DS3002 – Data Project 2 (Course Capstone)

25 points
Due 4/12/2022 11:59:59

The goal of the second data project, building upon the first project, is to further demonstrate (1) an understanding of and (2) competence creating and implementing basic data science systems such as pipelines, scripts, data transformations, APIs, databases and cloud services. Submit your project in your GitHub Repo or file drop on Collab. **Data Projects must be done individually.**

Putting it All Together: Data Integration & Analysis

Deliverable: Design and populate a dimensional data mart that represents a simple business process of your choosing. Examples might include retail sales, inventory, procurement, order management, transportation or hospitality bookings, medical appointments, student registration and/or attendance. You may select any business process that interests you, but remember that a dimensional data mart provides for the post hoc summarization and historic analysis of business transactions that reflect the interaction between various entities (e.g., patients & doctors, retailers & customers, students & schools/classes, travelers & airlines/hotels).

You project should demonstrate your understanding of the differing types of data systems (OLTP/OLAP), and how data can be **extracted** from various source systems (structured, semi-structured, unstructured), **transformed** (cleansed, integrated), and then **loaded** into a destination system that's optimized for post hoc diagnostic analysis. Your project should also demonstrate your knowledge of data integration patterns like ETL, ELT and ELTL, and architectures (e.g., lambda or kappa) for integrating batch and real-time (streaming) data sources.

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).*
- Your solution must include at least 3 additional dimension tables (e.g., buyers, sellers, products)
- Your solution must include at least 1 fact table that models the business process
- Your solution must populate its dimensions using data originating from multiple sources:
  - A relational database like MySQL, Oracle or SQL Server
  - A NoSQL database like MongoDB, Redis, Cassandra or HBase
  - An API that returns a message payload (e.g., JSON, CSV, text)
- Your solution must integrate datum of differing granularity (static and near real-time)
- 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

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).
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).
   a. Your solution must demonstrate the integration of streaming data for at least 3 intervals.
   b. 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.
4. You must submit all reference data used to populate the source databases, JSON/CSV files, etc.
5. You must submit all Python code needed to implement data integration, and any object creation.
6. You must submit all data visualization source files (e.g., Excel workbook, Power BI workbook).
7. You must submit screen-grabs of your data visualization(s)
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.*

Note: You may utilize any combination of on-premises and/or Cloud service technologies. For example, you can collect streaming data from a source API on the Internet, integrate it with reference data that's stored in another Cloud hosted database service (e.g., Mongo DB Atlas) using Databricks, and then load it into your dimensional data mart that's hosted on your laptop. Alternatively, you may choose to host your entire solution in the Cloud.

Grading:

- Successful deployment – 10 points.
- Functionality that meets all benchmarks – 12 points.
- Documentation – Describe your process, code, deployment strategy – 3 points.

Publicly-available sample databases:

- https://dataedo.com/kb/databases/mysql/sample-databases (Sample MySQL databases)
- https://docs.microsoft.com/en-us/sql/samples/sql-samples-where-are?view=sql-server-ver15 (Microsoft SQL samples)

Publicly-available datasets:

- https://www.kaggle.com/datasets
- https://data.world/
- https://www.data.gov/
- https://opendata.charlottesville.org/

Publicly-available APIs:

- https://docs.github.com/en/rest
- https://developer.twitter.com/en/docs/twitter-api
- HUGE LIST: https://github.com/public-apis/public-apis