GHCN-D Data Pipeline

NOAA Global Historical Climatology Network (Daily)

- Project and dataset
- Solution
- Setup project and IaC
- Data ingestion
- Data transformation (dbt)
- Data transformation (Dataproc/Spark)
- Dashboard

Project and Dataset

Repository:

https://github.com/MarcosMJD/ghcn-d

https://github.com/MarcosMJD/ghcn-d.git (clone)

Dataset: NOAA Global Historical Climatology Network (Daily)

https://registry.opendata.aws/noaa-ghcn/

https://noaa-ghcn-pds.s3.amazonaws.com/index.html (browse)

Countries (txt)

```
218 ZA Zambia
```

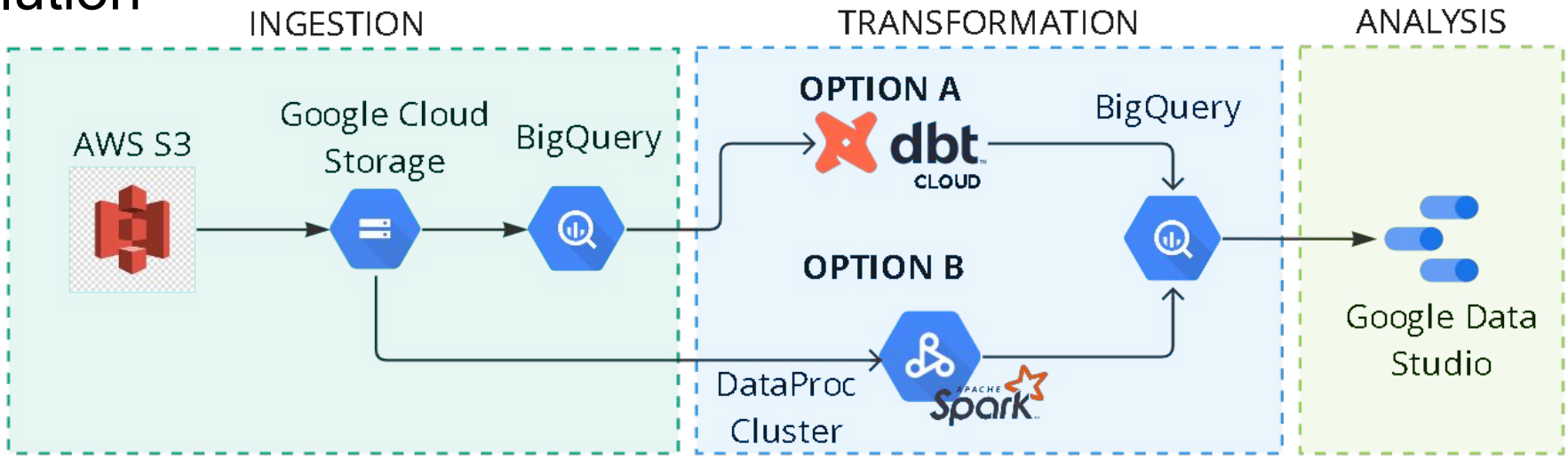
Stations (txt)

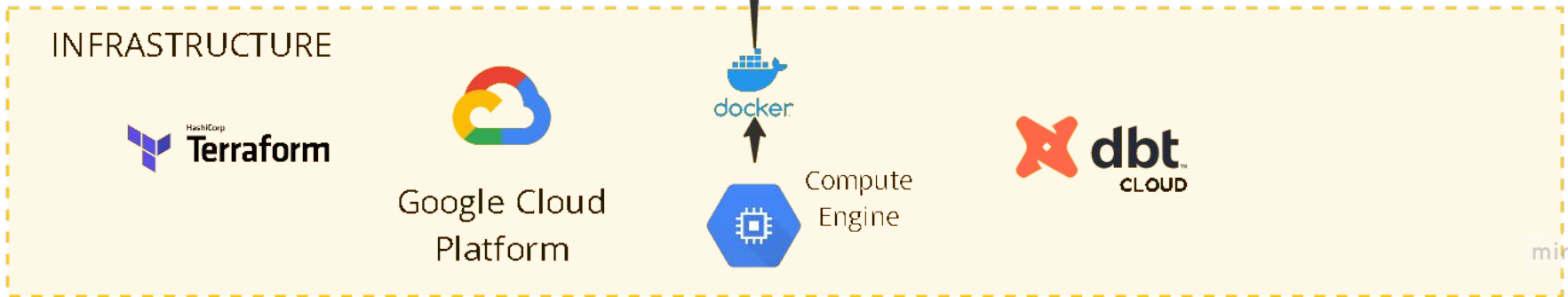
```
117838 ZI000067983 -20.2000   32.6160 1132.0    CHIPINGE                       GSN      67983
```

