Data Description for Stock and News Analysis Project

Our project utilizes two primary types of datasets: historical stock price data and related financial news articles. These datasets are essential in building a system that explores how external news events may influence stock market behavior.

1. Source and Access Link

Stock Price Data: Collected from Yahoo Finance using the yfinance Python library. This API provides historical market data for publicly traded companies, including metrics such as daily open, close, high, low, and volume.
Library: https://github.com/ranaroussi/yfinance

News Articles: Collected from MarketAux using their RESTful API. The MarketAux API provides real-time and historical financial news from a broad range of sources. Articles were filtered by company ticker and date.
API: https://www.marketaux.com/documentation

2. Dataset Size

Stock Data: Covers 12 weeks (from January 17 to April 4, 2025) of daily trading data for six companies: TSLA, NVDA, META, SPY, LMT, and BA. Each ticker has 60 records (one per trading day), and each record contains 7 attributes, which totals to around 360 records and 2,520 data points across all companies.

News Data: For the same 12 week period, we fetched one news article per weekday per ticker, resulting in approximately 60 news records per company. With six tickers, this totals around 360 news articles, each stored in its own JSON file.

3. Overview of Content and Structure

Stock Dataset: Stores in CSV files and ingested into PostgreSQL. Each record includes:

- ticker, date, openPrice, closePrice, highPrice, lowPrice, and volume

News Dataset: Stored as individual JSON files names by ticker and date (e.g.,

newsDataTSLA2025-02-18) and organized in folders per ticker. Each JSON file contains fields

that correspond with our news table such as:

- ticker, datePublished, headline, content, source, and url

Database Schema: The data is stored in a normalized PostgreSQL schema with four tables:

- Stocks: Daily stock metrics
- News: Financial news articles
- StockMetaData: Company background info (industry, market cap, etc.)
- Sentiment: Future support for sentiment based on news

Foreign key relationships link tickers and news IDs across tables, supporting integrated queries.

4. Potential Questions and Insights

- How do daily news headlines correlate with stock price fluctuations?
- Can sentiment analysis on financial news predict positive or negative price movement?
- Which companies are more sensitive to news exposure?
- Are certain news sources more influential on investor behavior?

By combining financial time series data with external textual data, our system supports exploratory data analysis and paves the way for machine learning models such as price prediction or sentiment impact scoring.