

Lab: Building and executing a pipeline graph in Cloud Data Fusion

Configuration(create data fusion instance)

- ✓ **Activate Google Cloud Shell**

gcloud auth list

gcloud config list project

disable datafusion

gcloud services disable datafusion.googleapis.com

- ✓ **Check project permissions**

- ✓ **Creating a Cloud Data Fusion instance**

Enable Cloud Data Fusion API

- ✓ **Data Fusion > Create an Instance.**



name: what you want





Edition type : Basic


Authorization : Grant Permission

It need 15 minutes to complete


- ✓ **Copy the service account to your clipboard**

 Google Cloud Platform 

 Data Fusion  Instance details  REFRESH  DELETE

 **ocbl-lab-017**

Instance ID	ocbl-lab-017
Instance URL	View Instance
Description	--
Edition	BASIC
Zone	us-west1-a
Created	Jun 10, 2019, 5:46:50 PM
Last updated	Jun 10, 2019, 6:00:14 PM
Stackdriver logs	Disabled
Stackdriver monitoring	Disabled
Service Account	cloud-datafusion-management-sa@xd69c932f9706fb3c-tp.iam.gserviceaccount.com
Version	6.0.1.0

Labels 

No Data Fusion labels configured

Configuration(add role)

- ✓ **IAM & Admin > IAM**
- ✓ 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.

Add members to "qwiklabs-gcp-26b5f140210c0060"

Add members, roles to "qwiklabs-gcp-26b5f140210c0060" project

Enter one or more members below. Then select a role for these members to grant them access to your resources. Multiple roles allowed. [Learn more](#)

New members

cloud-datafusion-management-sa@xd69c932f9706fb3c-tp.iam.gserviceaccount.com

Select a role

Data Fusion API Service

Cloud Data Fusion API Service Agent

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

[MANAGE ROLES](#)

Preparing>Loading the data, wrangler (clean data))

- ✓ `export BUCKET=$GOOGLE_CLOUD_PROJECT`
`gsutil mb gs://$BUCKET`
`gsutil cp gs://cloud-training/OCBL017/ny-taxi-2018-sample.csv gs://$BUCKET`
- ✓ #create a bucket for temporary storage items that Cloud data Fusion will create.
`gsutil mb gs://$BUCKET-temp`
- ✓ **View Instance** link on the Cloud Data Fusion instances page
- ✓ **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.
- ✓ Under **Google Cloud Storage**, select **Cloud Storage Default**.
- ✓ Click on the bucket corresponding to your project name.
- ✓ Select **ny-taxi-2018-sample.csv**. The data is loaded into the Wrangler screen in row/column form.

wrangler (clean data)

- ✓ **Parse > CSV**, select **Set first row as header** and then click **Apply**
- ✓ **Delete column** body
- ✓ column types have been loaded in as String
- ✓ **Change data type** trip_distance , total_amount column, **Float**.
- ✓ trip_distance column and select **Filter**. Click if **Custom condition** and input >0.0

