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 have a pipeline in place to load the data from source to destination fact tables
- This task requires the weather data of at least a few hours (5-6 hours), so ensure you have adequate data
    - To load the required data, you can run your pipeline 5-6 times manually per hour
    - In case of issue with data not persisting on cluster restart (in the Databricks community edition), please make sure to export the data from your tables to csv files so that you can load it manually next time
    - If you have the previous Pentaho pipeline intact, you can also schedule the pipeline to save data for the required city ids, export the data and load it to your delta tables

Tasks

1. Update your hourly fact weather table to include a new column '*is_forecasted_data*' (or similar name)
    a. This should be of Boolean Type or Integer Type
    b. This column will be used to separate the real data from calculated / forecasted data

2. Add a new module to forecast weather data using the existing data from the hourly fact table
    a. Forecast data for the next 5 hours using the average of last 4 or 5 hours each time
    b. Forecast the temperature data only for now (add the other measures later when you can)
    c. The forecasted data should be loaded to the same hourly fact table

*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.