Stock Price Prediction System using Airflow and Snowflake

Niranjan Rao, Ariel Hsieh
Department of Applied Data Science
San Jose State University
Email: niranjanrao.sarafsrinivasrao@sjsu.edu, ariel.hsieh@sjsu.edu

***Abstract*—This report presents the development of a stock price prediction system utilizing historical stock data, real-time APIs, Airflow pipelines, and machine learning models for forecasting. The system integrates data storage, processing, and forecasting, providing accurate predictions for investors. The forecasting models include LSTM and ARIMA, implemented in Snowflake, while Airflow is used for automating data ingestion and processing. This document details the problem statement, requirements, system design, data pipelines, SQL forecasting queries, and system implementation.**

I. Introduction

Stock price prediction plays a crucial role in guiding investment decisions. By leveraging machine learning models and real-time data pipelines, the goal of this project is to predict future stock prices for companies like Lockheed Martin (LMT) and McDonalds (MCD). This report outlines the development process, from setting up the data pipeline to deploying forecasting models in Snowflake.

II. Problem Statement, Requirements, and Specifications

Our team is building a stock price prediction system to forecast future stock prices based on historical data and technical indicators. This application helps investors make informed decisions by providing short-term and long-term stock price predictions. A database is essential for storing large volumes of historical stock data, and data pipelines are required to efficiently process, clean, and update the data for real-time predictions using machine learning models such as LSTM and ARIMA.

- **Data Storage:** A robust database to store historical stock prices, technical indicators, and prediction results.
- **Data Pipeline:** Automated ETL processes to gather, clean, and transform data from various sources, including APIs for real-time updates.
- **Predictive Modeling:** Implementation of machine learning models (e.g., LSTM, ARIMA) for accurate stock price forecasting.
- **Limitations:** Predictions may be affected by market volatility, external economic factors, and model accuracy. The system may not guarantee profits and should be used as a decision-support tool.

III. Functional Components of the Stock Price Prediction System

The stock price prediction system comprises several functional components that work together seamlessly to solve the problem of forecasting stock prices based on historical data. This section discusses each component's role and the interactions between them.

A. Data Sources

The primary data source for this system is the Alpha Vantage API, specifically utilizing the *TIME_SERIES_DAILY* endpoint. This API provides historical stock prices for selected companies, such as Lockheed Martin (LMT) and McDonald's (MCD), over the last 90 days. This rich dataset allows for comprehensive analysis and forecasting of stock price trends.

The Alpha Vantage API is crucial for obtaining real-time and historical stock data, ensuring that the system is equipped with the latest information to generate accurate forecasts. By employing this external data source, the system can leverage reliable financial data to inform its predictions, making it an essential component of the architecture.

B. Database Setup

The data retrieved from the Alpha Vantage API is stored in a Snowflake table named `stock_prices`. The schema of this table is designed to accommodate various attributes of stock prices, including the stock symbol, date, opening price, closing price, minimum and maximum prices for the trading day, and the trading volume.

This structured format ensures that the data is stored efficiently and can be easily queried for analysis. Snowflake's capabilities in handling large datasets make it an ideal choice for this application, allowing for rapid data retrieval and processing.

C. Data Ingestion Pipeline

The data ingestion process is managed by Apache Airflow, which automates the scheduling and execution of data workflows. A daily scheduled Directed Acyclic Graph (DAG) is created to fetch stock prices from the Alpha Vantage API, transform the data into the required format, and load it into the Snowflake table.

The data ingestion process involves several key steps:

- **Fetching Data:** The Airflow DAG calls the Alpha Vantage API to retrieve the latest stock price data.
- **Transforming Data:** The data fetched is parsed and reformatted according to the Snowflake table schema, ensuring that all fields are populated correctly.
- **Loading Data**: The transformed data is then loaded into the stock_prices table in Snowflake, making it available for analysis and forecasting.

This automated pipeline enhances the efficiency and reliability of the data ingestion process, reducing the need for manual intervention and minimizing the risk of errors.

D. Data Processing and Machine Learning

Once the data is stored in Snowflake, it is available for processing and analysis. Python scripts are implemented to interact with the API, parse the incoming data, and format it according to the specified schema before ingestion. This step ensures that the data is clean and ready for use in forecasting models.

In terms of forecasting, SQL queries in Snowflake are employed to set up machine learning models, such as ARIMA (Auto Regressive Integrated Moving Average) and LSTM (Long Short-Term Memory). These models are designed to analyze historical stock prices and generate predictions for the next 7 days based on that data.

- **ARIMA Model**: This statistical method is effective for time series forecasting, taking into account dependencies between an observation and a number of lagged observations.
- **LSTM Model**: A type of recurrent neural network (RNN), LSTMs are well-suited for sequence prediction problems and can learn long-term dependencies in time series data.

