Company Context

COVID-19 has created a new social, political and economic world order – one that is increasingly driven by cultural narratives and online discourse. NWO's bleeding-edge technology enables clients to surface the fears, motivations and demand drivers underlying various trends, providing them with unprecedented access to the why behind a narrative. NWO's platform is already in use by Fortune 500 brands to empower key resource allocation decisions.

Background

At nwo.ai, we are always looking to add proprietary data streams that enrich our dataset and give us access to insights that were previously not accessible. Since the Cambridge Analytica debacle, it has become exponentially more difficult to access certain data APIs that were previously accessible. At nwo.ai, we do not wish to use the data with malicious intent. Our technical mission is to leverage data to provide predictive signals.

The data platform team is focused on developing the best-in-class alternative data warehouse that powers our microtrend engine. The following case represents some of the foundational responsibilities expected for data engineers.

Instructions

- Review this document in its entirety. Immediately ask any questions that come up.
- Begin working on the case. You have up to 1 week to complete the case.
- At no later than 1 week (7 days) from the start time, email back the entire set of case files and the files listed in the Output section.
- The final submission should be made available to us as a GitHub repository (email us a link)

Points to Consider

- If you have questions, you may email **brian.h@nwo.ai**. We'll do our best to

respond ASAP. However, if we do not respond, **then feel free to make a reasonable assumption. State this assumption in your case report/summary**.

- Please do not discuss this case with anyone else. However, you may use any Internet resources for syntactical assistance only.
- Please give nwo.ai 3-5 days to evaluate case study submissions
- Feel free to provide detailed decision points in light of different scenarios and business conditions.

Scenario

You are working for nwo.ai as a Data Engineer on the Data Platform Team. Your objective is to acquire and operationalize data relating to the food and beverage industry. You have been tasked with architecting and developing an ETL pipeline that ingests Reddit posts and comments and associated metadata from relevant subreddits.

Design Constraints

- Python 3.8-3.10
- ETL solution should be scalable are able to run at any cadence
- All raw data files should be stored as json lines files (.jsonl) on your local filesystem
- Pytest for unit testing
- The production ready data should be hosted in a local instance of MongoDB
    - https://www.mongodb.com/docs/manual/tutorial/install-mongodb-community-with-docker/

Requirements

- You will receive weekly dumps of reddit submission data via a downloadable url
    - The latest file is found here
- Develop an ETL solution that can ingest this data weekly
    - You can use any known framework or structure to create the pipeline
- Develop an ETL pipeline that extracts raw data, performs necessary transformations and loads into storage

Deliverables

- Python source code: python class files and scripts containing both ETL pipelines. Including:
    - Documentation
    - Logging

    - Testing
- Docker, commands, and/or config files needed run and initialize your database
- BONUS - System architecture diagram and ERD. Be prepared to discuss the following:
    - Roadmap to operationalize your system in a production environment
    - Benefits and tradeoffs of your architecture (be specific with your choice of technologies)
    - Your choice of schema design pattern (i.e. STAR, 3NF, Snowflake) and why you believe it was appropriate for this scenario

Evaluation Criteria

- Efficacy - Is your solution effective? Does your code run (error / bug free)? Does it accomplish the task outlined above?
- Efficiency - Is your approach computationally efficient? Would it scale to a much larger corpus?
- Elegance - Is your code clear and legible, well constructed and commented? Is the thought process behind the approach clear? Would it be easy for another team member to work with?