Year Observations (csv). e.g. 2007.csv
(>1GiB in size)

```
37241182 ZI000067775,20071231,TMAX,223,,,S,
37241183 ZI000067775,20071231,TMIN,111,,,S,
37241184 ZI000067775,20071231,PRCP,0,,,S,
```

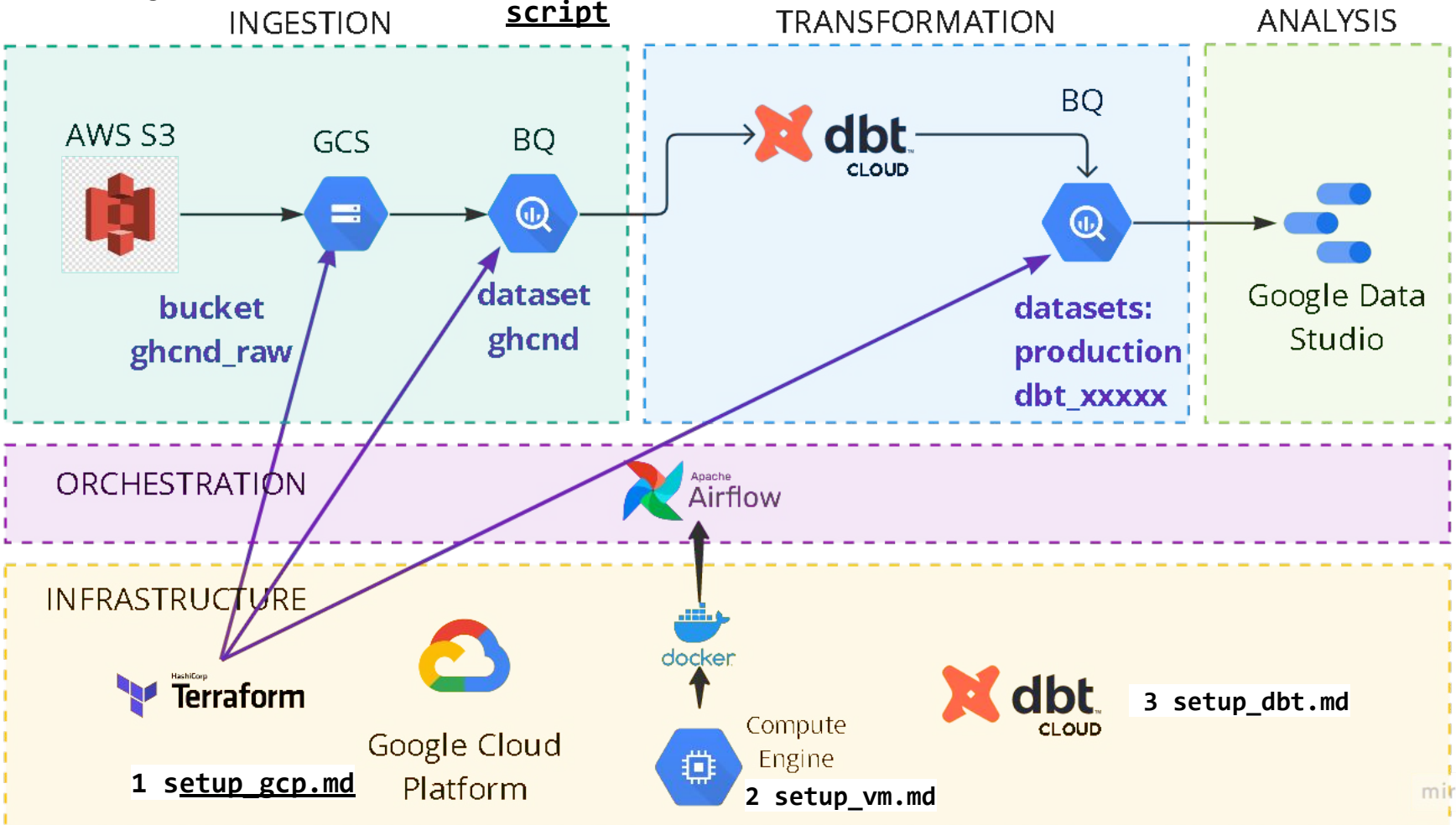
Solution



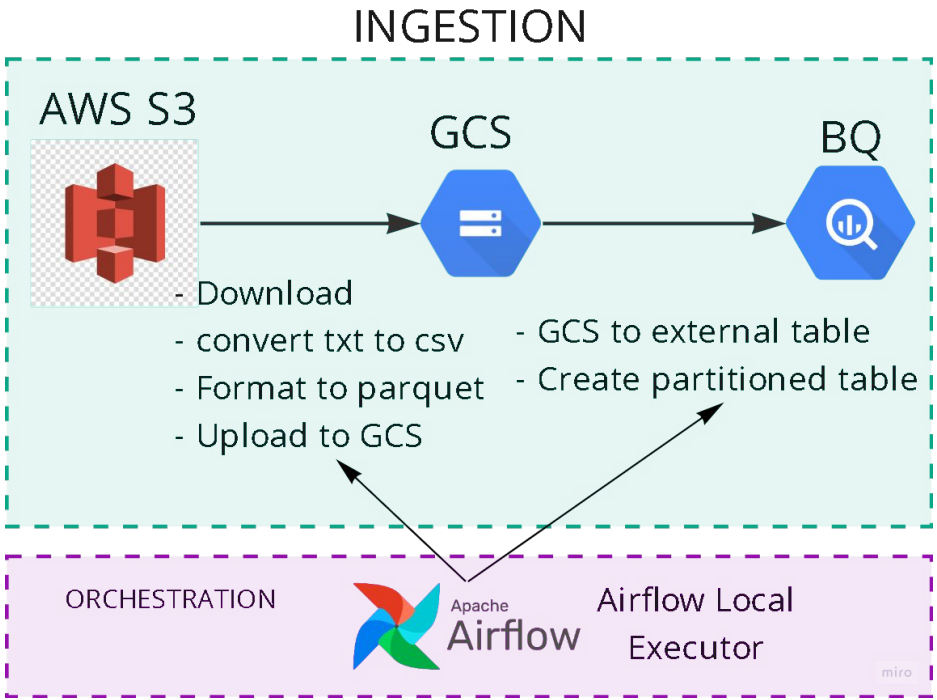


