Stock Price Prediction Analytics System

Jie Heng (018321914) Savitha Vijayarangan (018315986)

Abstract

This project builds a stock price prediction system using Yahoo Finance data. Historical prices for NVIDIA and Apple are stored in Snowflake and forecasted with machine learning for the next 7 days, automated via Airflow for daily updates.

Keywords

Stock Price Prediction, Yahoo Finance, Snowflake, ML Forecasting, Airflow, Data Pipeline.

I. Problem Statement

This project builds a stock price prediction system using Yahoo Finance API [1] data. Historical prices of NVIDIA and Apple are stored in Snowflake and analyzed with machine learning models to forecast the next 7 days, enabling automated financial analytics for informed investment decisions. A reliable database ensures consistent storage of historical data for accurate predictions, while automated data pipelines streamline daily ETL and forecasting, maintaining up-to-date data for analysis. The system stores historical stock data in Snowflake and uses Airflow to automate daily ETL and forecasting. This ensures up-to-date data, accurate predictions, and real-time trend analysis, providing investors and analysts with reliable insights for informed decision-making.

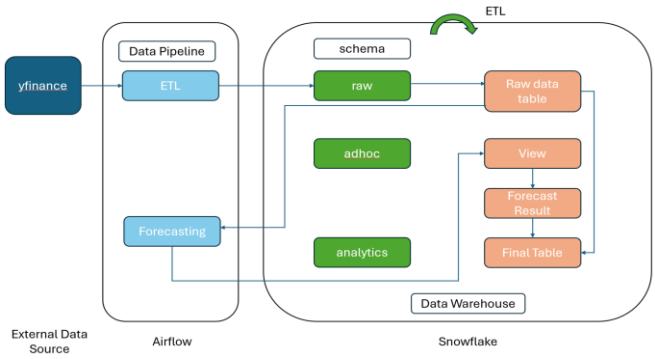
II. Requirements

Collect historical stock prices (180 days) for NVDA and AAPL. Store data into a Snowflake table with proper schema. Run a daily ETL pipeline to refresh stock prices. Apply ML forecasting to predict the next 7 days of stock prices for both companies. Store predicted prices into a dedicated table [2]. Forecast accuracy needs clean, complete data. It predicts only 7 days, limiting long-term use. Using the same model for all stocks may miss industry patterns, and reliance on Snowflake ML limits advanced model flexibility. Analysts can view trends and predictions, while investors assess future prices with indicators. The system supports adding more stocks with minor code and table changes.

III. Functional Analysis

Overall Architecture

Figure 1: Overall Architecture



Data Pipeline The proposed system consists of two main data pipelines, implemented as Airflow DAGs: ETL Pipeline: Fetches and loads historical stock data into Snowflake daily. Forecasting Pipeline: Trains a prediction model (e.g., ARIMA) and inserts the predicted stock prices into Snowflake.

Figure 2: Airflow DAGS

Functional Components

Use functions and tasks to solve the problem [3] [4].

Table 1: Function - return_snowflake_conn()

Name	Extract
Parameters	NA
Return	List of dictionaries (records) where each dictionary represents a stock record
Purpose	Extracting stock data from Yahoo Finance API for NVDA and AAPL (last 180 days)

Table 2: Task - extract

Name	return_snowflake_conn
Parameters	None
Return	Snowflake cursor object
Purpose	Establishes a connection to Snowflake via Airflow's SnowflakeHook
Name	transform
Parameters	extracted_data
Return	Dictionary with keys "symbol" and "records"
Purpose	Transform extracted stock data for consistency

Table 3: Task - transform

Name	load
Parameters	transformed_data, target_table
Return	None
Purpose	Load the transformed stock data into Snowflake

Table 4: Task - load

Name	get_snowflake_cursor
Parameters	None
Return	Snowflake cursor object
Purpose	Gets a new Snowflake connection cursor using Airflow SnowflakeHook

Table 5: Function - get_snowflake_cursor()

Name	train
Parameters	train_input_table (str), train_view (str), forecast_function_name (str)
Return	None
Purpose	Create a view from the input data and to set up a Snowflake machine learning forecast function for stock price predictions.

