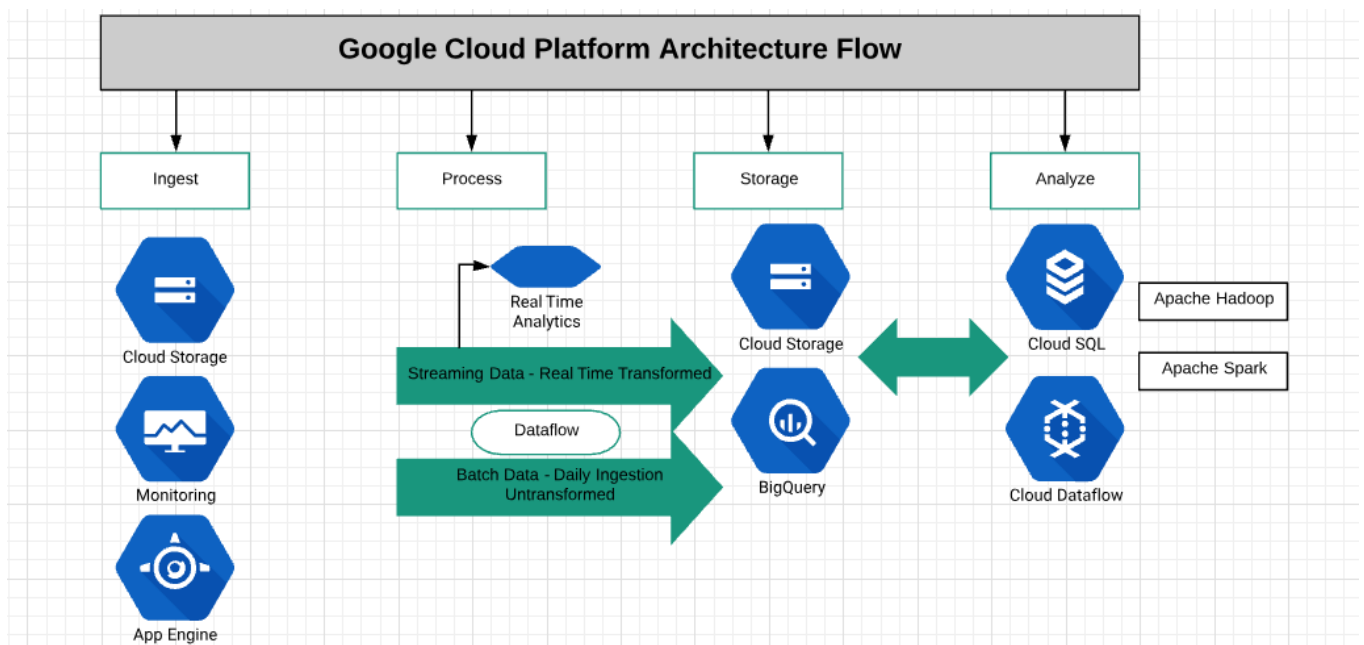


Brian Kolovich
MSDS 7346 Cloud Computing
Mini Project 7 – GCP

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.

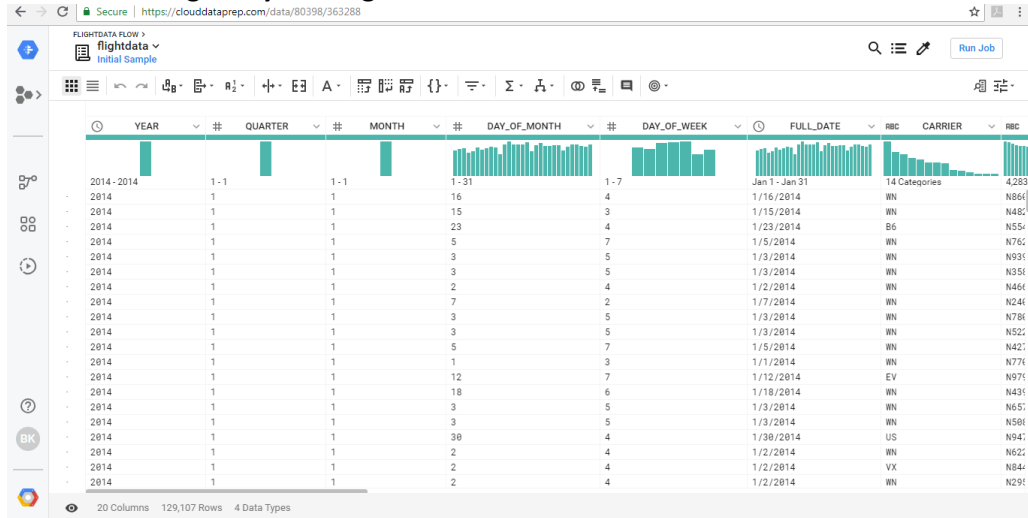
I built the diagram in the Google Cloud environment using the Lucidchart platform. I chose Google over Amazon due to its flexibility and efficiency.

