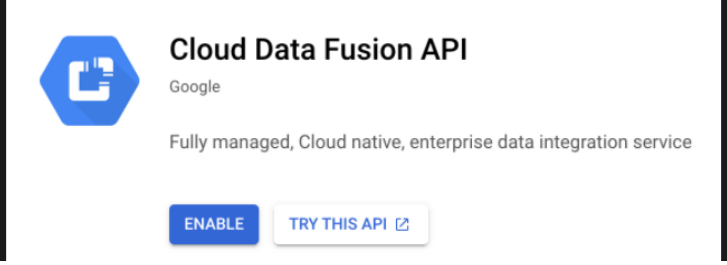
Your output will appear similar to the screenshot:



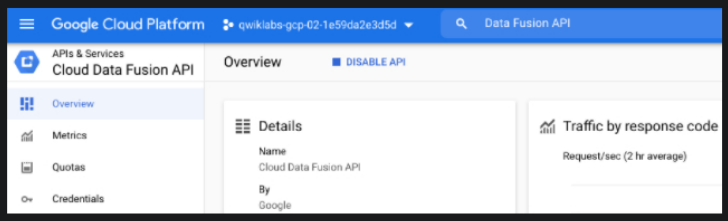
2. Next, restart the connection to the Cloud Data Fusion API.

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**.

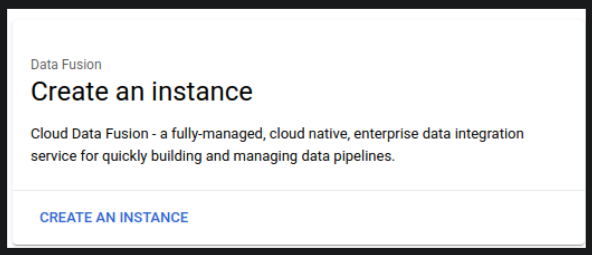


4. When the API has been enabled again, the page will refresh and show the option to disable the API along with other details on the API usage and performance.



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.

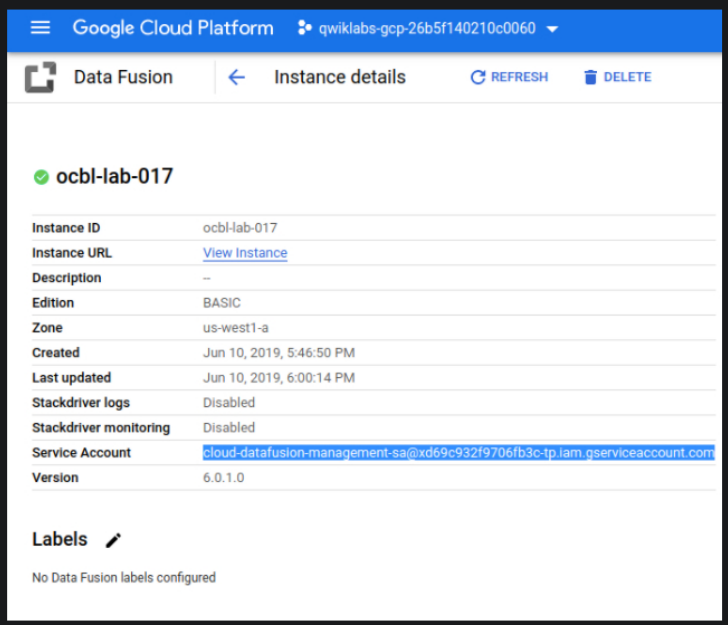
8. Select **Basic** for the Edition type.

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.

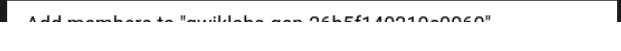
11. Once the instance is created, you need one additional step to grant the service account associated with the instance permissions on your project. Navigate to the instance details page by clicking the instance name.

