Assignment 5.2: Final Team Project Instructor Feedback Submission

Instructions:

At this point in the course, you should have a pretty good start on your final project. For this assignment, by the end of Day 7, you will submit a written (1-2 paragraph) summary/status of where your team is with the final project, any issues you have been having, and/or questions you may have that will help you complete the project. Cite resources and references appropriately in APA format.

It should also include a link to your project (GitHub) and any instructions for setting it up. Your instructor will provide feedback as to improvements or changes that might help make this project better. In the final submission in Module 7, there will be a section that looks at how well you were able to receive and incorporate the provided feedback into your final submission.
Only one member of your team needs to submit this assignment on the group's behalf.

Team 2 Final Project Response:

GitHub Repo: https://github.com/amcarr-ds/ads507_data_engineering
See the README "Getting started" for instructions on how to clone the repo.

Status Updates

Team 2 has made significant progress on the conceptualization of the pipeline, including meeting the following milestones:

- Confirmed and agreed upon all the datasets that will be included.
- Performed transformations on datasets (e.g., joins, common relationships/column(s), aggregating ,etc.)
  - UPDATE statements were needed to convert values to standard types
- Developed methods for getting the data into a SQL schema and building the Python methods for integration with MySQL.
- As we are using static datasets, uploading .csv files, our pipeline type will be ETL (extract, transform, and load) specifically.
- Established pipeline segments: API (Kaggle), transform using Python (connected to SQL), stored to GitHub/MySQL schema.
- The team has designated tasks, responsibilities, and goals to meet for the final project as stated in the requirements.
- Agreed upon applications (VS Code, Jupyter Notebook, MySQL Workbench)

Issues

While conceptualization of the overall pipeline and creating of individual ETL components have been progressing well, how to best link the E, T, and L (i.e., moving data via automatic methods) continues to be something Team 2 is grappling with.

- Still exploring best ways to implement triggers (e.g., investigating CRON jobs through Python).

Questions

We are seeking additional clarification on some of the final project requirements:

- How do you monitor the pipeline while it is running?
    - What exactly would be monitored (i.e., computational resources, alerts, database performance, etc.)?
    - Monitoring the performance query and/or if the pipeline breaks?
    - Is this a statement to show the different methods on how to in a theoretical sense?
- Gaps in system: Is the system secure?
    - How would we state whether or not the system is secure? Is it based on where this is being loaded or extracted to?

References

Banerjee, S. (2022, October 20). World population dataset. *Kaggle*. Retrieved February 4, 2023, from https://www.kaggle.com/datasets/iamsouravbanerjee/world-population-dataset

Sy, S. (2022, January 10). Temperature change. *Kaggle*. Retrieved February 4, 2023, from https://www.kaggle.com/datasets/sevgisarac/temperature-change

The Devastator. (2023, January 23). Emissions by Country. *Kaggle*. Retrieved February 4, 2023, from https://www.kaggle.com/datasets/thedevastator/global-fossil-co2-emissions-by-country-2002-2022