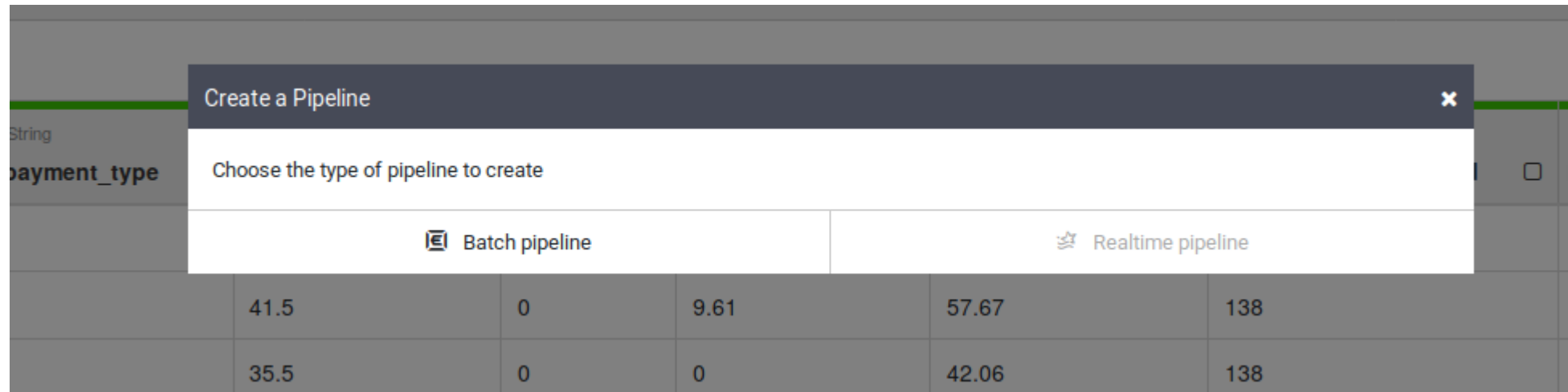
The screenshot displays the Google Cloud Data Fusion Wrangler interface. On the left, a sidebar shows the file 'ny-taxi-2018-sample.csv' selected from Google Cloud Storage. The main workspace shows a table with columns: pickup_datetime, dropoff_datetime, passenger_count, trip_distance, payment_type, fare_amount, extra, tip_amount, total_amount, pickup_location_id, and dropoff_location_id. All columns are currently of type 'String'. A context menu is open over the 'trip_distance' column, showing options like 'Parse', 'Set character encoding', 'Change data type', 'Format', 'Calculate', 'Custom transform', 'Filter', 'Send to error', 'Find and replace', 'Fill null or empty cells', 'Copy column', 'Delete column', 'Keep column', 'Join two columns', 'Swap two column names', 'Extract fields', 'Explode', 'Define variable', 'Set counter', 'Mask data', 'Encode', and 'Decode'. The 'Filter' option is selected, and a sub-menu is open showing 'Keep rows | Remove rows' with a 'Custom condition' input field containing 'trip_distance > 0.0'. The 'Apply' button is visible at the bottom of the filter menu.

	String pickup_datetime	String dropoff_datetime	String passenger_count	Float trip_distance	String payment_type
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	1.39	1
3	2018-03-30T08:54:43	2018-03-30T09:27:15	1	0.8	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	0.46	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	0.98	1
13	2018-12-05T05:52:16	2018-12-05T06:13:35	1	0.88	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	0.1	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

Creating pipeline

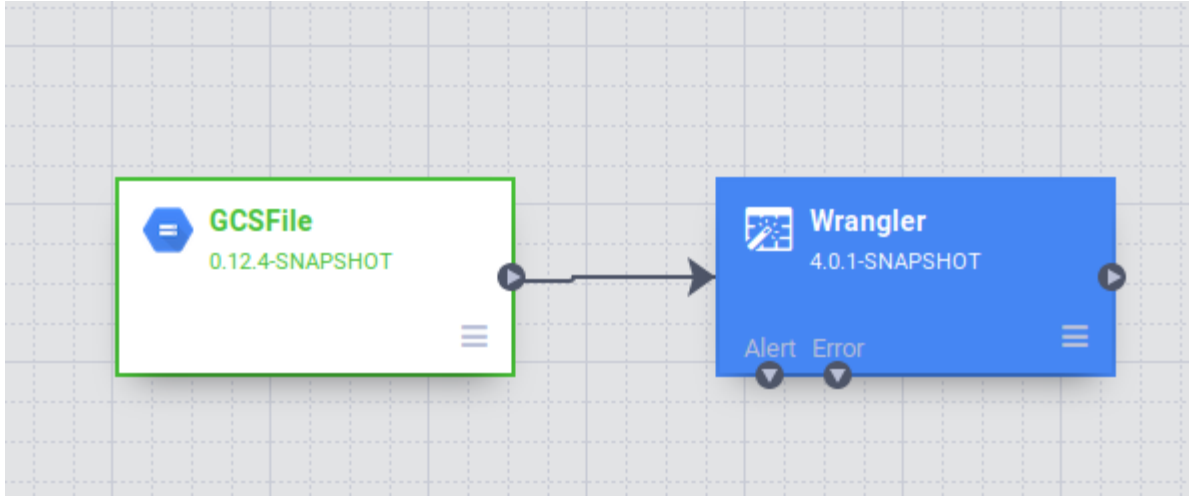
Cloud Data Fusion translates your visually built pipeline into an Apache Spark or MapReduce program that executes transformations on an ephemeral Cloud Dataproc cluster in parallel. This enables you to easily execute complex transformations over vast quantities of data in a scalable, reliable manner, 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**.