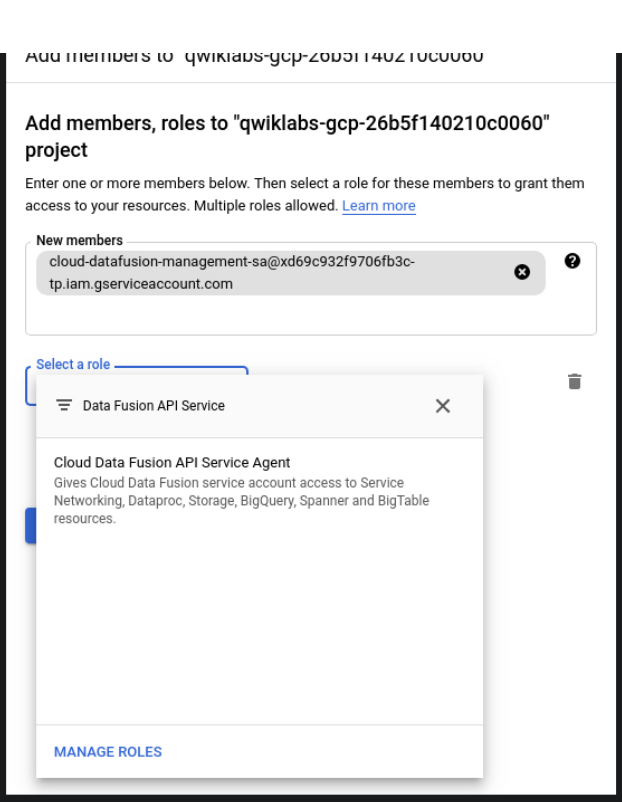


12. Copy the service account to your clipboard.

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

14. On the IAM Permissions page, add the service account you copied earlier as a new member and grant the **Cloud Data Fusion API Service Agent** role, by clicking the **Add** button.





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. Open a cloud shell console and execute the following commands to create a new bucket and copy the relevant data into it:

```
export BUCKET=$GOOGLE_CLOUD_PROJECT
gsutil mb gs://$BUCKET
gsutil 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.