The integration of these models allows for both short-term and long-term forecasting, providing a comprehensive outlook on stock price trends.

E. User Documentation

The system includes comprehensive user documentation that covers:

- Instructions on how to modify stock symbols to fetch new data.
- Guidelines for viewing and interpreting stock price predictions generated by the machine learning models.
- Details on how to troubleshoot common errors in data ingestion and processing.

This documentation ensures that users can easily interact with the system, customize it to suit their needs, and understand the outputs of the forecasting models.

F. Database and Data Pipeline Interactions

The overall data pipeline is as follows:

1) Data is retrieved daily from the Alpha Vantage API using a Python script.
2) This data is transformed and loaded into the Snowflake *stock_prices* table via the Airflow DAG.
3) Snowflake serves as both a data storage and processing engine. SQL queries are run to extract historical stock prices for model training.
4) Machine learning models (e.g., ARIMA, LSTM) are trained on the historical data and generate future stock price predictions.
5) These predictions are either stored back into Snowflake or presented to the user for analysis.

This interaction between the Alpha Vantage API, Python scripts, Airflow, Snowflake, and machine learning models ensures a smooth flow of data from raw ingestion to predictive analysis, solving the problem of automated stock price forecasting in a scalable and maintainable manner.

IV. Overall System Diagram

The system is composed of several key components: an Alpha Vantage API for stock data, an Airflow pipeline for data ingestion, and Snowflake for data storage and machine learning. Figure 1 illustrates the overall architecture.

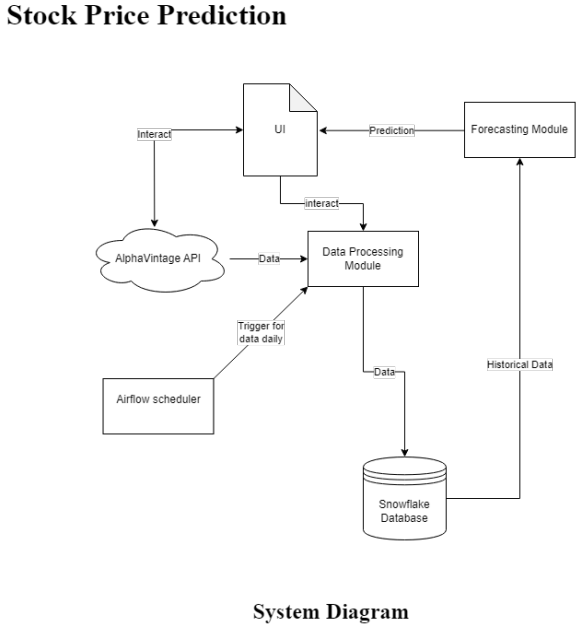

Fig. 1. System Diagram for Stock Price Prediction

V. Table Structures

The stock prices, which are essential for various financial analyses and forecasting activities, are systematically organized and stored in a Snowflake database, utilizing a well-defined table schema that outlines the structure and data types of the information captured within each column. This schema is designed to facilitate efficient data retrieval and management, ensuring that all relevant attributes of stock prices are accurately represented for each trading day.

TABLE I
STOCK PRICES TABLE STRUCTURE

Column Name	Data Type	Description
stock_symbol	STRING	Ticker symbol of the stock (e.g., LMT, MCD)
date	DATE	The date of the stock price record
open	FLOAT	The opening price of the stock on that date
close	FLOAT	The closing price of the stock on that date
min	FLOAT	Lowest price of the stock during the trading day
max	FLOAT	Highest price of the stock during the trading day
volume	INTEGER	Number of shares traded on that date

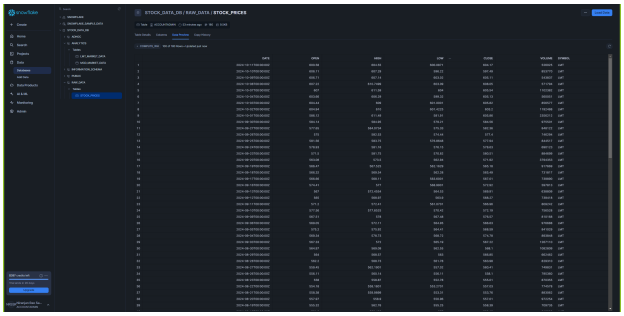

Fig. 2. Data Extracted and Loaded into Snowflake

VI. AIRFLOW DATA PIPELINE

In this section, we will discuss the implementation of the Airflow data pipeline that facilitates the automation of our stock price prediction system. The pipeline is designed to streamline the process of fetching, transforming, and loading stock price data from the Alpha Vantage API into our Snowflake database.