- 3) Clearly state your assumptions.

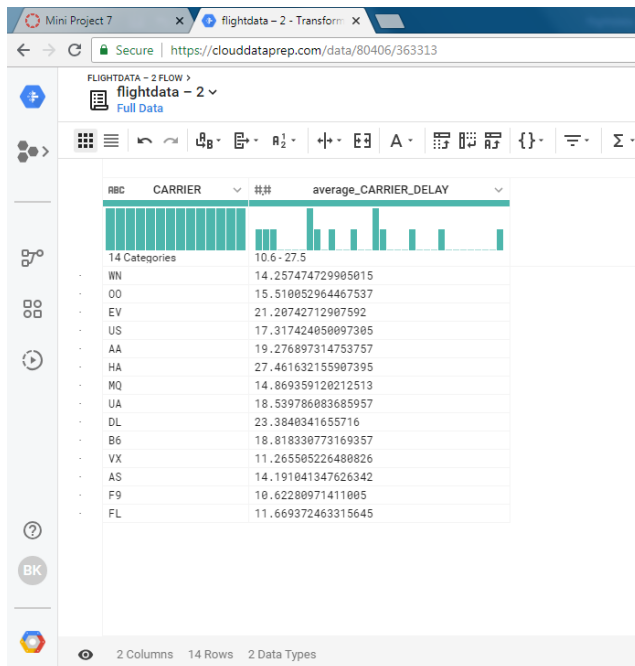
There are four main stages to the design. The first stage begins with ingesting the data using cloud storage, Google App Engine, and Cloud Monitoring. Stage two is processing the data into two workstreams. The first workstream handles the batch, untransformed data and the second workstream handles the streaming data that outputs real time data analytics. The third phase is storage. In this stage, we leverage Cloud Storage for handling files and Big Query for handling large tables. The last stage is analysis, where Apache Hadoop and Spark analyze the data via the Cloud SQL environment.

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.

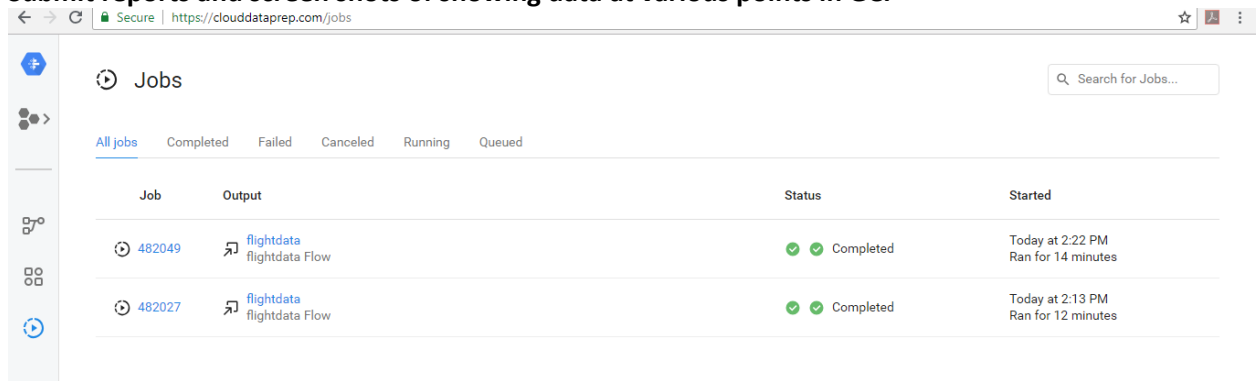
1) Load data in BigQuery in Google Cloud Platfrom



