DS3002 – Data Project 2 (Course Capstone)

Demonstration of requirement fulfillment

**Requirements:**

Your solution (database schema) needn’t be complex, but should meet the following requirements:

* Your solution must include a **Date dimension** to enable the analysis of the business process over various intervals of time *(the code for creating this in MySQL has already been provided for you).*
  + *In the Jupyter notebook fact table*
* Your solution must include at least 3 additional dimension tables (e.g., buyers, sellers, products)
  + *Historic, Sector, and Price tables created in both Databricks and Jupyter notebooks*
* Your solution must include at least 1 fact table that models the business process
  + *Fact table in both Databricks and Jupyter notebooks*
* Your solution must populate its dimensions using data originating from multiple sources:
  + A relational database like MySQL, Oracle or SQL Server
    - *Imported data to Azure SQL Server and queried it in Databricks*
  + A NoSQL database like MongoDB, Redis, Cassandra or HBase
    - *Imported data to MongoDB and retrieved it in Databricks*
  + An API that returns a message payload (e.g., JSON, CSV, text)
    - *Accessed API data through “AlphaVantage”*
* Your solution must integrate datum of differing granularity (static and near real-time)
  + *Static data from Kaggle CSV, real-time data from “AlphaVantage” API*
* Your solution must include one or more visualizations that demonstrate the business value of your solution. For example, a “dashboard” developed using Excel, Power BI, Tableau or other data visualization tool capable of demonstrating the use of PivotTables and/or Pivot Charts
  + *Pandas bar chart visualizations Jupyter notebook*

**Benchmarks:**

1. Your solution must demonstrate at least one additional batch execution (i.e., provide some sample source [SQL & NoSQL] data to demonstrate loading at least one incremental data load).
   1. *Function created to impute incremental data in Jupyter notebook*
2. Your solution must demonstrate accumulating data that originates from a real-time (streaming) data source for a predetermined interval (mini-batch), integrating it with reference data, and then using the product as a source for populating some aspect of your dimensional data mart. (i.e., implement something like the Databricks bronze, silver, gold architecture).
   1. Your solution must demonstrate the integration of streaming data for at least 3 intervals.
      1. *Three intervals five minutes apart in Jupyter notebook*
   2. Your data visualization(s) need NOT reflect the integration of data in real-time.
3. You must submit all SQL code, including all data definition and data manipulation statements.
   1. *In Databricks notebook*
4. You must submit all reference data used to populate the source databases, JSON/CSV files, etc.
   1. *MMM.JSON file, other data taken from cloud*
5. You must submit all Python code needed to implement data integration, and any object creation.
   1. *Jupyter notebook and python in Databricks*
6. You must submit all data visualization source files (e.g., Excel workbook, Power BI workbook).
   1. *N/A*
7. You must submit screen-grabs of your data visualization(s)
   1. *2 screenshots of pandas graphs*
8. Please submit all code, and other artifacts, in a standalone GitHub repository in your account.

*If you opt to use any cloud-hosted services then please identify them so we may faithfully replicate your project.*

*Submitted in Github repo at my Github site* *https://github.com/kmrowed/ds3002\_final*