# MSDS 7346
# Cloud Computing
# Mini Project 6 – Data Analysis using GCP

Name: Mooyoung Lee

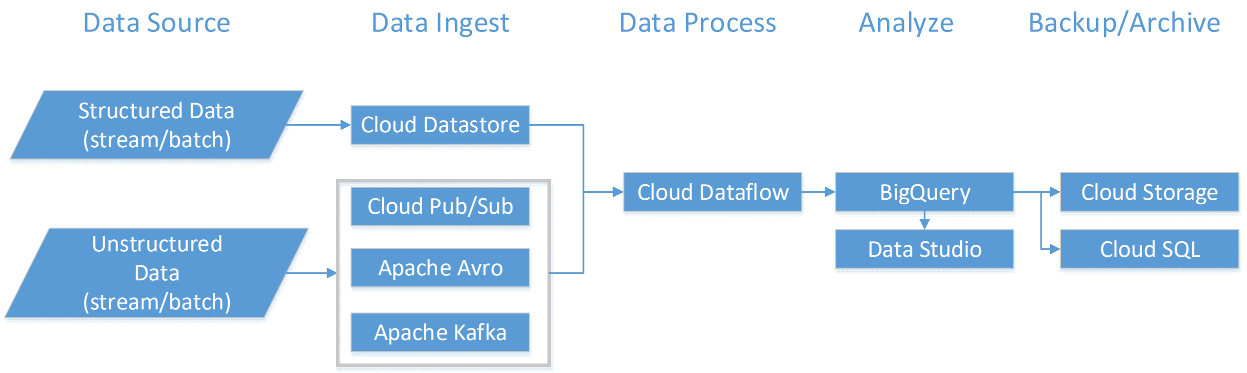
A marketing firm provides services to different clients. As part of marketing services they obtain data from variety of sources both from Enterprise Data Warehouse as well as Social media. For the purposes of this exercise we will assume that each source provides a daily batch file that includes data for all client to the marketing company. Some sources of data will be in **a star schema** while others may be **unstructured** or **semi-structured data**. In addition to the batch data, some or all of these systems will also stream real-time updates, you can consider them as IOT type devices.

The Marketing company is looking for some recommendations on how to set up a system that allows them to analyze trends and patterns over time and across clients. **Both batch and streaming data** is coming in, batch files are coming **once per day**. The batch data can be used as is without any transformation but the **streaming data requires some transformation** and need to be **analyzed real-time**. They also need to **provide reporting** on this data as well as have the **ability to export the data** for each individual client.

The data needs to be **backed up for 30 days and after 30 days it is archived for 5 years**.

**Question 1 : Design the architecture that can efficiently store and analyze both structured and unstructured data sets**

1) Develop an architecture diagram to solve this problem. I am looking for a one page block-level architecture using either AWS or Google Cloud public offering. You need to show how you would handle two data streams, transformation etc. This is your pitch to the marketing company how you will handle their data architecture. Since the information provided to you is very sparse, please make assumption, include those assumptions as part of the presentation.



2) Provide a very brief description of why you choose certain services - there is not one right answer, I am looking for your reasoning.

Above diagram is a direct copy from the product information page of google cloud dataflow. There was no need to modify the diagram so I re-used the existing one.

| Target Spec | Google Cloud Tool | Description |
|---|---|---|
| Structured data ingest | Cloud Datastore | Cloud datastore is good for highly available structured data at scale. |
| Unstructured data ingest | Cloud Pub/Sub,<br>Apache Kafka,<br>Apache Avro | Cloud pub/sub stores event streams and send data to Cloud Dataflow for analysis.<br>Apache Kafka is a software to handle real-time data feed.<br>Apache Avro serialize data and store w/ JSON format. |
| Process real-time and batch data; Transformation | Dataflow | Dataflow is a service helping to transform data in stream and batch modes. |
| Analyze real-time | BigQuery | It is a service that handle 100,000 streaming rows per second and provide ad-hoc real-time analysis using standard SQL. |
| Provide reporting | Data Studio | Data studio provide informative dashboards and reports. |
| Export data | BigQuery/<br>Data Studio | Selected data and reports can be exported with BigQuery and Data Studio. BigQuery can export data to CSV, JSON, and Avro formats. |
| Backup 30 days and archive 5 years | Cloud Storage,<br>Cloud SQL,<br>BigQuery | Cloud Storage provide backup or archive solution for unstructured data and binary data.<br>Cloud SQL is a distributed MySQL database which allows automated backups and point-in-time restoration.<br>BigQuery can also be used for data archive. |

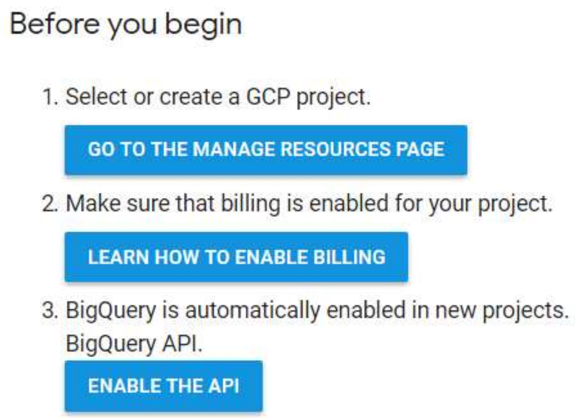
3) Clearly state your assumptions.