Creating pipeline



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.

Adding a data source

- ✓ Create Dataset trips
- ✓ in **bigquery editor** More > Query Settings > Set a destination table for query results. Also, under Table name input zone_id_mapping.
- ✓ Run query
SELECT
zone_id,
zone_name,
borough
FROM
`bigquery-public-data.new_york_taxi_trips.taxi_zone_geom`
- ✓ Cloud Data Fusion > Source > bigquery (add it) and click on properties (see next slide to change setting)

Query settings

Destination

- ☐ Save query results in a temporary table
☒ Set a destination table for query results

Project name: qwiklabs-gcp-02-1c3a502cef0b
Dataset name: trips
Table name: zone_id_mapping

- Destination table write preference
☒ Write if empty
☐ Append to table
☐ Overwrite table

- Results size ?
☐ Allow large results (no size limit)

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			

PropertiesDocumentation

Label *
BigQuery

Basic

Reference Name *
zone_mapping

BROWSE

Project ID
auto-detect

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

Dataset *
trips

Table *
zone_id_mapping

GET SCHEMA

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-00-a1ffccd2637e-temp

Credentials

Service Account File Path
auto-detect

Label is bigquery

Reference :
zone_mapping

Dataset : trips

Table : zone_id_mapping

Temporary Bucket Name name of your
project followed by "-temp"

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

Output schema

Name	Type	Null
zone_id	string	<div><div></div><div>✓</div></div>
zone_name	string	<div><div></div><div>✓</div></div>
borough	string	<div><div></div><div>✓</div></div>

Apply

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

Joining two sources(add ,properties)

- ✓ **Analytics** section > **Joiner**
- ✓ Drag a connection arrows as shown in graph
- ✓ **Properties** of **Joiner**
 - * **Join Type :Inner**
 - * **fill join condition**

