DS-2002 – Data Project 1

100 points

The goal of this project is to 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.

ETL Data Processor

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.

Deliverable:

1. **Design a dimensional data mart that represents a simple business process of your choosing.**
    a. Examples might include retail sales, inventory management, procurement, order management, transportation or hospitality bookings, medical appointments, student registration and/or attendance.
    b. 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).

2. **Develop an ETL pipeline that extracts, transforms, and loads data into your data mart.**
    a. Extract data from one or more SQL database tables; hosted locally or in the Cloud.
    b. Retrieve a data file, either from a remote or local file system, converting its original format (e.g., CSV, JSON) into a SQL database table.
    c. Modify the number of columns from each source to the destination.
    d. Provide error messages wherever an operation fails (i.e., Try/Except error handlers).

3. **Author one or more SQL queries (SELECT statements) to demonstrate proper functionality.**
    a. SELECT data from ***at least*** 3 tables (two dimensions; plus the fact table).
    b. Perform some type of aggregation (e.g., SUM, COUNT, AVERAGE). This, of course, necessitates some form of grouping operation (e.g., GROUP BY *<customer.last_name>*).

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 2 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 use data originating from at least 3 of the following sources:
    - A relational database like MySQL, Oracle or SQL Server
    - A NoSQL database like MongoDB, Redis, Cassandra or HBase
    - A file system (either local or remote).
    - An API that returns a message payload (e.g., JSON, CSV, text)

Benchmarks:

1. You must submit all data used to populate the data mart (source databases, JSON/CSV files, etc.)
2. You must submit all SQL code, including all data definition and data manipulation statements.
3. You must submit all Python code needed to implement data integration, and any object creation.

    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.

Grading:

- Successful deployment – 40%.
- Functionality that meets all benchmarks – 50%.
- Documentation – Describe your process, code, deployment strategy – 10%.

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