- Marketing company will verify this architecture with cloud service provider to make sure this configuration is the most efficient, in both performance and price, for the company's application.
- There is no hidden target specification that is not listed on the problem statement.
- Cloud system can fail and lose all data but it is much better reliable than company's own server.
- The data and information on cloud can be stolen which also can happen from company's own server.
- Everything works well as the google documentation explained from below cites.

Resources:
https://cloud.google.com/dataflow/
https://cloud.google.com/solutions/iot-overview
https://cloud.google.com/pubsub/
https://cloud.google.com/datastore/docs/concepts/overview
https://en.wikipedia.org/wiki/Apache_Kafka
https://en.wikipedia.org/wiki/Apache_Avro
https://cloud.google.com/data-studio/
https://cloud.google.com/solutions/big-data/stream-analytics/
https://cloud.google.com/solutions/designing-a-disaster-recovery-plan

**Question 2 : Store, process, and analyze the provided data**
I have provided you a public dataset. Load this dataset in the BigQuery to do interactive analysis.
1) Load data in BigQuery in Google Cloud Platfrom

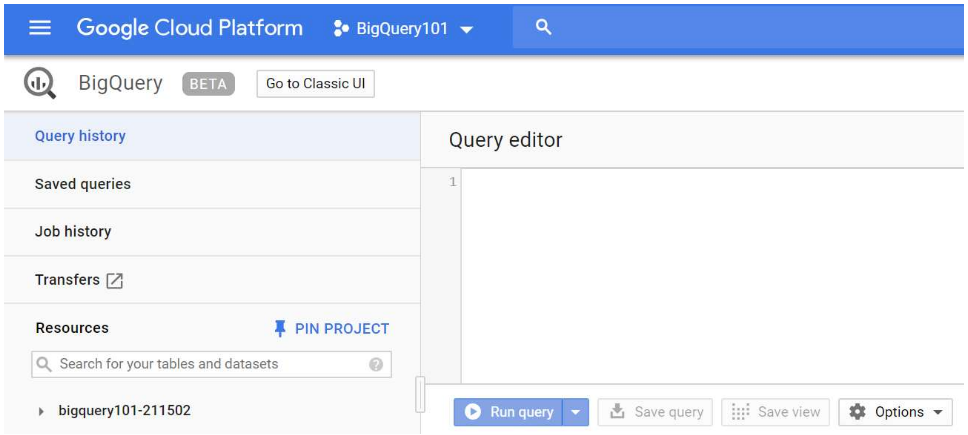
- Go to below cite. Create a CGP project and Enable the API. https://cloud.google.com/bigquery/docs/quickstarts/quickstart-web-ui



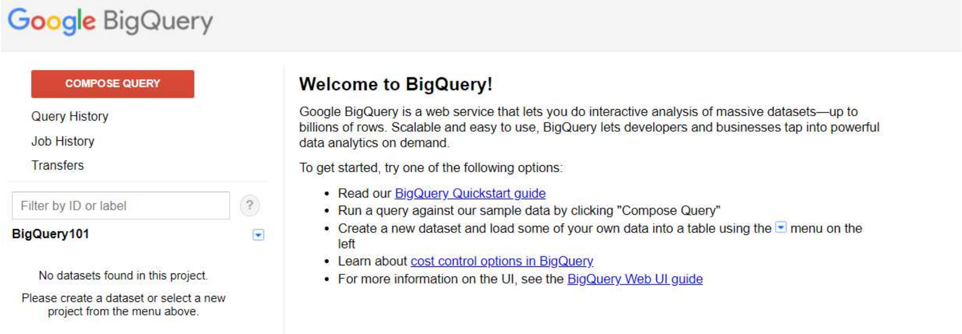
Before you begin

1. Select or create a GCP project.
2. Make sure that billing is enabled for your project.
3. BigQuery is automatically enabled in new projects. BigQuery API.

- Go to BigQuery page by typing 'bigquery' and
- Hit 'Go to Classic UI'



Classic view below



No datasets found in this project.

Please create a dataset or select a new project from the menu above.

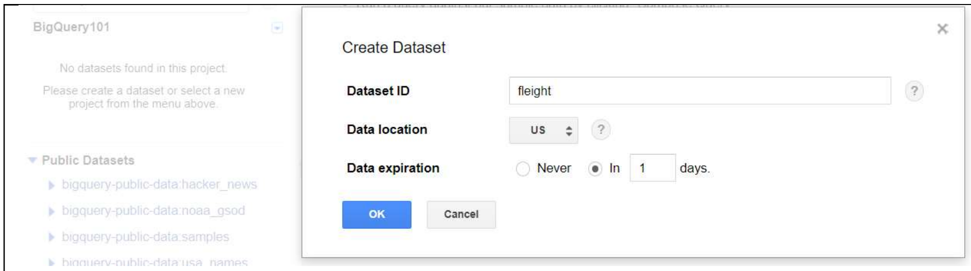
**Welcome to BigQuery!**

Google BigQuery is a web service that lets you do interactive analysis of massive datasets—up to billions of rows. Scalable and easy to use, BigQuery lets developers and businesses tap into powerful data analytics on demand.