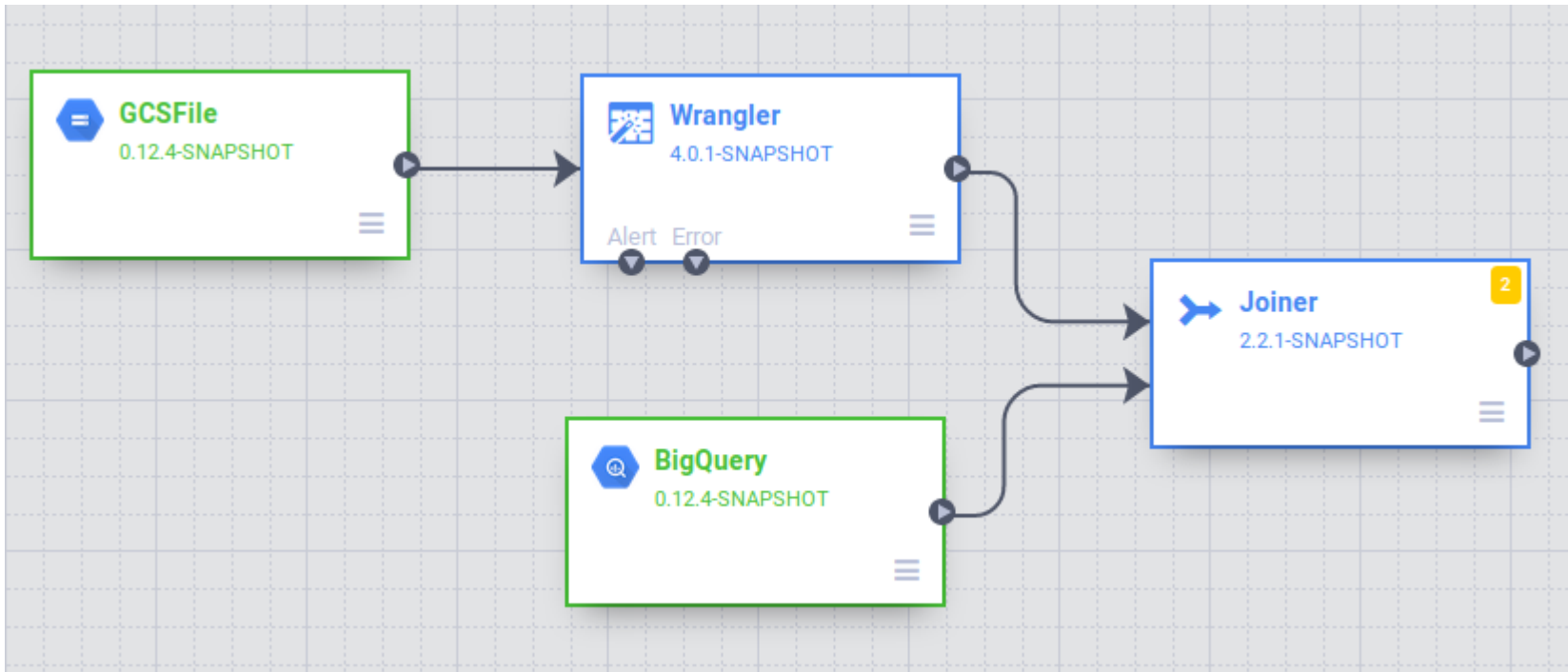
Join Condition *

Wrangler	<input type="text" value="pickup_location_id"/>	=
BigQuery	<input type="text" value="zone_id"/>	

?

+

M



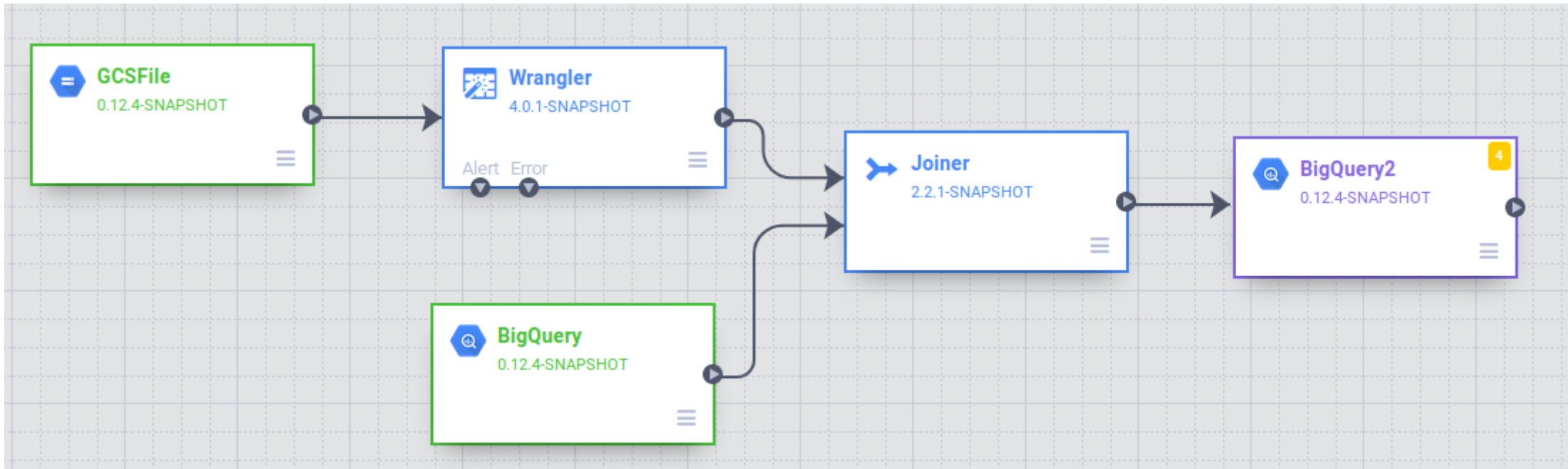
Joining two sources(properties : remove unnecessary columns)

