Complete instructions and steps to run our DE project

Tools and files required

1) Download all the files from the provided google drive link [https://drive.google.com/drive/folders/1hh386ZfRX6DH05hrmgPU2y3-a_bWHvqJ?usp=sharing] containing folders airflow_intro and terraform_intro and other respective files
2) Ensure that terraform is installed in your laptop

Google cloud Setup:

3) Login in your google cloud account
4) Create a new project ID and a service account with all the required permissions and enable the required API's
5) Create a key json file and save it in the google folder of airflow_intro and copy the same file in the .google folder. [create google folder and .google folder under airflow_intro if not available]
   (please check the material file named Airflow Deep Dive under files section of teams for the google cloud setup)- https://bitsiserlohn.sharepoint.com/:b:/r/sites/msteams_90bdd5/Class%20Materials/Airflow%20Deep%20Dive.pdf?csf=1&web=1&e=6GdozZ, **slides 17-22**

Run terraform

6) Open the terraform folder and ensure only main.tf and variables.tf files are there in folder and update the following
   - Main.tf
     1) Replace the file path of the key json file in credentials variable (inside google folder in airflow_intro) with the one downloaded in your system in line 14
     2) Replace PROJECT ID name "arbaazdeproject" with your project ID name in line 46
   - Variables.tf
     1) Replace PROJECT ID name "arbaazdeproject" with your project ID name in line 7
     2) Replace JSON File name with your created json file's name in line 19
     3) Replace PROJECT ID name "arbaazdeproject" with your project ID name in line 31

7) Open command prompt from the terraform folder and run the following
- terraform init
- terraform plan
- terraform apply

- enter yes when asked and then it will create the google storage bucket and the template folder structure in the Bigquery using your created credentials

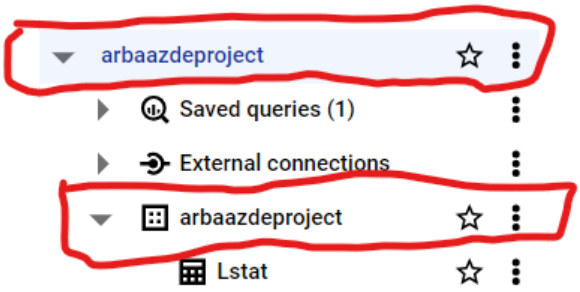
Airflow_intro folder

1) Replace the google cloud credentials with yours from line 56 to 60 in docker-compose.yaml file
   56- replace /opt/airflow/.google/arbaazdeproject-e42a28616ca0.json with /opt/airflow/.google/YOUR JSON KEY NAME
   57-replace google-cloud-platform://?extra__google_cloud_platform__key_path=/opt/airflow/.google/arbaazdeproject-e42a28616ca0.json withgoogle-cloud-platform://?extra__google_cloud_platform__key_path=/opt/airflow/.google/YOUR JSON KEY NAME
   58- replace arbaazdeproject with YOUR PROJECT ID
   59 - replace demo_data_lake_arbaazdeproject with your bucket name(check in google cloud if required)
   60 - replace arbaazdeproject with your bigquery dataset name (check in google cloud if required)

2) Open great_expectations_bigquery.py
   - Replace arbaazdeproject with your bigquery dataset name in line 25
   - Replace arbaazdeproject with your bigquery table name in line 26 (check in google cloud if required)

3) Open config_variables.yml under resources folder
   - bigquery://arbaazdeproject/arbaazdeproject with bigquery://YOUR PROJECT ID/YOUR BIGQUERY DATASET NAME [line 2]

4) Open demo_taxi_fail_chk.yml and demo_taxi_pass_chk.yml under checkpoints folder in ge (inside config folder)
   - Replace arbaazdeproject.arbaazdeproject with YOUR BIGQUERY NAME.DATASET ID in both the files mentioned above [line 26]



Execution

1) Follow the steps mentioned in the README.MD file inside the airflow_intro (ensure docker is opened)

```
docker-compose build
docker-compose up airflow-init
docker-compose up
```

2) After running and checking the docker (if airflow scheduler is running) , open the local host
3) Enter the login credentials as below in the local host
   username- airflow
   password- airflow
4) Run the great_expectations_bigquery.py pipeline

Further steps

1) Check the dataset table being created in the bigquery dashboard of the google cloud
2) After this , open the DE project sql queries.doc from the zip folder and copy all the queries and run them separately in bigquery which will create more tables
3) The created tables and the main dataset table are used for building visualizations and dashboard in the lookerstudio.
4) The visualization and the dashboard files (in pdf) are present in the given google drive to view.

NOTE :
For viewing the visualization on the looker studio, we need to provide permission access. Therefore, please contact us over teams by sharing your email ID in order to view the dashboard and the other visualizations.

We humbly request you guys to contact us if you find any difficulty in running our project.

Thanks and we are looking forward to hearing from you.

Group 18 -
Chetan Harshal Tote
Joshil Fernandes
Soham Sanjay Vaidya
Surabhi Kailas Sangore
Arbaaz Khaja Qutubuddin
Carolyn Gundimi
Aju Thomas
Devarsh Rajesh Bende
Omama Mashhood Ur Rahman and Kiran Kumar Pinjare