To get started, try one of the following options:

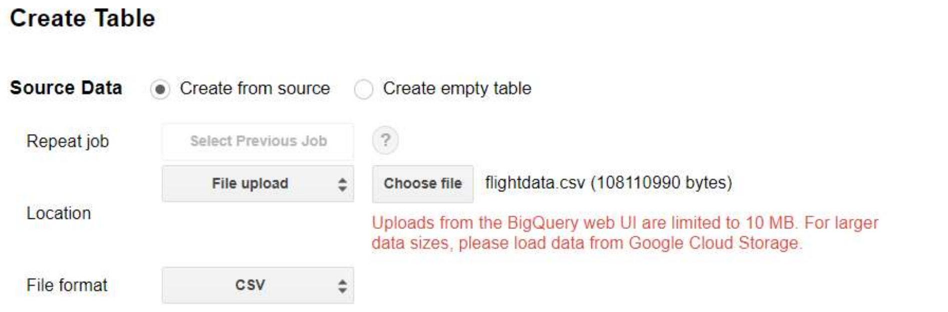
- Read our BigQuery Quickstart guide
- Run a query against our sample data by clicking "Compose Query"
- Create a new dataset and load some of your own data into a table using the ▾ menu on the left
- Learn about cost control options in BigQuery
- For more information on the UI, see the BigQuery Web UI guide

- Hit the blue down arrow button right next to the project name
- Hit the 'Create Dataset' and give a name for the dataset
- It allows file size upto 10MB. If the file size is larger, use Cloud Storage.


BigQuery101
No datasets found in this project.
Please create a dataset or select a new project from the menu above.
Public Datasets
bigquery-public-data:hacker_news
bigquery-public-data:noaa_gsod
bigquery-public-data:samples
bigquery-public-data:usa_names
OK
Cancel

Create Dataset

| | |
|---|---|
| Dataset ID | fleight |
| Data location | US |
| Data expiration | ○ Never ● In 1 days. |

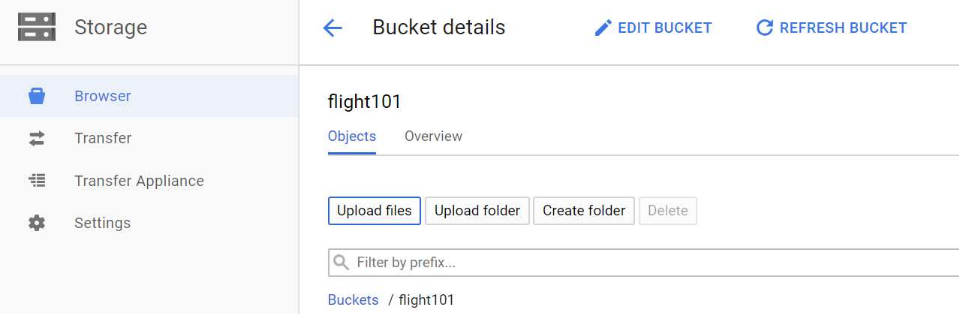

Create Table

Source Data ● Create from source ○ Create empty table

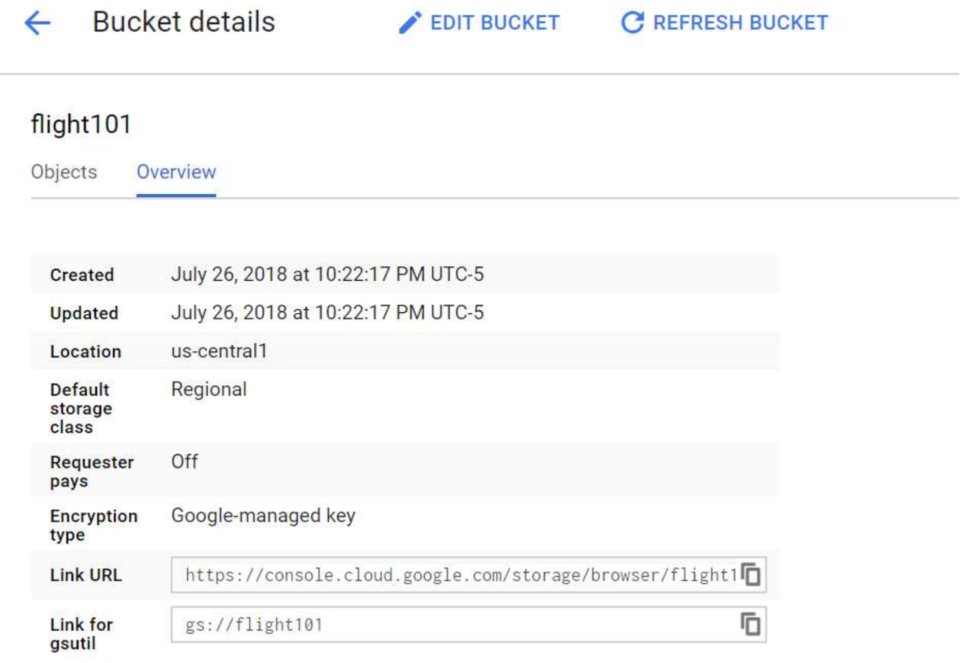
| | |
|---|---|
| Repeat job | Select Previous Job |
| Location | File upload · Choose file flightdata.csv (108110990 bytes) |
| File format | CSV |

Uploads from the BigQuery web UI are limited to 10 MB. For larger data sizes, please load data from Google Cloud Storage.