Setup project and IaC

4. setup_project.sh script

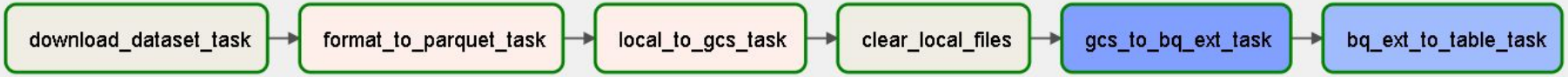


Data ingestion pipeline



Data ingestion pipeline

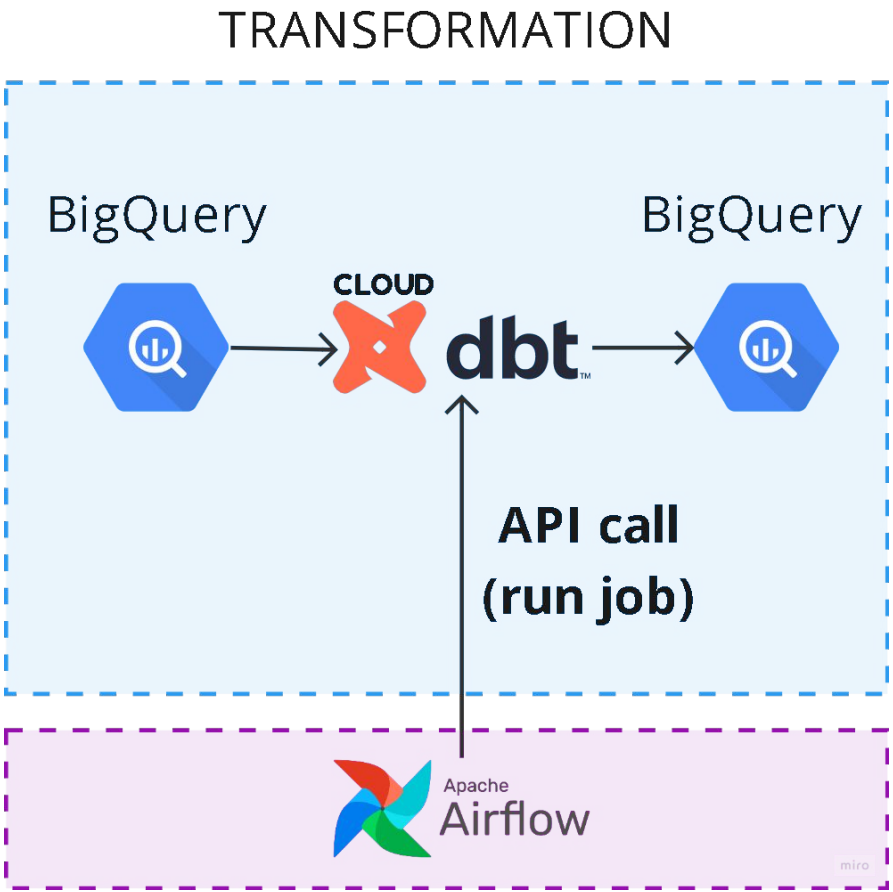
- **‘year of observations’ ingestion pipelines**: **‘PAST_YEARS’** and **’CURRENT_YEAR’**



