Challenges Faced During the Project

Throughout the course of this project, we encountered several challenges related to working with Apache Spark, Delta Lake, Hive external tables, and the data integration process. This section outlines the key difficulties we faced and the approaches taken to address them.

# 1. Spark Dependencies and File Path Issues

Working locally with Apache Spark introduced several issues regarding dependencies and file paths, especially when using Delta-Spark and creating Hive external tables. We had to manually download required files and place them in specific locations for the setup to function correctly. This required additional research and troubleshooting, and we still encountered a number of conflicts during the process.

# 2. Hive Error During Consecutive Pipeline Runs

Running the three pipelines consecutively often resulted in Hive errors. For some reason, it was necessary to restart the notebook before starting another pipeline. This introduced inefficiencies and required manual intervention to ensure the correct execution of each pipeline.

# 3. Efficient Implementation of Slowly Changing Dimensions (SCD)

Implementing Slowly Changing Dimensions (SCD) efficiently posed challenges, particularly due to the large number of tables involved. Initially, the first run of the SCD process took around 50 minutes to complete. To optimize this, we introduced a new hash column that stores the hash of the key columns. This optimization reduced the execution time to less than a minute.

# 4. ID Creation for Integrated Data Between NYC and Random Data

The process of generating new IDs for the flat table in the integrated database, which merged the NYC dataset with Faker-generated random data, required logical problem-solving. The solution involved using the DENSE\_RANK function on unique columns, which allowed us to efficiently assign new IDs.

# 5. Installing dependencies

There were issues with installing pandas and scikit-learn due to dependency conflicts. The solution was to create a virtual environment and use other tools.

# 6.Data corruption

Some data was corrupted during the gathering process, which caused delays in preprocessing and misleading in the model training.

# 7. External Factor

Time constraints and external factors impacted the pace of development.

# 4. Deployment error

Facing several deployments error in databriks while creating three notebooks (ml-to-deploy , deploy-to-azure and azure-to-predict). These error are related to some libraries installation, workspace creation, giving the access to the workspace, deploy the model, some logical error in scoring.py file and error in deploying the model to