Prepare Data for ML APIs on Google Cloud: Challenge Lab

Project Goal

Demonstrate the integration and use of Google Cloud services including **Dataflow**, **Dataproc**, **Speech-to-Text API**, and **Cloud Natural Language API** by processing structured, unstructured, and audio data.

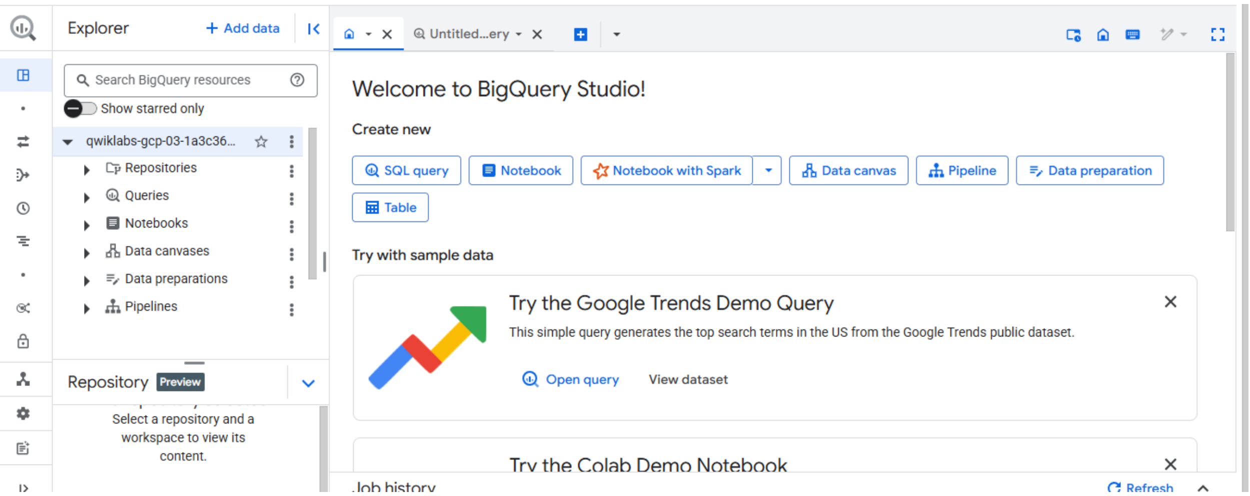
Task 1. Run a simple Dataflow job

In this task, you use the Dataflow batch template **Text Files on Cloud Storage to BigQuery** under "Process Data in Bulk (batch)" to transfer data from a Cloud Storage bucket (gs://cloud-training/gsp323/lab.csv). The following table has the values you need to correctly configure the Dataflow job.

You will need to make sure you have:

- Create a BigQuery dataset called BigQuery Dataset Name with a table called Output Table Name.
- Create a Cloud Storage Bucket called Cloud Storage Bucket Name.

Field	Value
Cloud Storage input file(s)	`gs://cloud-training/gsp323/lab.csv`
Cloud Storage location of your BigQuery schema file	`gs://cloud-training/gsp323/lab.schema`
BigQuery output table	`Output Table Name`
Temporary directory for BigQuery loading process	`Temporary BigQuery Directory`
Temporary location	`Temporary Location`
Optional Parameters > JavaScript UDF path in Cloud Storage	`gs://cloud-training/gsp323/lab.js`
Optional Parameters > JavaScript UDF name	`transform`
Optional Parameters > Machine Type	`e2-standard-2`



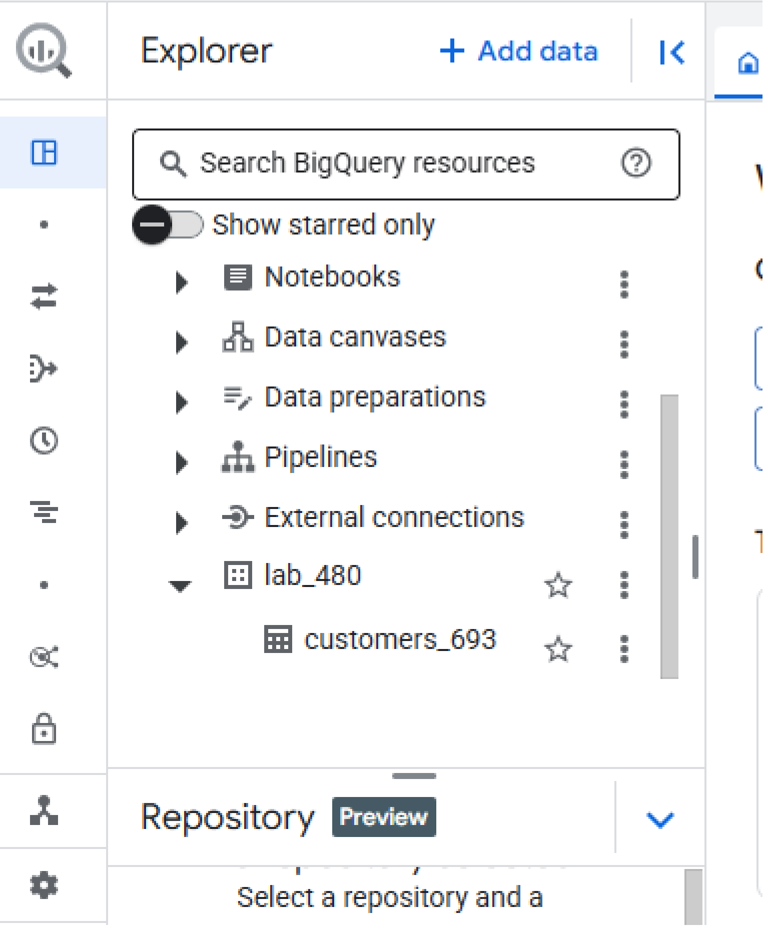
Welcome to BigQuery Studio!