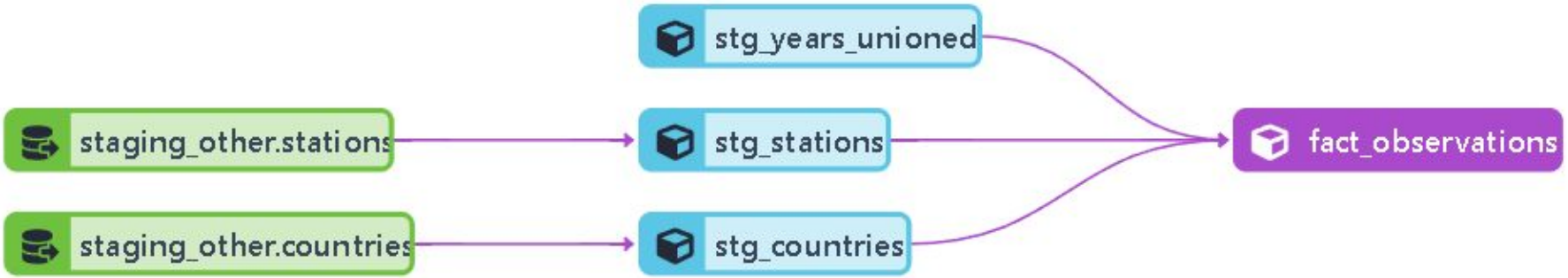
Partitioning and clustering: Observations tables (one per year) partitioned by date of observation and clustered by station

- **countries and stations ingestion pipeline**: **aws_gcs_other_datasets_dag**

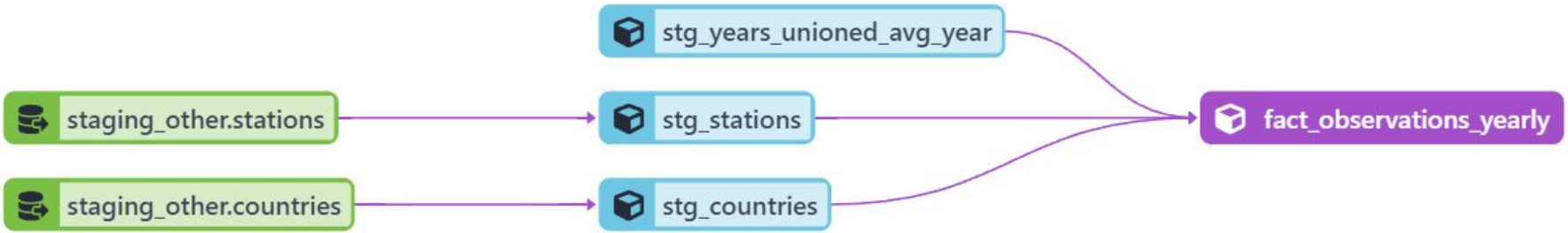
Transformation (dbt)



