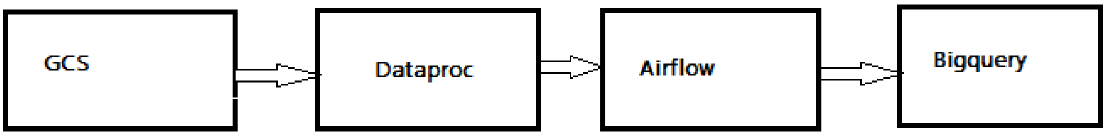
Bigquery lake project:



- the airlines sample data which contains details of travel made by employees. This file is also attached with the mail.
- Consider the sample data plced in your local.
- Build etl script in python scripting language to extract data and apply transformations and then load it in any open source database for this assignment

Transformations:

Following transformations should be applied :

1) All date column values should have uniform date format : dd-mm-yyyy

2) Pax_name field values should be cleaned . It has some places Mr. and MR or MRS . Remove all Prefix.

For eg: MR. ABG001 then it should be ABG001

3)Airline Field has data repetition .. Somewhere value is like Indigo/Indigo .. Where as it should be only Indigo.

4) For all the values First letter should be in caps and others in small.

For eg: Spicejet

5) For Airline Field name of Flights should be same For eg: if somewhere Indigo is written and another place Indgo then it should be proper and correct name – Indigo**.**

Composer Airflow:

-Automate the pyspark job using Airflow and schedule the job to run on daily basis

-Load the processed data to bigquery

-If any error occurs send an email alert using on_failure_callback function