- Go to Storage and Create a bucket to upload a file

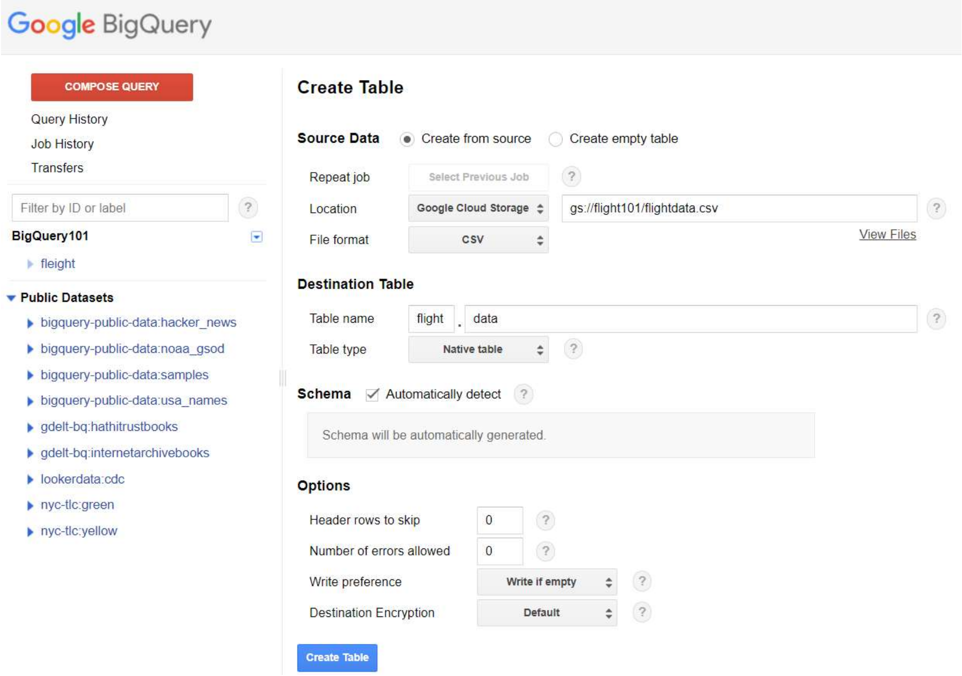


- Go to Overview to find the links for the file



| | |
|---|---|
| Created | July 26, 2018 at 10:22:17 PM UTC-5 |
| Updated | July 26, 2018 at 10:22:17 PM UTC-5 |
| Location | us-central1 |
| Default storage class | Regional |
| Requester pays | Off |
| Encryption type | Google-managed key |
| Link URL | https://console.cloud.google.com/storage/browser/flight1 |
| Link for gsutil | gs://flight101 |

- Retry to create a table using the following settings



Create Table

Source Data ● Create from source ○ Create empty table

Repeat job

Location: Google Cloud Storage gs://flight101/flightdata.csv

File format: CSV

Destination Table

Table name: flight . data

Table type: Native table

Schema ☑ Automatically detect

Schema will be automatically generated.

Options

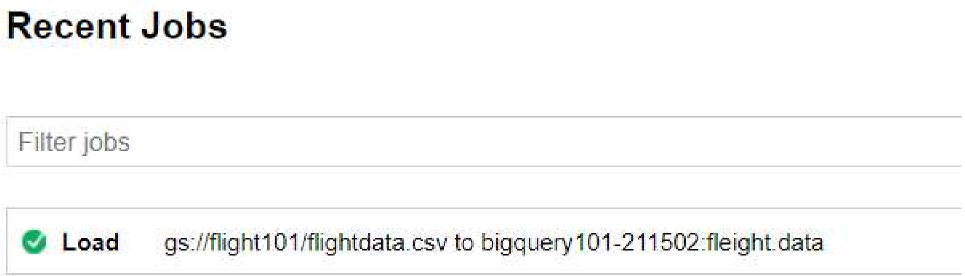
Header rows to skip: 0

Number of errors allowed: 0

Write preference: Write if empty

Destination Encryption: Default

*the destination table name has to be match with the BigQuery table name.