A. Code Repository

The complete code for the Airflow data pipeline has been meticulously organized and made available for public access on GitHub. This repository includes all necessary scripts, configuration files, and documentation required to set up and run the pipeline effectively. Interested users can access the code by following the link below:

- **GitHub Code:** https://github.com/NiranjanRao07/Stock-Price-Prediction

By exploring the GitHub repository, users can gain insights into the design and functionality of the pipeline, as well as the best practices implemented in the code.

B. Airflow Web UI

To provide a visual representation of the data pipeline, a screenshot of the Airflow Web UI is presented below. This interface is instrumental in monitoring and managing the workflow of our data pipeline, allowing users to observe the various tasks involved in the process, their status, and any dependencies that may exist between them.

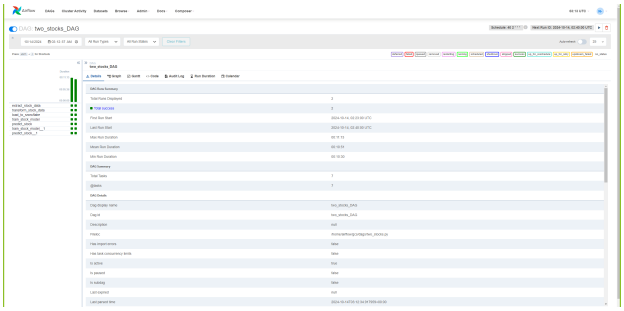

Fig. 3. Airflow Web UI Pipeline

The Airflow Web UI is not only a powerful tool for overseeing the execution of the pipeline but also serves as a platform for debugging and optimizing the workflow. Through this interface, users can easily trigger manual runs of the pipeline, inspect logs for individual tasks, and gain valuable insights into the performance of the data ingestion and transformation processes.

C. Airflow Connections and Variables

The pipeline uses Airflow connections and variables to manage API keys and Snowflake credentials. These features help manage sensitive information and configurations in a secure and flexible way, and allow secure access without hardcoding details in the DAGs. This setup ensures centralized management, security, and flexibility in managing different environments, making it easier to adapt workflows as needed. Screenshots of these settings are shown below:

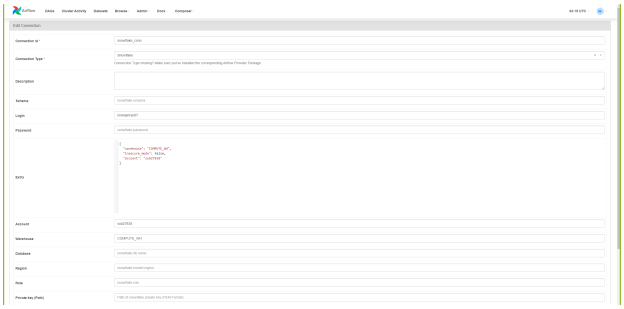

Fig. 4. Airflow Connections for Snowflake and Alpha Vantage

Fig. 5. Airflow Variables for API Keys and Credentials

VII. SQL QUERIES FOR STOCK PRICE FORECASTING

A. SQL Code

Stock price forecasting was implemented using SQL queries in Snowflake, utilizing models like ARIMA and LSTM. The code is available in the following GitHub repository:

- **GitHub Code:** https://github.com/NiranjanRao07/Stock-Price-Prediction

The Airflow tasks create a pipeline to train a stock price forecasting model and generate predictions for a specific stock symbol. The train_stock_model task prepares the data for training by creating a view that filters stock prices by the symbol. It then creates a model using Snowflake's machine learning capabilities to forecast stock prices. The predict_stock task uses the trained model to predict the next 7 days of stock prices and stores the forecasted data in a table. This process helps in automating the prediction and evaluation of stock price forecasts using Snowflake's capabilities for data storage, view creation, and machine learning.

B. Prediction Results

The stock price predictions for companies such as Lockheed Martin (LMT) and McDonald's (MCD) are shown in the following screenshots:

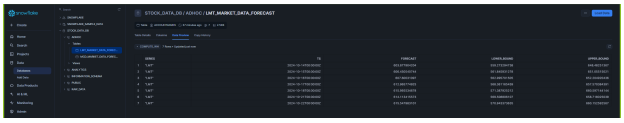

Fig. 6. Stock Price Forecast for Lockheed Martin

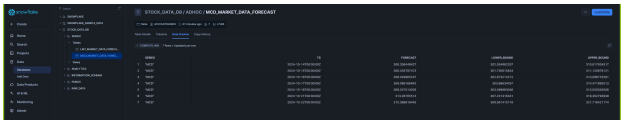

Fig. 7. Stock Price Forecast for McDonald's

VIII. Conclusion

In conclusion, this stock price prediction system integrates Airflow for automated data pipelines and Snowflake for data storage and machine learning. By using models such as LSTM and ARIMA, we can provide both short-term and long-term stock price forecasts. The system is designed to support investor decision-making by delivering accurate predictions based on historical and real-time data.