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.

This will be due on March 21th at 11:59 PM. Submit it to Git, copy the invite /link to me and Isabelle.

**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 output files.
2. Benchmarks:
   1. 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, 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 JSON to CSV, from CSV to JSON, from JSON into a SQL database table, etc. I want the option to convert any source to any target. So, if I get a CSV as an input, I want the user to choose an output)
         1. **EXTRA** – Use an 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 (local file) 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
   2. The processor should produce informative errors should it be unable to complete an operation. (Try / Catch with error messages)
3. Grading:
   1. o Successful build of the solution (I recommend Python…but you can use whatever you like. I just need to be able to run it)
   2. o Functionality that meets all benchmarks – 10 points
   3. o Creativity / Innovation / Quality – 2 points
   4. o 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/>  
You can Choose/find data from anywhere you like…these are just suggestions.

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