```
gsutil mb gs://$BUCKET-temp
```

Note: The created bucket name is your project id followed by "-temp".

3. Click the **View Instance** link on the Cloud Data Fusion instances page, or the details page of an instance. 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.

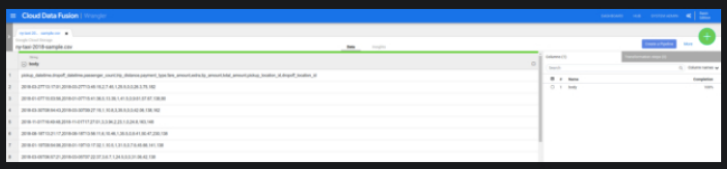
4. **Wrangler** is an interactive, visual tool that lets you see the effects of transformations on a small subset of your data before dispatching large, parallel-processing jobs on the

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 **Google Cloud Storage**, 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.



Task 3: Cleaning the data

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

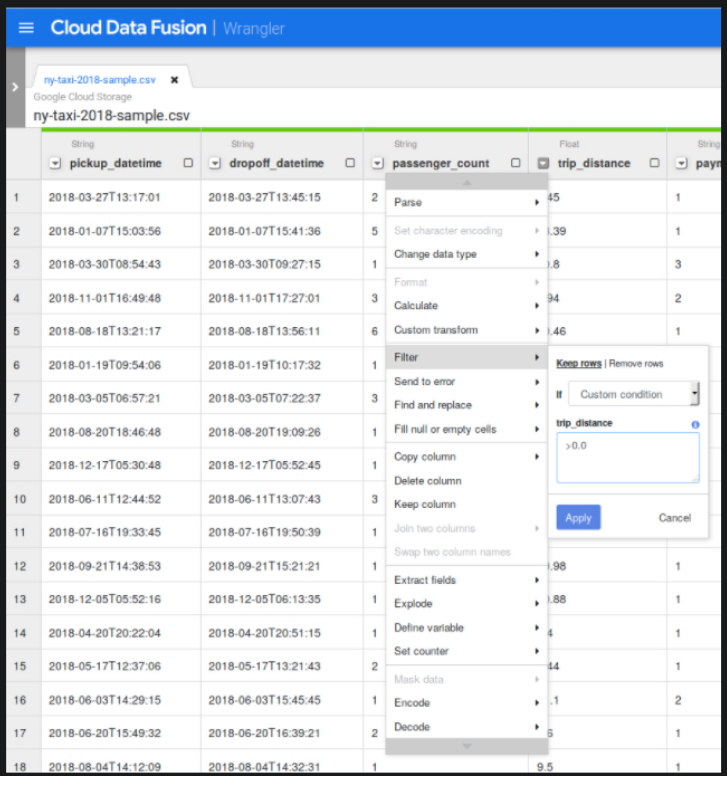
1. To the left of the `body` column, click the **Down** arrow.

2. Click **Parse > CSV**, select **Set first row as header** and then click **Apply.** The data splits into multiple columns.

3. Because the `body` column isn't needed anymore, click the **Down** arrow next to the `body` column and choose **Delete column**.

4. You'll notice that all of the column types have been loaded in as `String`. 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.

5. 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`



6. Click on **Apply**.