- 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.

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_locat	string	▼	✓	🗑️	+
zone_id	string	▼	✓	🗑️	+
zone_name	string	▼	✓	🗑️	+
borough	string	▼	✓	🗑️	+

Storing the output to BigQuery

- ✓ **Sink** section >> **BigQuery**.
- ✓ Drag a connection arrow
- ✓ Click on properties (see next slide)
- ✓ Close the window



BigQuery Properties 0.14.2

This sink writes to a BigQuery table. BigQuery is Google's serverless, highly scalable, enterprise data warehouse. Data is first written to a temp...

Validate



Properties

Documentation

dropoff_datetime	string	▼	✓
passenger_count	string	▼	✓
trip_distance	float	▼	✓
payment_type	string	▼	✓
fare_amount	string	▼	✓
tip_amount	string	▼	✓
total_amount	string	▼	✓
dropoff_location_id	string	▼	✓
zone_name	string	▼	✓
borough	string	▼	✓

Basic

Reference Name *

bq_insert



Project ID

auto-detect



Dataset *

trips



Table *

trips_pickup_name



Temporary Bucket Name

qwiklabs-gcp-03-b70d69904768-temp



Deploying and running the pipeline

☰

Cloud Data Fusion | Studio

Data Pipeline - Batch

▼

Filter

☰

☒

«

▶ Source


15

▶ Transform

23

▶ Analytics

6



MyPipeline

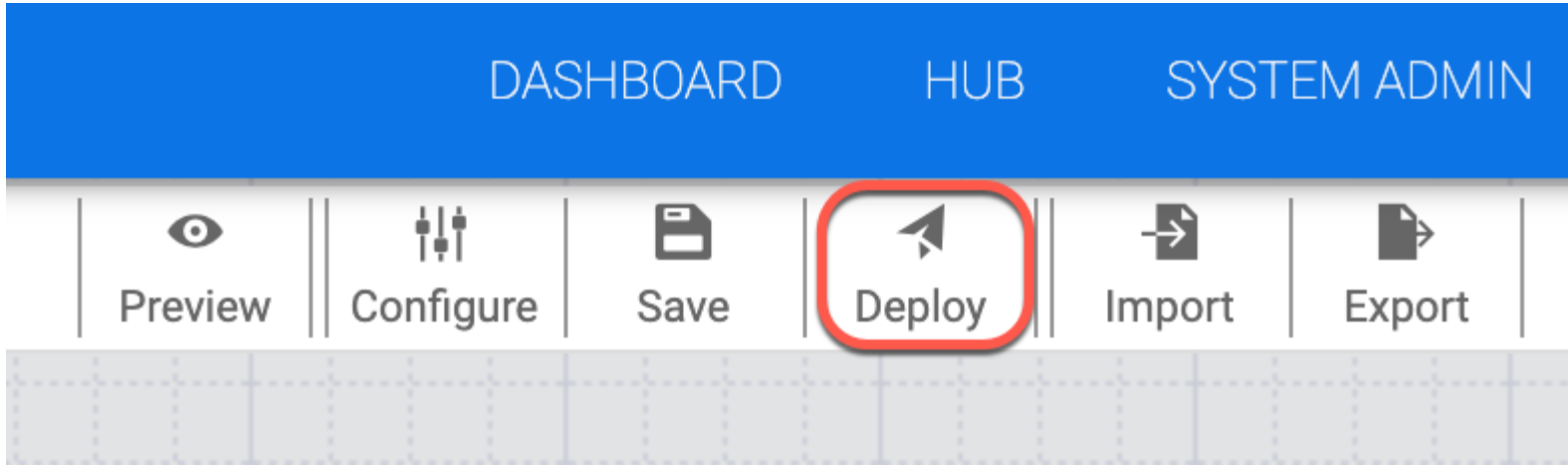
