UTILIZATION OF TESTING TOOLS

Software Requirement & Specification

BigQuery - Data Warehousing Solution Python - For combing the delay and cancellation CSVs gsutil - For uploading combined main dataset to GCS Google Data Studio - To visualize the insights obtained from data analysis Choosing BigQuery The combined data set for Delays and Cancellation was huge. Querying such large datasets with database management solution like MySQL or Postgres takes nearly a minute. In a professional environment, such large latency might decrease performance and efficiency.Therefore, a high throughput low latency solution i.e. BigQuery was adopted.

BigQuery is a serverless, highly scalable, and cost-effective data warehouse designed to help you turn big data into informed business decisions. [1]. BigQuery can analyse terrabytes of data with seconds.

Google Cloud Storage The overall size of combined dataset was 7 GB. Such large size data files cannot be uploaded directly to BigQuery.Therefore, GCS was was used as an intermediate data storage through which data was further load to BigQuery. The GCS bucket that was used : gs://airline-data.

The following command was used to upload the combined CSV file:

```
gsutil cp <source_to_CSV_dir>/combined_data.csv gs://airline-data
```

Using BigQuery The two datasets i.e. Delay & Cancellation and Global Airport Database had been stored in two separate tables in a single dataset under the GCP project. As the datasets were clean and difinite, the schema for both table was generated with Auto-Detect Schema feature of BigQuery