✓ Load gs://flight101/flightdata.csv to bigquery101-211502:fleight.data

- Hit data table to see the schema, details, and preview of dataset

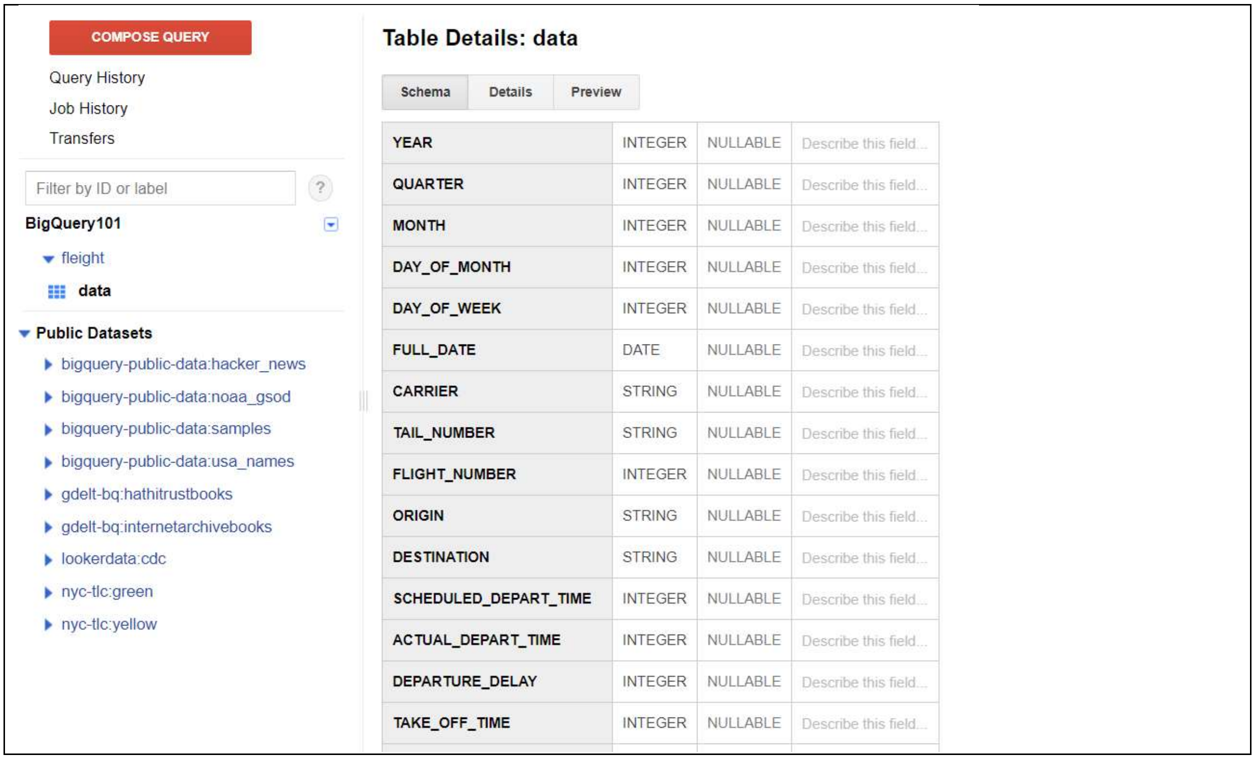
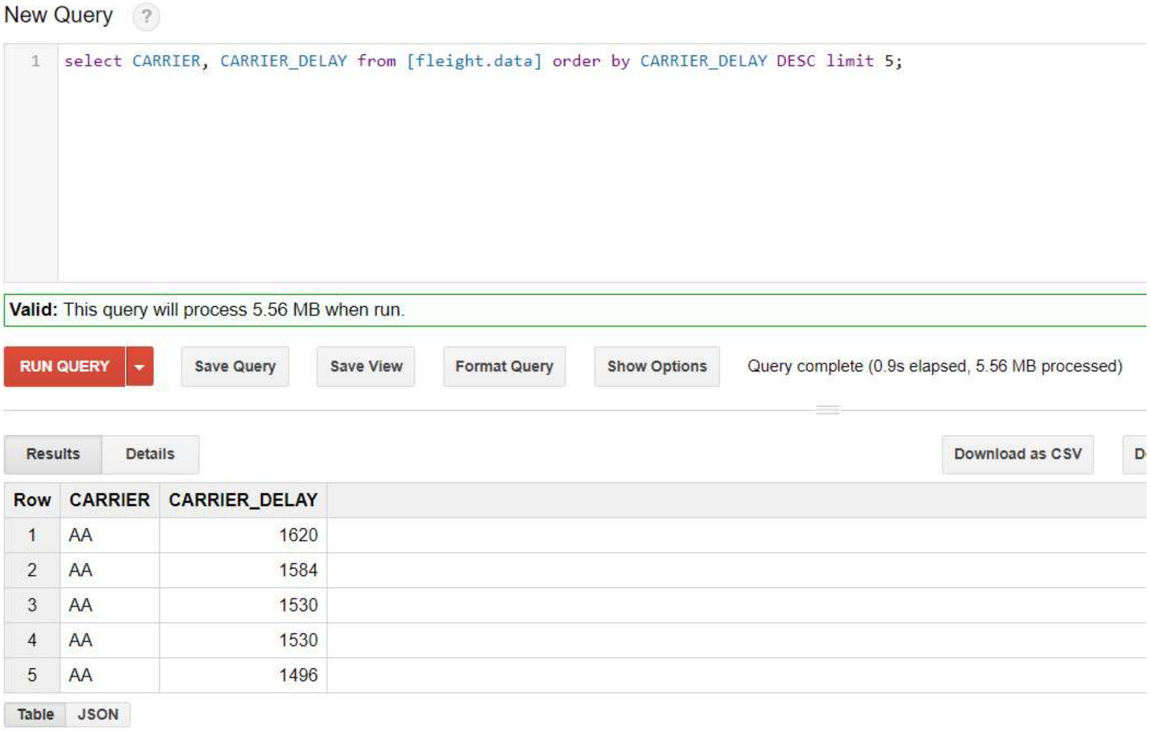


