Data to Insights: Ingesting and Querying New Datasets v1.1

30 minutes
1 Credit

Rate Lab

Overview

In this lab, you will ingest new data sources into Google BigQuery and learn how to query external data sources

directly.

Objectives

- Ingest Data from Google Cloud Storage into Google BigQuery

- Query from a CSV in Cloud Storage directly as an External Data Source

Ingesting Data from Google Cloud Storage

Upload a dataset to Google Cloud Storage

Step 1

Download this CSV: NAICS_digit_2017_codes.csv.

Step 2

Open the Google Cloud Console.

Step 3

Go to **Storage** in the **Navigation menu** (left-side navigation).

Step 4

Click **Create Bucket** (or use an existing bucket).

Step 5

In the Create a bucket window that will appear, add a unique bucket name and leave the remaining settings at their default values.

Step 6

Click **Create**.

Step 7

Click **your-bucket-name**.

Step 8

Click **Upload Files**.

Step 9

Navigate to the CSV file you downloaded from Step 1 and

upload it.

Wait for the file to upload.

Step 10

Confirm the file has uploaded to your storage bucket.

Ingesting a CSV into a Google BigQuery Table

Open BigQuery Console

In the Google Cloud Console, select **Navigation menu** > **BigQuery**:

The **Welcome to BigQuery in the Cloud Console** message box opens. This message box provides a link to the quickstart guide and lists UI updates.

Click **Done**.

Step 1

In the BigQuery console, click on the name of your project,

then click **Create Dataset**.

Name it `irs_990`. Leave the options at their default values (Data Location, Default table Expiration). Click **Create dataset**.

Step 2

Select `irs_990` dataset and click on **Create table**.

Step 3

Populate the following Create Table options:

- In **Source > Create table from:**, change the drop down to **Google Cloud Storage**.

- Copy and Paste the below GCS path and change `<your-bucket-name>` to your bucket:

```
gs://<your-bucket-name>/NAICS_digit_2017_codes.csv
```

- Confirm `irs_990` is the default dataset selected.

- Leave Table type as Default (Native table).

- For **Table name** type `naics_digit_2017_codes`.

- For Schema, check **Schema and input parameters** to Auto Detect the schema.

- Leave the Other Options as Default.

- Click **Create table**.

Step 4

Confirm the new `naics_digit_2017_codes` schema looks similar to the below:

Step 5

Click on **Preview** to see sample data values.

It looks like we've potentially ingested unnamed or blank columns, we can clean these up using SQL or Cloud Dataprep as we learned in previous labs.

Reading a CSV as an External Data Source in Google BigQuery Table

Instead of ingesting and storing the CSV data table in Google BigQuery, you decide you want to query the underlying data source directly.

The process is essentially the same as before except for changing the Table Type.

Step 1

Select `irs_990` dataset and click on **Create table**.

Step 2

Populate the following Create Table options:

- In **Source > Create table from:**, change the drop down to **Google Cloud Storage**.

- Copy and Paste the below GCS path:

```
gs://data-insights-course/labs/lab5-ingesting-and-querying/irs990_code_lookup.csv
```

-

- Confirm `irs_990` is the default dataset selected.

- Change the Table type to **External table** .

- For **Table name** type `irs990_code_lookup`.

- Populate the **Schema** as follows by clicking **Add Field** and filling out the input boxes.

Name	Type	Mode
irs_990_field	STRING	NULLABLE
code	STRING	NULLABLE
description	STRING	NULLABLE

- Under Advanced Options, **Header rows to skip** put `1`.

- Leave the Other Options as Default.

- Click **Create table**.

Wait for the table to be created.

Step 3

Copy and Paste the below query into the **Query editor**.

Step 4

Change the **Project name** from data-to-insights to your own.

```
#standardSQL
# Lookup what IRS code values mean
SELECT
  irs_990_field,
  code,
  description
FROM
  `your-project.irs_990.irs990_code_lookup` # change
WHERE
  irs_990_field IN ('elf','subcd')
```

Step 5

Click **Run**.

Step 6

Insights: Read through the query results.

What does the field `elf` mean?

`elf` denotes how the return was filed: E for Electronic, P for Paper.

Are the subsection (subcd) codes unique?

No, they are not. Code 3 is used multiple times to denote 8 possible different Organization types (Charitable Corporation, Educational Organization, etc..). This insight will become particularly important when we use this as a lookup value for our individual filings. We will learn how to handle this in our upcoming labs.

Performance Pitfall: Creating and querying from External Data Sources directly (e.g. CSVs stored on Google Cloud Storage) has performance impacts as Google BigQuery has less control over data outside of its fully-managed data warehouse.

Congratulations! You have learned how to ingest data into Google BigQuery and query external data sources directly. In future labs, we will merge these data sources together for

a single enriched reporting data source.

Congratulations!

You have completed the second part of the **BigQuery Data Ingestion** lab.

Learning Review

- Google BigQuery supports ingesting data from many sources. Popular ones are Google Cloud Storage, CSV, JSON, ARVO, Cloud BigTable and more.

- You can query External Data Sources directly but there are limitations (particularly around performance)

- Auto-detecting the data schema when Creating a Table may lead to fields you did not want to include in the dataset. Consider manually spelling out the schema for more control.

References

- Loading Data from Cloud Storage
- Querying External Data Sources

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.