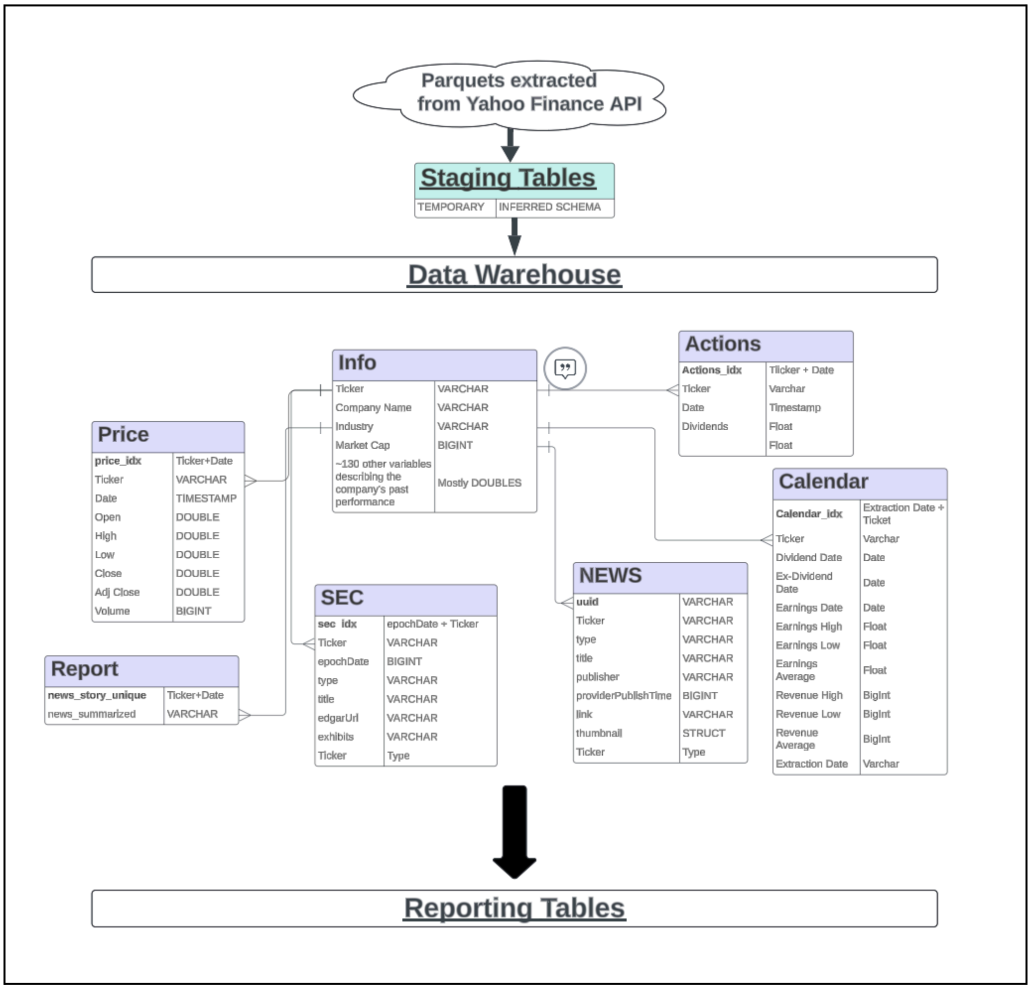
Team Project Phase 2 - Team 7

Objective: To develop a unified, automated dashboard that consolidates real-time stock market data for user-selected stocks, updating daily after market close. The dashboard incorporates sentiment analysis to classify the latest news as Positive, Negative, or Neutral, enabling investors to make more informed decisions. Additionally, it leverages Generative AI (GenAI) to summarize and analyze critical market information, including SEC filings, providing users with actionable insights and a comprehensive overview of market trends.



Info	
Ticker	VARCHAR
Company Name	VARCHAR
Industry	VARCHAR
Market Cap	BIGINT
~130 other variables describing the company's past performance	Mostly DOUBLES

Actions	
Actions_idx	Ticker + Date
Ticker	Varchar
Date	Timestamp
Dividends	Float
	Float

Price	
price_idx	Ticker+Date
Ticker	VARCHAR
Date	TIMESTAMP
Open	DOUBLE
High	DOUBLE
Low	DOUBLE
Close	DOUBLE
Adj Close	DOUBLE
Volume	BIGINT

Calendar	
Calendar_idx	Extraction Date + Ticket
Ticker	Varchar
Dividend Date	Date
Ex-Dividend Date	Date
Earnings Date	Date
Earnings High	Float
Earnings Low	Float
Earnings Average	Float
Revenue High	BigInt
Revenue Low	BigInt
Revenue Average	BigInt
Extraction Date	Varchar

NEWS	
uuid	VARCHAR
Ticker	VARCHAR
type	VARCHAR
title	VARCHAR
publisher	VARCHAR
providerPublishTime	BIGINT
link	VARCHAR
thumbnail	STRUCT
Ticker	Type

