Start Lab 01:30:00

Analyzing data using AI Platform Notebooks and BigQuery

1 hour 30 minutes Free Rate Lab

Overview

Duration is 1 min

In this lab you analyze a large (70 million rows, 8 GB) airline dataset using Google BigQuery and AI Platform Notebooks.

What you learn

In this lab, you:

- Launch AI Platform Notebooks
- Invoke a BigQuery query
- Create graphs in AI Platform Notebooks

This lab illustrates how you can carry out data exploration of large datasets, but continue to use familiar tools like Pandas and Juypter. The "trick" is to do the first part of your aggregation in BigQuery, get back a Pandas dataset and then work with the smaller Pandas dataset locally. AI Platform Notebooks provides a managed Jupyter experience, so that you don't need to run notebook servers yourself.

Setup

For each lab, you get a new Google Cloud 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 the Google Cloud Console.

   Open Google Console

   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.

   Username
   google2876526_student@qwiklabs.n

   Password
   TG959yrKDX

   GCP Project ID
   qwiklabs-gcp-0855e773352d3560

   New to labs? View our introductory video!

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.

Deployment Manager

This lab is using a deployment manager script to create the Cloud AI Platform instance you will need for this exercise.

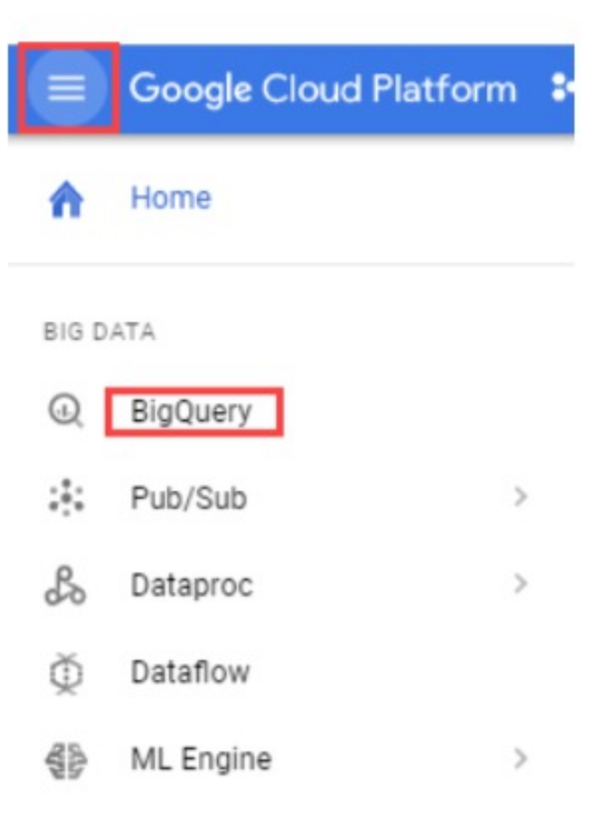
It should take 2 - 3 minutes for the instance to be ready.

Please wait before launching the Jupyter notebook, otherwise the script may be interrupted and the repository may not be cloned.

Invoke BigQuery

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 query textbox, type:

```
#standardSQL
SELECT
  departure_delay,
  COUNT(1) AS num_flights,
  APPROX_QUANTILES(arrival_delay, 5) AS arrival_delay_quantiles
FROM
  `bigquery-samples.airline_ontime_data.flights`
GROUP BY
  departure_delay
HAVING
  num_flights > 100
ORDER BY
  departure_delay ASC
```

Click **Run**.

What is the median arrival delay for flights left 35 minutes early?

(Answer: the typical flight that left 35 minutes early arrived 28 minutes early.)

Step 2 (Optional)

Can you write a query to find the airport pair (departure and arrival airport) that had the maximum number of flights between them?

Hint: you can group by multiple fields.

One possible answer:

```
#standardSQL
SELECT
  departure_airport,
  arrival_airport,
  COUNT(1) AS num_flights
FROM
  `bigquery-samples.airline_ontime_data.flights`
GROUP BY
  departure_airport,
  arrival_airport
ORDER BY
  num_flights DESC
LIMIT
  10
```

Draw graphs in AI Platform Notebooks

Step 1

Click on the **Navigation Menu**. Navigate to **AI Platform**, then to **Notebooks**.

Step 2

Click on **Open JupyterLab** for the instance named as **python-notebook**.

Step 3

In JupyterLab, start a new notebook by clicking on **Notebook** > **Python 3**.

Step 4

In the first cell of the notebook type the following to install the `google-cloud-bigquery` with version `1.25.0`, then click **Run**.

```
!pip install google-cloud-bigquery==1.25.0
```

Restart the kernel by clicking **Kernel** > **Restart Kernel**.

Step 5

In the next cell in the notebook, type the following, then click **Run**.

```
query="""
SELECT
  departure_delay,
  COUNT(1) AS num_flights,
  APPROX_QUANTILES(arrival_delay, 10) AS arrival_delay_deciles
FROM
  `bigquery-samples.airline_ontime_data.flights`
GROUP BY
  departure_delay
HAVING
  num_flights > 100
ORDER BY
  departure_delay ASC
"""

from google.cloud import bigquery
df = bigquery.Client().query(query).to_dataframe()
df.head()
```

Note that we have gotten the results from BigQuery as a Pandas dataframe.

In what Python data structure are the deciles in?

Step 6

In the next cell in the notebook, type the following, then click **Run**.

```
import pandas as pd
percentiles = df['arrival_delay_deciles'].apply(pd.Series)
percentiles = percentiles.rename(columns = lambda x : str(x*10) + "%")
df = pd.concat([df['departure_delay'], percentiles], axis=1)
df.head()
```

What has the above code done to the columns in the Pandas DataFrame?

Step 7

In the next cell in the notebook, type the following, then click **Run**.

```
without_extremes = df.drop(['0%', '100%'], 1)
without_extremes.plot(x='departure_delay', xlim=(-30,50), ylim=(-50,50));
```

Suppose we were creating a machine learning model to predict the arrival delay of a flight. Do you think departure delay is a good input feature? Is this true at all ranges of departure delays?

Hint: Try removing the xlim and ylim from the plotting command.

Summary

In this lab, you learned how to carry out data exploration of large datasets using BigQuery, Pandas, and Jupyter. The "trick" is to do the first part of your aggregation in BigQuery, get back a Pandas dataset and then work with the smaller Pandas dataset locally. AI Platform Notebooks provides a managed Jupyter experience, so that you don't need to run notebook servers yourself.

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.