Task

Continue with the weather data pipeline using PySparks in Databricks. It is recommended to structure the project properly creating separate notebooks for separate modules. Link

Pre-requisites

- Ensure you have completed the previous tasks, and that the data is loaded to the cleaned tables

Tasks

1. Create the required dimension and fact tables
    a. Generate / load data into the dimension tables (minimum tables required are for location (city), date, and time (hour))
2. Modularize your codebase from the previous task and separate out different functionalities into different notebooks (e.g. logger -> related to recording the logs, weather_api -> related to data extraction from OpenWeatherMap, so on...)
3. Add new modules to load data into the fact (reporting) tables
    a. The pipeline should have minimum two fact tables: one with hourly weather data and another with daily weather data

*In case you have issues with retaining data from the tables in the **Databricks Community Edition**, export your data to csv files, and save them locally.*

Then load the data from these files to the table the next time you start up the cluster.

You may need to do this for the raw data and the log data only as the cleaned / fact tables can be generated using the raw data.