DS3002 Data Project 1
25 Points

The Goal of this project is to demonstrate (1) an understanding of and (2) competence of implementing and using basic data science systems rooted in SQL and other data sources like flat files (CSV), Open Data and other relational and data sources as well as APIs and data transformation. For this project you will use GitHub to store and manage your code. You can get a crash course in Git here →

ETL data processor

1. Deliverable: Author a segment of an ETL pipeline that will ingest or process raw data. You must also submit a URL to a GitHub repository for your solution. In python you'll need to know how to open files, iterate files, pattern match and and output files.
2. Benchmarks:
    i. Your data processor should be able to ingest a pre-defined data source and perform at least three of these operations:
        1. Fetch / download / retrieve a remote data file by URL, S3 key, or ingest a local file mounted. Suggestions for remote data sources are listed at the end of this document.
        2. Convert the general format and data structure of the data source (from TSV to CSV, from CSV to JSON, from JSON into a SQL database table, etc.) EXTRA – Use and API (like twitter) to pull information realtime.
        3. Modify the number of columns from the source to the destination, reducing or adding columns.
        4. The converted (new) file should be written to disk, or pushed to S3, or written to a SQL database.
        5. Generate a brief summary of the data file ingestion including:
            1. Number of records
            2. Number of columns
    ii. The processor should produce informative errors should it be unable to complete an operation.
3. Grading:
    i. ☐ Successful build of the solution (I recommend Python…but you can use whatever)
    ii. ☐ Functionality that meets all benchmarks – 10 points
    iii. ☐ Creativity / Innovation / Quality – 2 points
    iv. ☐ Documentation – Describes how to use the data processor and the elements that make it operational – 3 points

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