SEC	
sec_idx	epochDate + Ticker
Ticker	VARCHAR
epochDate	BIGINT
type	VARCHAR
title	VARCHAR
edgarUrl	VARCHAR
exhibits	VARCHAR
Ticker	Type

Report	
news_story_unique	Ticker+Date
news_summarized	VARCHAR

Phase 1 Overview:

In Phase 1, we developed a unified dashboard for tracking the top 10 U.S. stocks, updating automatically after market close. The data pipeline was designed to ingest raw data from the Yahoo Finance API, process it into parquet files, and load it into staging tables using Cloud Run functions. These data sets were transformed and upserted into a cloud data warehouse, with reporting enabled through Superset. Prefect was employed to orchestrate the workflow, ensuring reliable and scalable daily updates. Additionally, we tested an analytics pipeline leveraging web scraping and GPT-4 to summarize news articles, offering actionable insights on stock performance. Our development practices were streamlined through GitHub for version control and efficient task management.

Team Project Phase 2 - Team 7

Sentiment Analysis:

In Phase 2, we introduced sentiment analysis to classify textual data into Positive, Negative, and Neutral categories, helping users quickly assess the sentiment of news articles and SEC filings for better investment decisions.

Model Design and Training:

We developed a sentiment analysis model pipeline using supervised learning techniques. The pipeline combines text preprocessing (tokenization, stopword removal, and vectorization using TF-IDF) with logistic regression for classification. Initially, the model was trained on a sample Kaggle dataset to validate its functionality and refine the pipeline before applying it to our custom dataset.

Serverless ML on Sample Dataset:

The initial training and deployment were done using Cloud Run functions:

- **ml-postwc-train-sentiment**: Automates training on a sample dataset, saves the model as a .joblib file, and uploads it to Google Cloud Storage (GCS).
- **ml-postwc-serve-sentiment**: Serves the trained model to perform real-time sentiment classification on incoming textual data.

Next Steps for Sentiment Integration:

While these models currently operate independently, we plan to fully integrate them into our pipeline. This involves training on the complete dataset (news and SEC filings) and automating sentiment classification during daily Prefect runs. The resulting sentiment scores will be seamlessly incorporated into reporting tables for visualization in Superset.

Pipeline Design:

For the current phase, we developed ML functions for sentiment analysis. Although these functions operate independently, they are set to be integrated into the pipeline in Phase 3. Prefect orchestrates the entire pipeline, ensuring seamless execution of daily updates. Once integrated, sentiment analysis will enrich the reporting tables with real-time insights, enhancing the user experience.

Challenges Faced:

Throughout Phase 2, we encountered significant challenges, particularly in integrating machine learning (ML) into our data pipeline. Initially, we attempted to implement a custom ML model tailored to our dataset, as recommended by the professor. Despite extensive debugging efforts, the code consistently failed to execute as expected, with errors related to data formatting, model configuration, and environment dependencies. These persistent issues consumed considerable time and resources, ultimately leading us to pivot to an alternative solution.

We then adopted serverless ML, which provided a more streamlined and scalable approach. Using Cloud Run functions, we successfully trained and deployed a sentiment analysis model on a sample dataset from Kaggle. However, when we attempted to train the model using our custom dataset collected through the pipeline, new challenges emerged. Although the model trained successfully, integrating it with our existing infrastructure, specifically MotherDuck and Prefect, proved to be complex and required further refinement.

The next step involved connecting the trained model to our reporting tables in MotherDuck and automating its execution within the Prefect workflow. However, Prefect encountered compatibility issues when orchestrating ML workflows alongside data transformations. This introduced delays and necessitated reconfiguring our pipeline to resolve the conflicts.

Additionally, several objectives we set for Phase 2 were not completed, such as summarizing web articles using GPT-4o and fully integrating sentiment analysis with Prefect. These features remain critical to our project vision and will be prioritized in Phase 3.

Development Practices: Throughout our project, we used GitHub for version control and project management. We established a repository management structure to track tasks, assign responsibilities, and monitor progress, where one team member serves as the primary repository administrator, responsible for reviewing and managing merge requests to avoid overlapping changes.
Link: https://github.com/tangyum/BA-882-Pipeline-Project

Next Steps: For Phase 3, we aim to enhance our pipeline and data warehouse to support comprehensive stock market analysis while establishing a solid technical foundation for future development.

Key Goals for Phase 3:

- **Data Expansion**
  Broaden our dataset to include additional stocks, sector-level information, as well as detailed content from SEC filings for granular market analysis.
- **Advanced Summarization and Analysis**
  Refine our summarization process using advanced LLMs to offer daily recaps for selected stocks. This will improve user experience by providing concise insights and potentially delivering data-driven investment recommendations.
- **Prefect Integration and Automation**
  Enhance Prefect workflows to incorporate sentiment analysis, enabling fully automated ML predictions within the daily pipeline updates.
- **Interactive Reporting**
  Improve the Superset dashboard and explore integrating interactive features like a chatbot. If Superset cannot handle these features, we will consider alternatives like Streamlit or Chainlit.