Task 4: Creating the pipeline

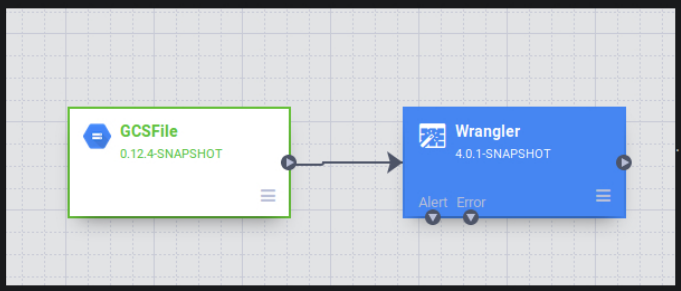
Basic data cleansing is now complete and you've run transformations on a subset of your data. You can now create a batch pipeline to run transformations on all your data.

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. To close the Wrangler tool click the **X** button in the top right corner.

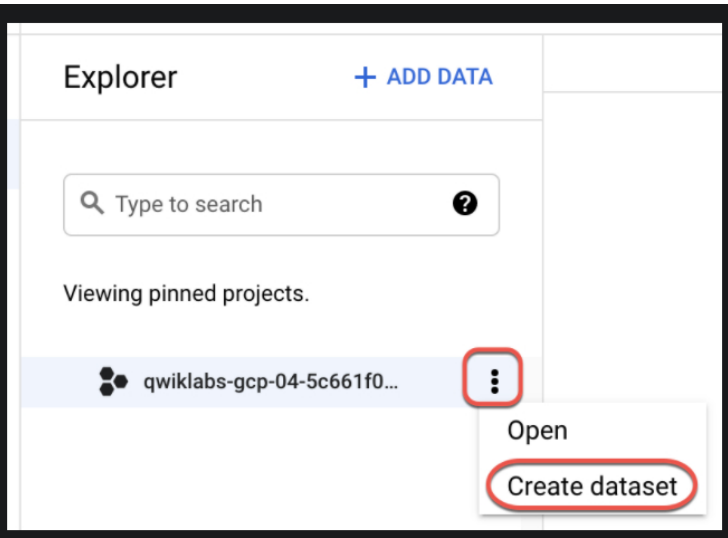
Task 5: Adding a data source

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 GCP 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).

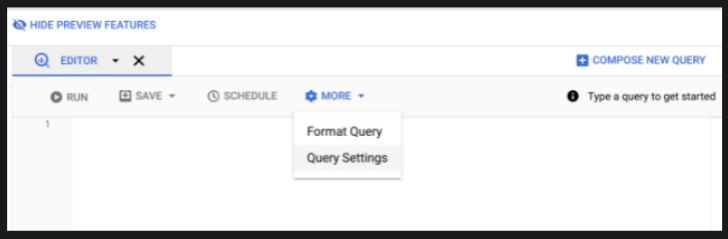
3. On the menu that appears click the **Create dataset** link.



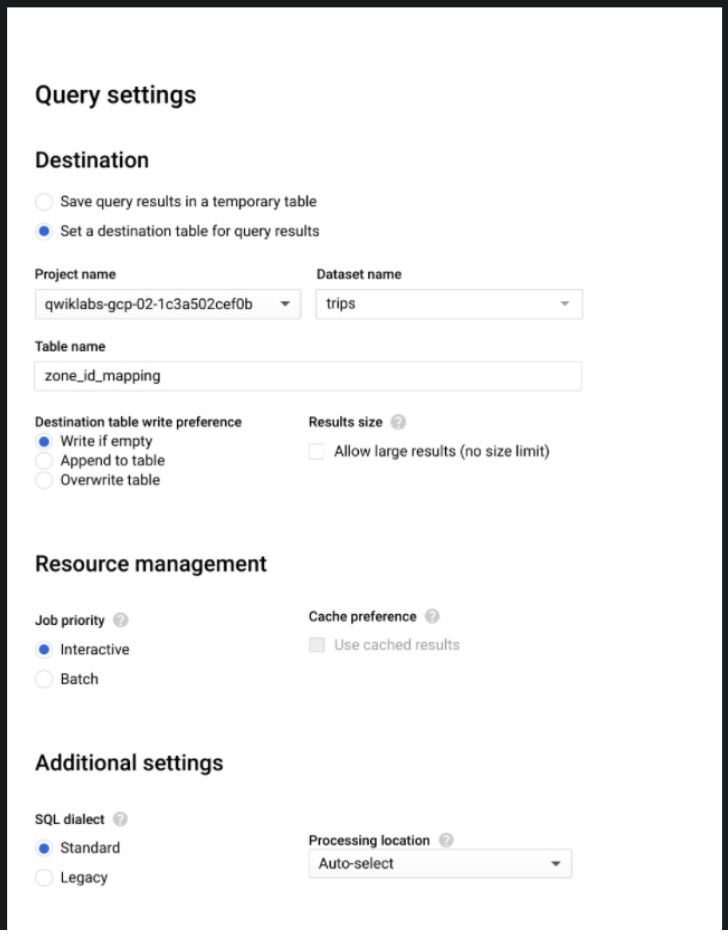
4. In the **Dataset ID** field type in `trips`.

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**. Also, under **Table name** input `zone_id_mapping`. Click **Save**.



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.

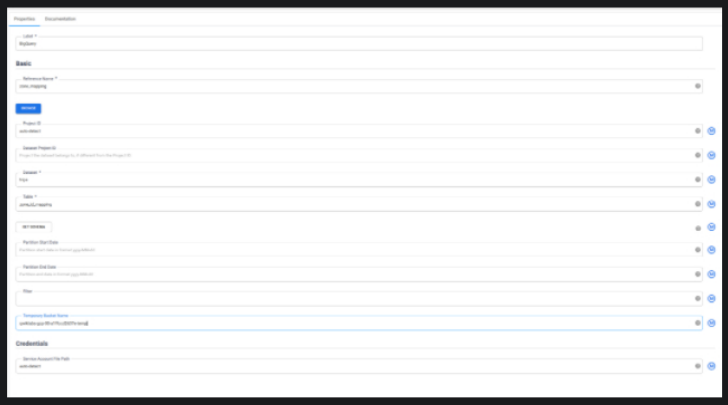
Job information | Results | JSON | Execution details

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

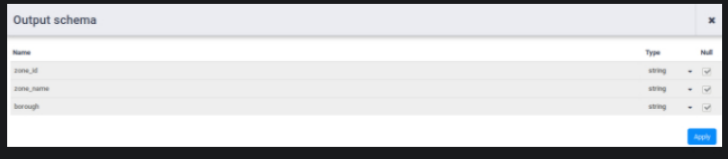
9. Now, you will add a source in your pipeline to access this BigQuery table. Return to tab where you have Cloud Data Fusion open, from the Plugin palette on the left, select **BigQuery** from the **Source** section. A BigQuery source node appears on the canvas with the two other nodes.