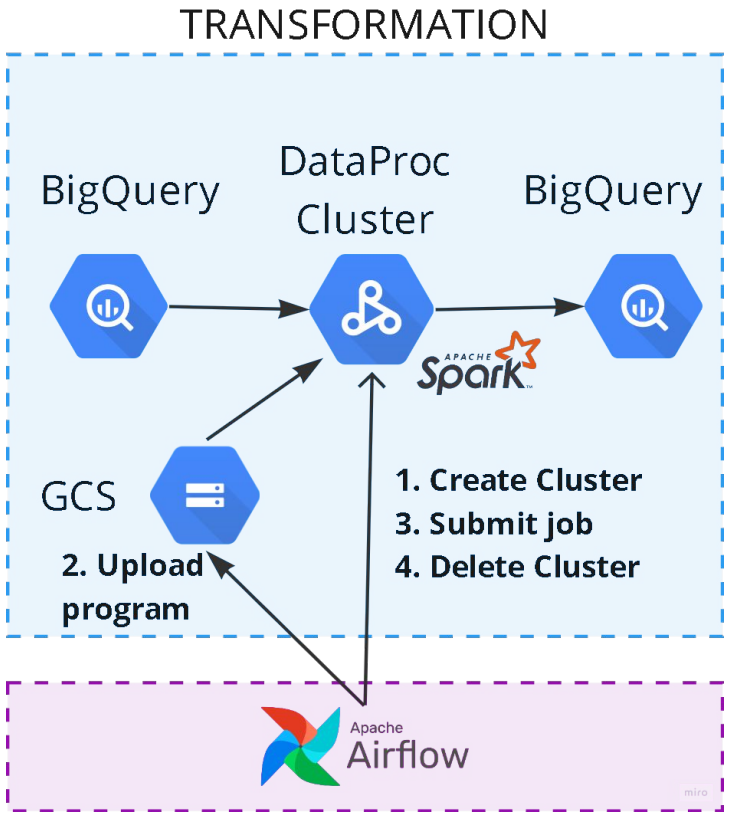
Transformation (dbt)

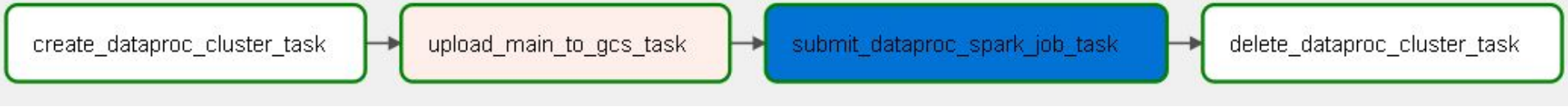


Transformation (dbt)



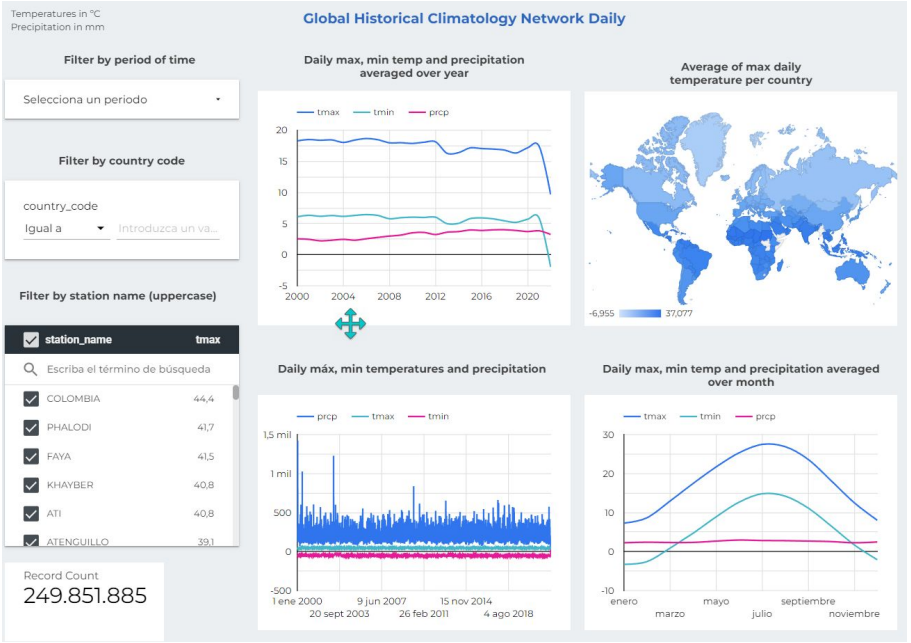
Transformation (DataProc/Spark)





Improvements

- Incremental solution in dtb to avoid the limitation of max character limit in sql query when creating the facts_observations models.
- CI/CD.
- Tests and documentation.
- Performance.
- Cost analysis.
- Security.
- Fix some bugs.
- Explore other cloud providers.



Thanks!

https://github.com/MarcosMJD