Try the Google Trends Demo Query

This simple query generates the top search terms in the US from the Google Trends public dataset.

Try the Colab Demo Notebook

Select a repository and a workspace to view its content.



Select a repository and a

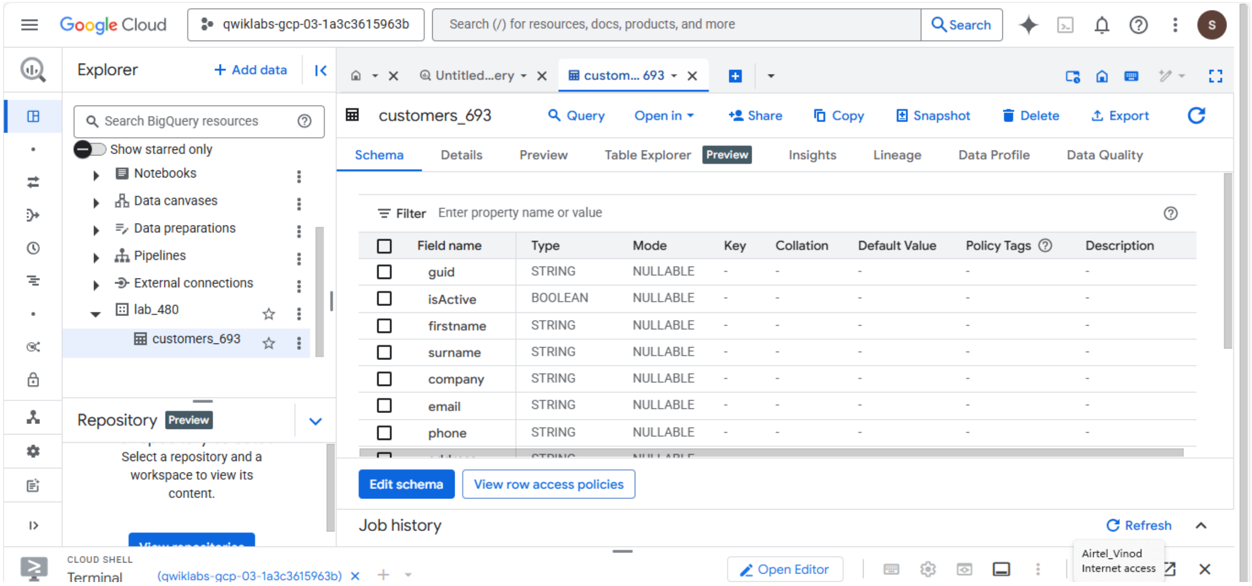
Task 2. Run a simple Dataproc job

In this task, you run an example Spark job using Dataproc.

Before you run the job, log into one of the cluster nodes and copy the /data.txt file into hdfs (use the command hdfs dfs -cp gs://cloud-training/gsp323/data.txt /data.txt).