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. 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.



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



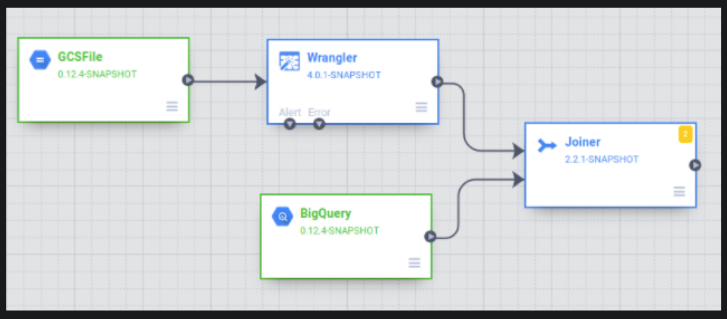
13. To close the BigQuery Properties window click the **X** button in the top right corner.

Task 6: Joining two sources

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

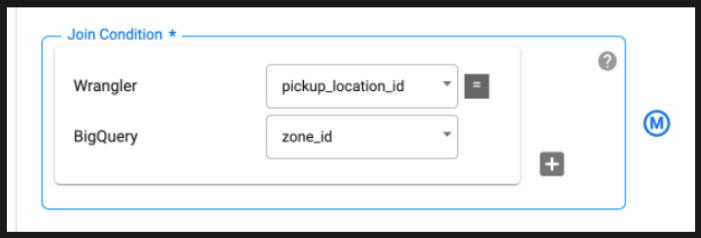
1. Under the **Analytics** section in the Plugin Palette, choose **Joiner**. A **Joiner** node appears on the canvas.

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.

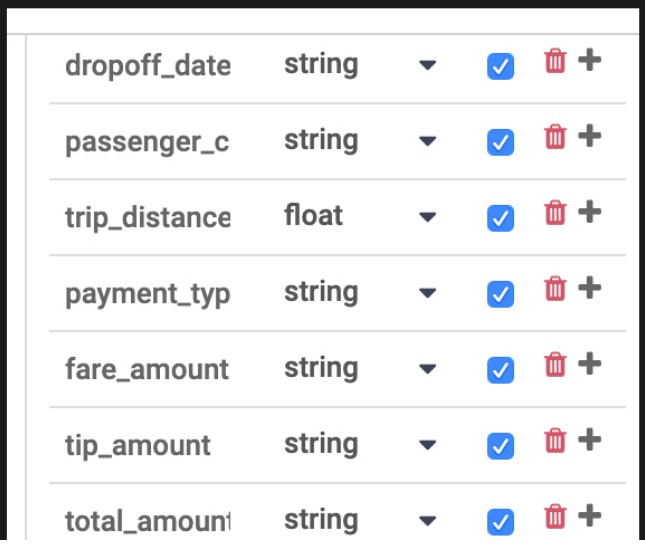


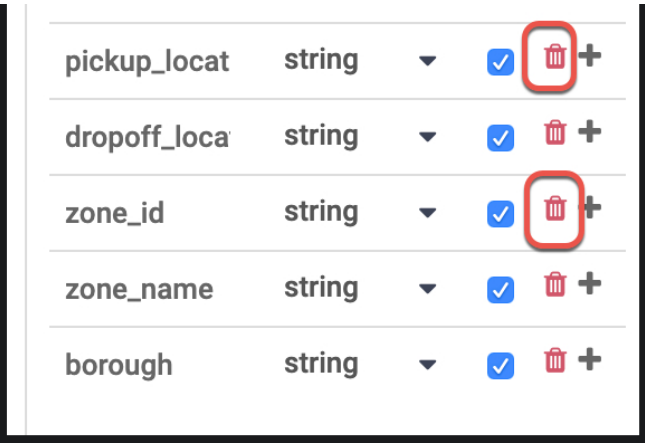
3. To configure the Joiner node, which is similar to a SQL JOIN syntax:

- 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.



- 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.





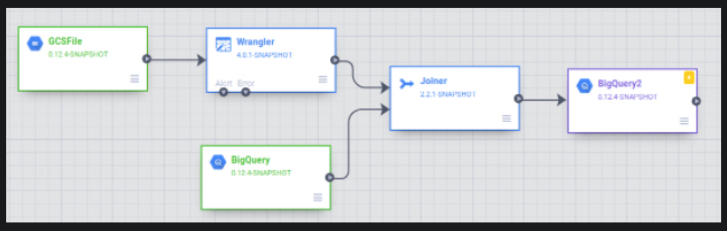
- Close the window by clicking the **X** button in the top right corner.

Task 7: Storing the output to BigQuery

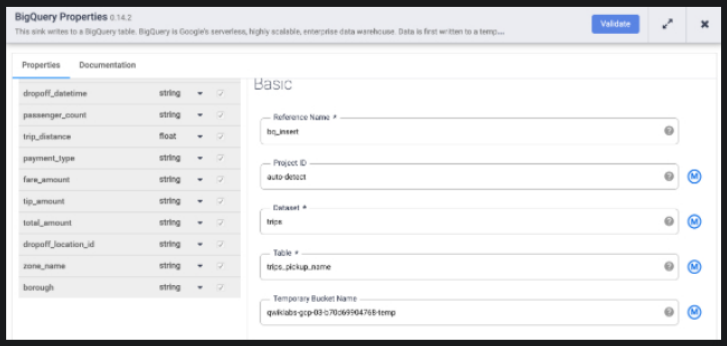
You will store the result of the pipeline into a BigQuery table. Where you store your data is called a sink.

