Analyzing data using AI Platform Notebooks and BigQuery

Overview

In this lab, you analyze a large (70 million rows, 8 GB) airline dataset using 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 explore 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, have a Pandas dataset returned, and then work with the smaller Pandas dataset locally. AI Platform Notebooks provides a managed Jupyter experience, so you don't need to run notebook servers yourself. For more information about how to visualize BigQuery data in a Jupyter notebook, see Visualizing BigQuery data in a Jupyter notebook.

Setup

For each lab, you get a new Google Cloud project and set of resources for a fixed time at no cost.

1. Sign in to Qwiklabs using an **incognito window**.

2. Note the lab's access time (for example, 02:00:00), and make sure you can finish within that time. 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 (**Username** and **Password**). You will use them to sign in to the Google Cloud 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 receive errors or **incur charges**.

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

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

Deployment Manager

This lab uses a Cloud Deployment Manager script to create the Cloud AI Platform instance you will need for this exercise. The instance should be ready in 2 or 3 minutes.

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

Invoke BigQuery

Open BigQuery Console

1. In the Google Cloud Console, on the **Navigation menu** , click **BigQuery**. The **Welcome to BigQuery in the Cloud Console** dialog opens. This dialog provides a link to the quickstart guide and lists UI updates.

2. Click **Done** to close the dialog.

Step 1

In the BigQuery **Editor**, 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
```

Copied!
content_copy
Click **Run**.

What is the median early arrival for flights that 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 solution:

```
#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
```

Copied!

content_copy

Draw graphs in AI Platform Notebooks

Step 1

In the Google Cloud Console, on the **Navigation Menu**, click **AI Platform > Notebooks**.

Step 2

For the **python-notebook** instance, click **Open JupyterLab**.

Step 3

In JupyterLab, to start a new notebook, click **Notebook > Python 3**.

Step 4

In the first cell of the notebook, to install BigQuery version 1.25.0, type the following, and then click **Run**.

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

Copied!
content_copy
Note: Restart your **kernel** to use updated packages. Ignore the deprecation warnings and incompatibility errors related to Cloud Storage.
Step 5

In the next cell of the notebook, type the following, and 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()
```

Copied!

content_copy
Note that the results from BigQuery are returned as a Pandas dataframe.

What Python data structure are the deciles in?

Step 6

In the next cell of the notebook, type the following, and 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()
```

Copied!
content_copy
What did this code do to the columns in the Pandas dataframe?

Step 7

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

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

Copied!
content_copy
If you were creating a machine learning model to predict the arrival delay of a flight, would a departure delay be 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. For more information about working with the Pandas library, see Pandas: How to Read and Write Files.