4 Infrastructure and Data Models Description

4.1 Data Lake

The system utilizes a centralized Data Lake to store, access, and manage financial data.

- **Data Download:** *Quandl*: primarily used to fetch financial reports; *YFinance:* used to gather intraday price and volume data; *NewsAPI:* provides access to financial news.
- **Data Storage:** Currently, only stock-related data is collected, organized by ticker. Boths raw and processed data is stored in Parquet format.
- **Data Access:** Data can be accessed using a combination of the ticker symbol, dataset name, and a specified time range (start and end dates).

4.2 Data Catalog

Since the Data Lake classifies data only by ticker, the Data Catalog organizes datasets by categories such as price-volume data, fundamentals, and news for easier indexing and search.

- **Categorization:** Categories and key attributes of datasets are manually defined and stored in JSON files, which facilitates maintaining existing catalogs and adding new data or categories.
- **Data Search:** Data can be efficiently searched using self.categories = {}, which provides an indexed approach to locate and retrieve relevant datasets.

4.3 Data Workbench

The platform is primarily responsible for extracting, aggregating, and cleaning data.

- **Retrieve data:** The system implements three main functions to facilitate efficient data retrieval.
  retrieve_data is used to retrieve specific data by providing the *ticker* and *dataset_name*.
  retrieve_data_by_dataset aggregates data for all instruments within a given dataset. This is particularly useful for analyzing trends or patterns across multiple tickers.
  retrieve_data_by_category aggregates all datasets under a specific category for a given ticker. This allows for a holistic view of related datasets, enabling comprehensive analysis.
- **Clean data:** Handle Missing Values
- **Transform data:** Extracts data from the Data Lake and applies a transformation function to it. For example, researchers can pass a function to take the logarithm of all numerical columns.

4.4 Quant Data Models

4.4.1 Intraday Data Model

The Intraday Data Model is designed to handle intraday stock price and volume data.

Key Attributes: ***timestamp:*** Records the timestamps of intraday data (type: pd.Series). ***close:*** Closing price data (type: pd.Series). ***volume:*** Volume data (type: pd.Series). **symbol:** Stock ticker symbol (type: str).

Key Methods:

- ***aggregate_by_interval:*** Uses the resample method in pandas to aggregate time-series data at specified intervals (e.g., 60min, 1D). Processes intraday data and returns a pd.DataFrame.

4.4.2 News Data Model

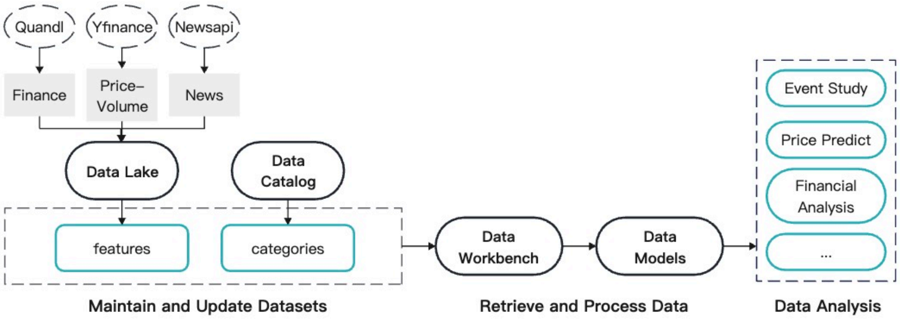
The News Data Model is designed to analyze news sentiment and identify key news articles. Its core functionality lies in sentiment analysis, leveraging the FinBERT model to quantify sentiment.

Key Attributes: ***timestamp:*** Timestamp of the news article (type: pd.Timestamp). ***title:*** News headline (type: str). ***sentiment_score:*** Sentiment score of the news (type: float). ***classifier:*** Shared FinBERT model instance used for sentiment analysis.

Key Methods:

- ***analyze_sentiment:*** Uses FinBERT to analyze the sentiment of the current news headline and generates a sentiment score. Positive sentiment results in a positive score, while negative sentiment results in a negative score.
- ***analyze_dataframe:*** Applies FinBERT sentiment analysis row by row to a DataFrame containing multiple news. Returns a DataFrame with an additional column for sentiment scores.
- ***filter_by_sentiment:*** Filters the current news based on a specified sentiment score threshold.
- ***filter_dataframe_by_sentiment:*** Filters a DataFrame of multiple news articles based on a sentiment score threshold, returning a DataFrame of articles that meet the criteria.

4.5 System Flow



The data, as shown above, has been transformed and organized before being stored in the designated location within the Data Lake for the corresponding ticker. It is categorized and tagged with key attributes in the Data Catalog. When needed, researchers can retrieve this data for a specific time range or combine it with other datasets using the Data Workbench. The Workbench also supports data cleaning and transformation. Once processed, the data is passed to the appropriate Quant Data Model based on its category, where it is instantiated as an object. This object can then be used for analyses such as Event Studies or Price Predictions.