Table Details: data

| | | | |
|---|---|---|---|
| YEAR | INTEGER | NULLABLE | Describe this field... |
| QUARTER | INTEGER | NULLABLE | Describe this field... |
| MONTH | INTEGER | NULLABLE | Describe this field... |
| DAY_OF_MONTH | INTEGER | NULLABLE | Describe this field... |
| DAY_OF_WEEK | INTEGER | NULLABLE | Describe this field... |
| FULL_DATE | DATE | NULLABLE | Describe this field... |
| CARRIER | STRING | NULLABLE | Describe this field... |
| TAIL_NUMBER | STRING | NULLABLE | Describe this field... |
| FLIGHT_NUMBER | INTEGER | NULLABLE | Describe this field... |
| ORIGIN | STRING | NULLABLE | Describe this field... |
| DESTINATION | STRING | NULLABLE | Describe this field... |
| SCHEDULED_DEPART_TIME | INTEGER | NULLABLE | Describe this field... |
| ACTUAL_DEPART_TIME | INTEGER | NULLABLE | Describe this field... |
| DEPARTURE_DELAY | INTEGER | NULLABLE | Describe this field... |
| TAKE_OFF_TIME | INTEGER | NULLABLE | Describe this field... |

2) Create a simple query to find top 5 Carriers with highest Carrier delays.

- Top 5 delays and corresponding carriers


New Query

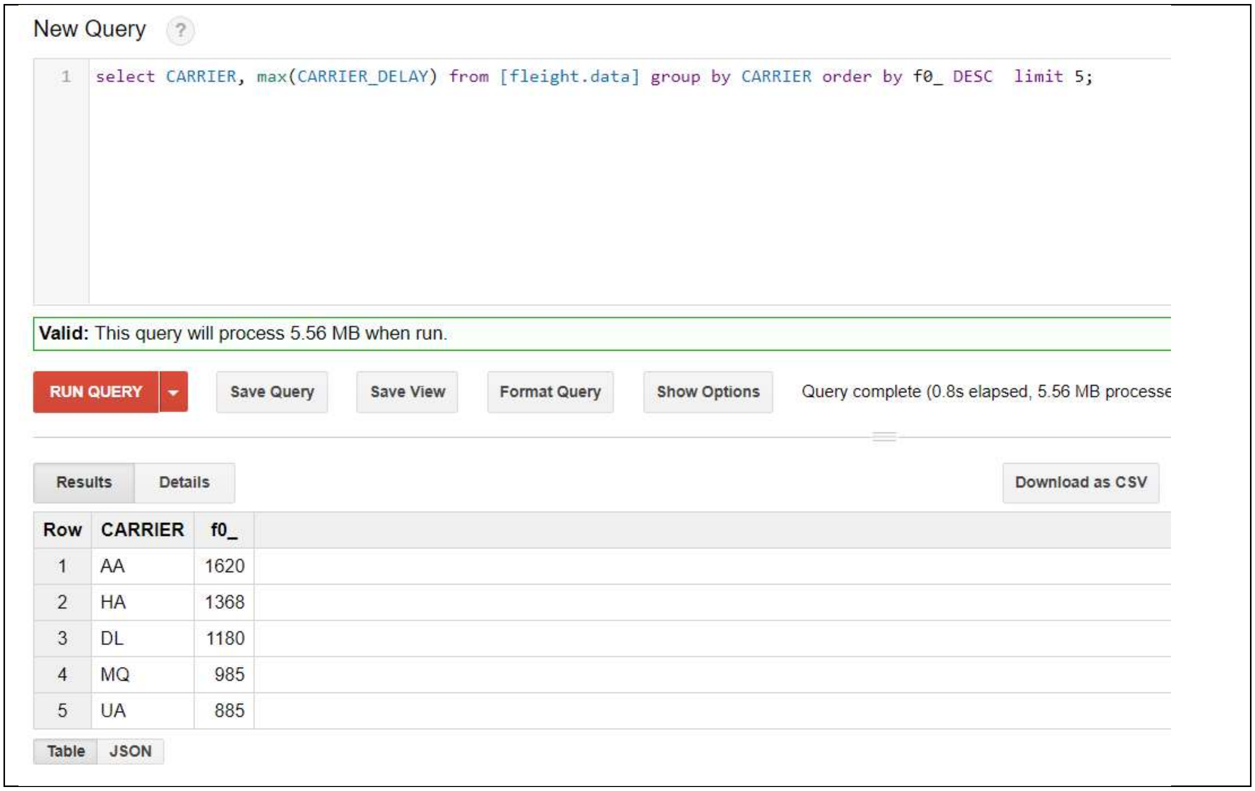
```
select CARRIER, CARRIER_DELAY from [fleight.data] order by CARRIER_DELAY DESC limit 5;
```