Name your pipeline

Enter a description for your pipeline.

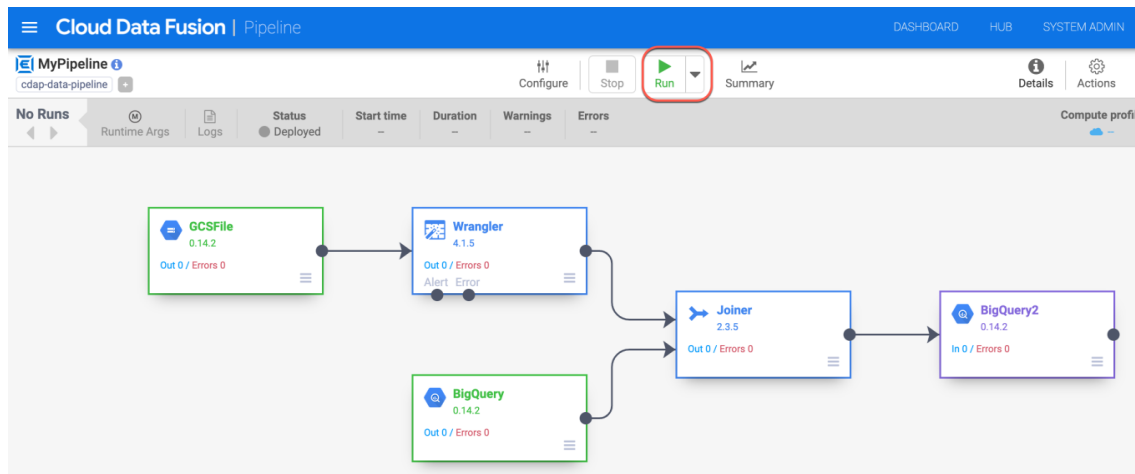
Save

Cancel

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.



Cloud Data Fusion | Pipeline

DASHBOARD HUB SYSTEM ADMIN

MyPipeline **cdap-data-pipeline**

Configure Stop **Run** Summary Details Actions

No Runs

Runtime Args Logs Status Deployed Start time Duration Warnings Errors Compute profile

GCSFile 0.14.2 Out 0 / Errors 0

Wrangler 4.1.5 Out 0 / Errors 0 Alert Error

BigQuery 0.14.2 Out 0 / Errors 0

Joiner 2.3.5 Out 0 / Errors 0

BigQuery2 0.14.2 In 0 / Errors 0

Run pipeline

Cloud Data Fusion | Pipeline

DASHBOARD HUB SYSTEM ADMIN Basic Edition

MyPipeline **cdap-data-pipeline**

Configure Stop **Run** Summary Details Actions

Run 1 of 1

Runtime Args Logs **Status** Succeeded Start time 05-22-2020 02:39:40 PM Duration 5 mins 8 secs Warnings 6 Errors 0 Compute profile Dataproc

GCSFile 0.14.2 Out 4,925 / Errors 0

Wrangler 4.1.5 Out 4,890 / Errors 0 Alert Error

BigQuery 0.14.2 Out 263 / Errors 0

Joiner 2.3.5 Out 4,817 / Errors 0

BigQuery2 0.14.2 In 4,817 / Errors 0

Success process

Viewing result in bigquery

```
SELECT * FROM `trips.trips_pickup_name`
```