Shreya Nagabhirava
DS 2002

Process

1. First I imported the required libraries
2. Instantiate global variables
3. Define global functions
    a. Get_sql_dataframe
    b. Get_mongo_dataframe
    c. Set_mongo_collection
4. Populate dimensions by ingesting reference data
    a. Fetch the reference data from an azure sql database
        i. Create a new databricks metadata database
        ii. Create a new table that sources its data from a view in an azure sql database
        iii. Create a new table that sources its data from the other table in my azure sql database
    b. Fetch reference data from a mongoDB atlas database
        i. View the data files on the databricks file system
        ii. Create a new MongoDB database and load json data into a new MongoDB collection
        iii. Fetch data from the new MongoDB collection
        iv. Use the spark dataframe to create a new table in the databricks metadata database
        v. Query the new table in the databricks metadata database
5. Fetch data from a file system
    a. Use pyspark to read from a csv file
    b. Verify the dimension tables
6. Integrate reference data with real-time data
    a. Use autoloader to process streaming data
    b. Process the raw JSON data
    c. Perform aggregations