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 BigQuery 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` .

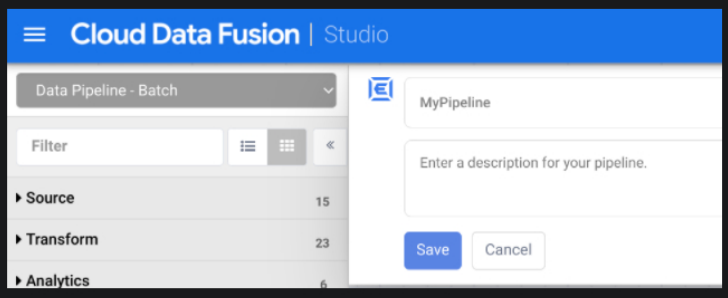


4. Close the window by clicking the **X** button in the top right corner.

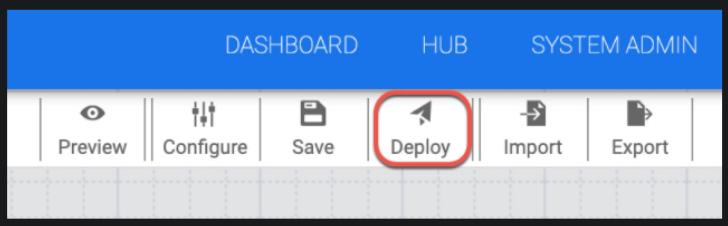
Task 8: Deploying and running the pipeline

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

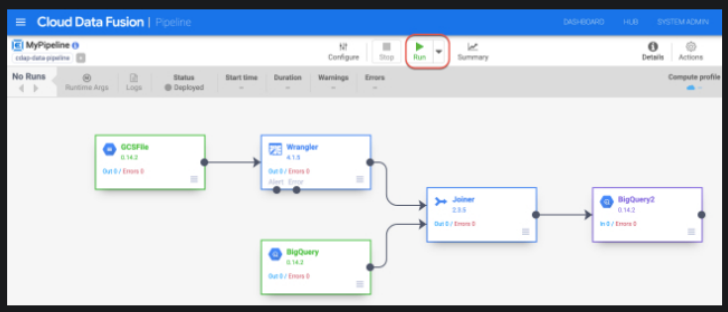
1. Name your pipeline in the upper left corner of the Data Fusion UI and click **Save**.



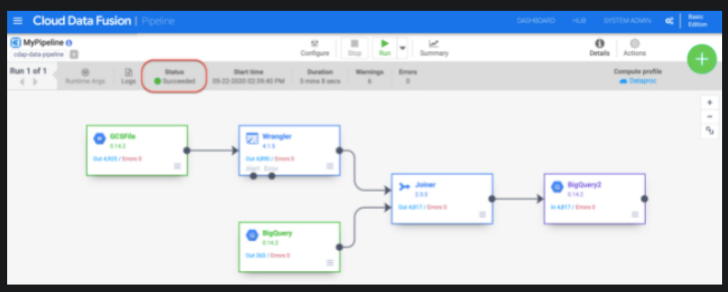
2. Now you will deploy the pipeline. In the upper-right corner of the page, click **Deploy**.



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.



Note: The pipeline can take 10-15 minutes to get succeeded.

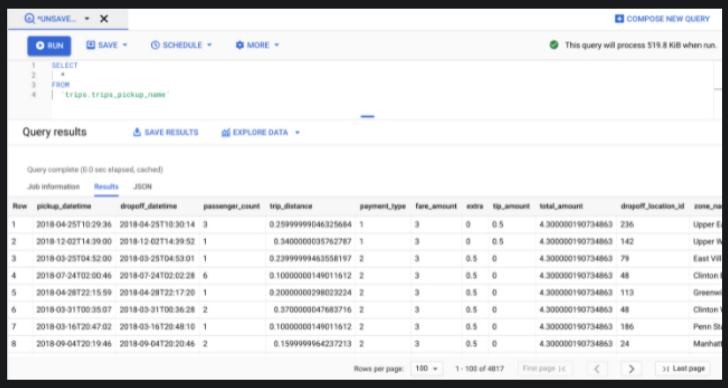
Task 9: Viewing the results

To view the results after the pipeline runs:

1. 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



End your lab

When you have completed your lab, click **End Lab**. Qwiklabs 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.

Copyright 2021 Google LLC All rights reserved. Google and the Google logo are trademarks of Google LLC. All other company and product names may be trademarks of the respective companies with which they are associated.