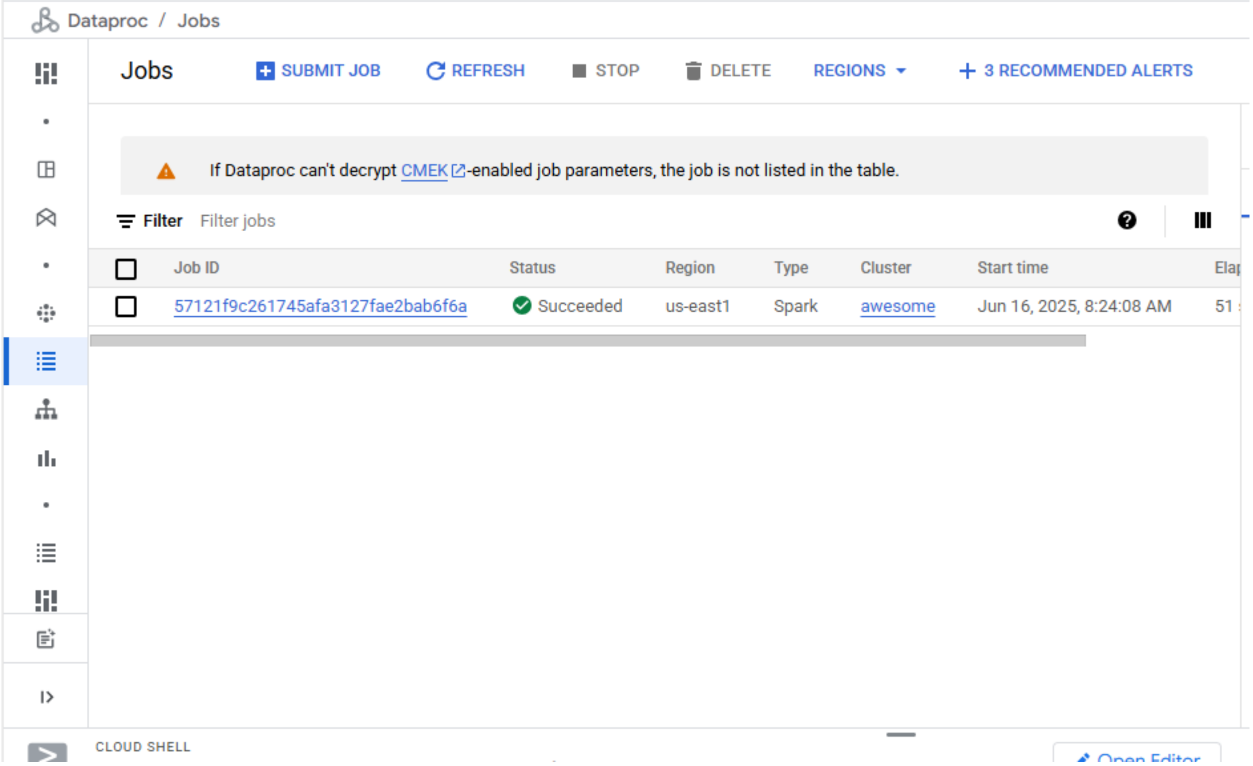
Run a Dataproc job using the values below.

Field	Value
Region	`Region`
Job type	`Spark`
Main class or jar	`org.apache.spark.examples.SparkPageRank`
Jar files	`file:///usr/lib/spark/examples/jars/spark-examples.jar`
Arguments	`/data.txt`
Max restarts per hour	`1`
Dataproc Cluster	`Compute Engine`
Region	`Region`
Machine Series	`E2`
Manager Node	Set **Machine Type** to **e2-standard-2**
Worker Node	Set **Machine Type** to **e2-standard-2**
Max Worker Nodes	`2`
Primary disk size	`100 GB`
Internal IP only	`Deselect "Configure all instances to have only internal IP addresses`



customers_693

☐	Field name	Type	Mode	Key	Collation	Default Value	Policy Tags	Description
☐	guid	STRING	NULLABLE	-	-	-	-	-
☐	isActive	BOOLEAN	NULLABLE	-	-	-	-	-
☐	firstname	STRING	NULLABLE	-	-	-	-	-
☐	surname	STRING	NULLABLE	-	-	-	-	-
☐	company	STRING	NULLABLE	-	-	-	-	-
☐	email	STRING	NULLABLE	-	-	-	-	-
☐	phone	STRING	NULLABLE	-	-	-	-	-



If Dataproc can't decrypt CMEK -enabled job parameters, the job is not listed in the table.

☐	Job ID	Status	Region	Type	Cluster	Start time	Elap
☐	57121f9c261745afa3127fae2bab6f6a	Succeeded	us-east1	Spark	awesome	Jun 16, 2025, 8:24:08 AM	51

Task 3. Use the Google Cloud Speech-to-Text API

- Use Google Cloud Speech-to-Text API to analyze the audio file gs://cloud-training/gsp323/task3.flac. Once you have analyzed the file, upload the resulting file to: Cloud Speech Location

Task 4. Use the Cloud Natural Language API

- Use the Cloud Natural Language API to analyze the sentence from text about Odin. The text you need to analyze is "Old Norse texts portray Odin as one-eyed and long-bearded, frequently wielding a spear named Gungnir and wearing a cloak and a broad hat." Once you have analyzed the text, upload the resulting file to: Cloud Natural Language Location