**Valid:** This query will process 5.56 MB when run.

Query complete (0.9s elapsed, 5.56 MB processed)

| Row | CARRIER | CARRIER_DELAY |
|---|---|---|
| 1 | AA | 1620 |
| 2 | AA | 1584 |
| 3 | AA | 1530 |
| 4 | AA | 1530 |
| 5 | AA | 1496 |

- Distinct carriers by max delay order



```
select CARRIER, max(CARRIER_DELAY) from [fleight.data] group by CARRIER order by f0_ DESC  limit 5;
```

**Valid:** This query will process 5.56 MB when run.

Query complete (0.8s elapsed, 5.56 MB processe

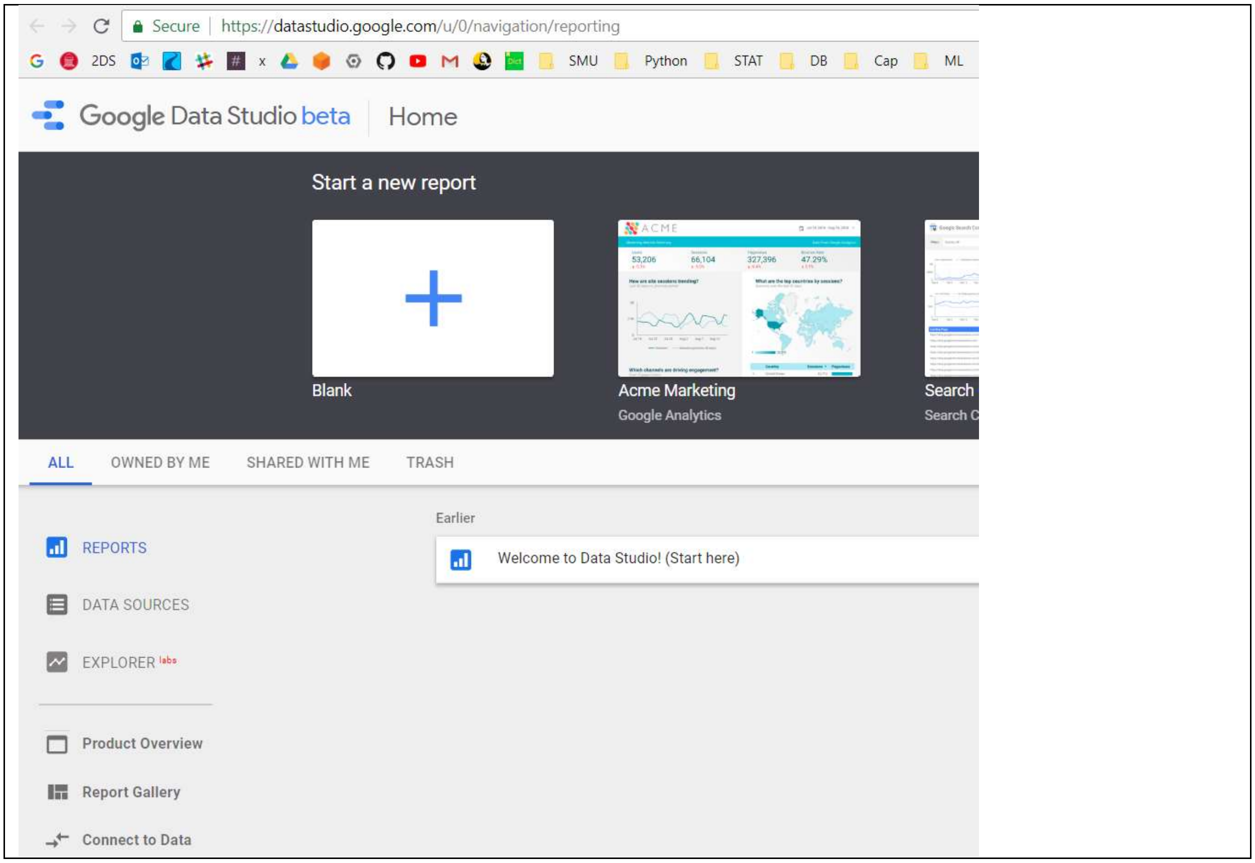
| Row | CARRIER | f0_ |
|---|---|---|
| 1 | AA | 1620 |
| 2 | HA | 1368 |
| 3 | DL | 1180 |
| 4 | MQ | 985 |
| 5 | UA | 885 |

3) Submit reports and screen shots of showing data at various points in GCP
4) Data set is available separately as flightdata.csv