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



3) Submit reports and screen shots of showing data at various points in GCP



The screenshot shows the 'Jobs' page in Cloud Dataprep. It features a sidebar with navigation icons and a main content area with a search bar and tabs for job status. A table lists two jobs, both completed successfully.

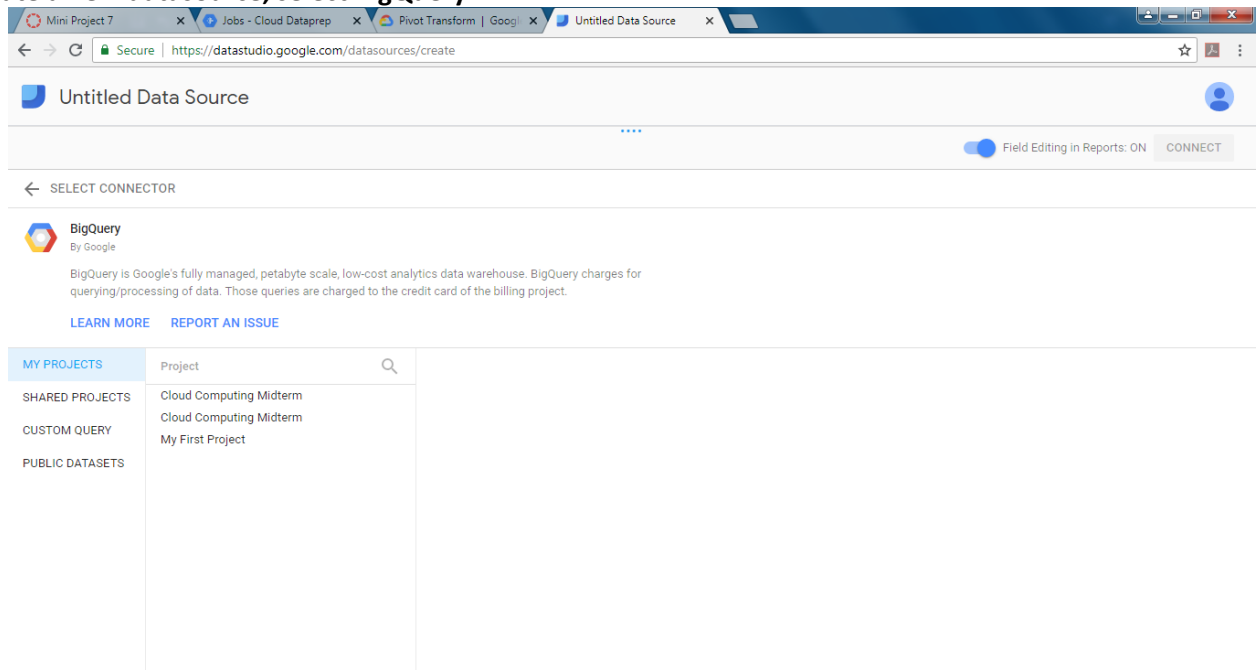
Job	Output	Status	Started
482049	flightdata flightdata Flow	Completed	Today at 2:22 PM Ran for 14 minutes
482027	flightdata flightdata Flow	Completed	Today at 2:13 PM Ran for 12 minutes

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



4) Select flight information table that you created in the previous question

flight_data

Field Editing in Reports: ON USING OWNER'S CREDENTIALS CREATE REPORT EXPLORE

EDIT CONNECTION ADD A FIELD

Index	Field	Type	Aggregation	Description
1	Record Count	123 Number	Auto	
2	CARRIER_DELAY	123 Number	None	
3	CARRIER	RBC Text	None	

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.