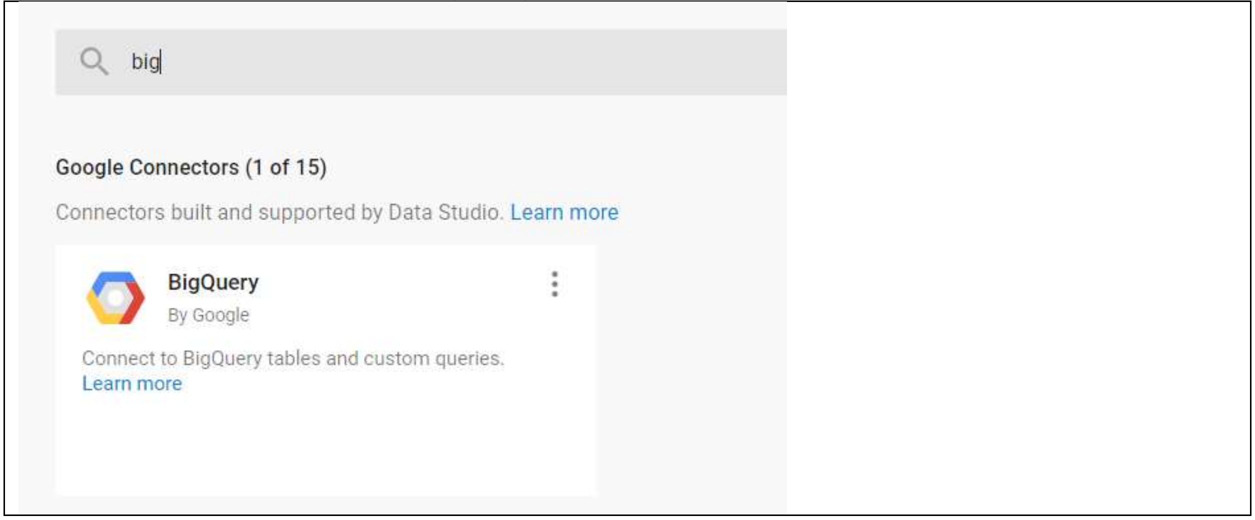
Submission: Submit different screen shots to show completion of each steps

Collaborators: None
Resources:
https://cloud.google.com/bigquery/docs/quickstarts/quickstart-web-ui

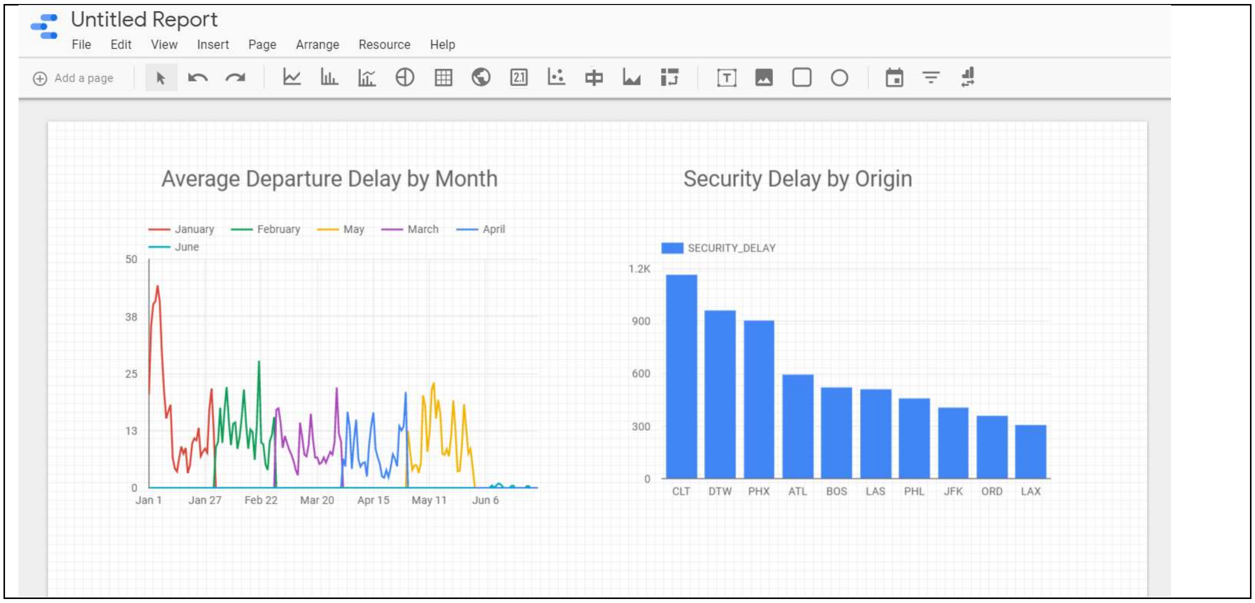
**Question 3 : Create report using Datastudio**
Use the dataset that you have loaded in the previous question to develop a dashboard.
You can choose the content of the dashboard.
1) Go to Google Data Studio by https://datastudio.google.com



2) Click on data source
3) Create a new datasource, select BigQuery


big
Google Connectors (1 of 15)
Connectors built and supported by Data Studio. Learn more
BigQuery
By Google
Connect to BigQuery tables and custom queries.
Learn more

4) Select flight information table that you created in the previous question
5) Have fun with creating a report. This is strictly an exercise of learning this tool. Look at the data and you should be able to generate on dashboard like report.



Submission: Submit different screen shots to show completion of each steps
Collaborators: None
Resources:

https://cloud.google.com/bigquery/docs/visualize-data-studio