Table 6: Task - train

Name	predict
Parameters	forecast_function_name (str), train_input_table (str), forecast_table (str), final_table (str)
Return	None
Purpose	Uses the trained forecast function to generate predictions and merges historical data with predictions into a final table

Table 7:Task – predict

IV. Table Structure

MYLABDATABASE.RAW.STOCK_DATA

Field Name	Data Type	Attributes	Constraints	Description
SYMBOL	VARCHAR	Not null	Primary Key with DATE	NVDA, AAPL
DATE	DATE	Not null	Primary Key with SYMBOL	Trading date
OPEN	FLOAT			Opening price
CLOSE	FLOAT			Closing price
LOW	FLOAT			Lowest price
HIGH	FLOAT			Highest price

MYLABDATABASE.ADHOC.STOCK_DATA_VIEW

Field Name	Data Type	Attributes	Constraints	Description
DATE	DATE	Not null	Primary Key, Foreign Key	Trading date
CLOSE	FLOAT / NUMBER		Foreign Key	Stock closing price
SYMBOL	VARCHAR	Not null	Primary Key, Foreign Key	NVDA, AAPL

MYLABDATABASE.ADHOC.STOCK_DATA_FORECAST

Field Name	Data Type	Attributes	Constraints	Description
SERIES	VARCHAR	Not null	Primary Key, Foreign Key	Stock symbol
TS	FLOAT	Not null	Primary Key	Forecasted date
FORECAST	FLOAT			Predicted closing price
LOWER_BOUND	FLOAT			Lower bound of the confidence interval
UPPER_BOUND	FLOAT			Upper bound of the confidence interval

MYLABDATABASE.ANALYTICS.MARKET_DATA

Field Name	Data Type	Attributes	Constraints	Description
SYMBOL	VARCHAR	Not null	Primary Key with DATE, Foreign Key	NVDA, AAPL
DATE	DATE	Not null	Primary Key with SYMBOL, Foreign Key	Trading date or Forecasted date
ACTUAL	FLOAT			Actual closing price (present for historical data, NULL for predictions)
FORECAST	FLOAT		Foreign Key	Forecasted closing price (NULL for historical data, present for predictions)
LOWER_BOUND	FLOAT		Foreign Key	Lower bound of the confidence interval
UPPER_BOUND	FLOAT		Foreign Key	Upper bound of the confidence interval

V. Implementation (Python Codes & SQL Queries)

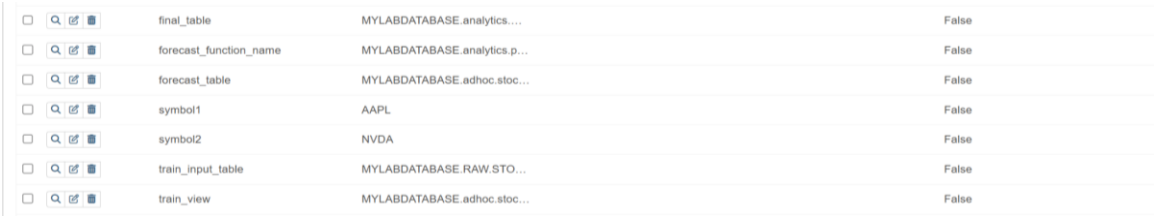
Refer to github url:

data226/lab1_etl.py at main · SunnyJaneH/data226

data226/lab1_trainpredict.py at main · SunnyJaneH/data226

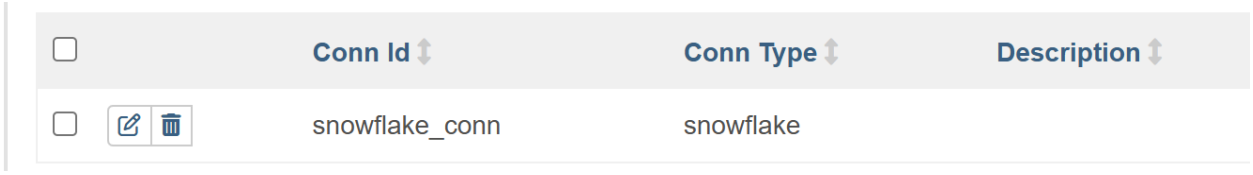
Airflow UI Screenshots

Figure 3: Variables



final_table	MYLABDATABASE.analytics....	False
forecast_function_name	MYLABDATABASE.analytics.p...	False
forecast_table	MYLABDATABASE.adhoc.stoc...	False
symbol1	AAPL	False
symbol2	NVDA	False
train_input_table	MYLABDATABASE.RAW.STO...	False
train_view	MYLABDATABASE.adhoc.stoc...	False

Figure 4: Connection



Conn Id	Conn Type	Description
snowflake_conn	snowflake	

Figure: Airflow UI Screenshots

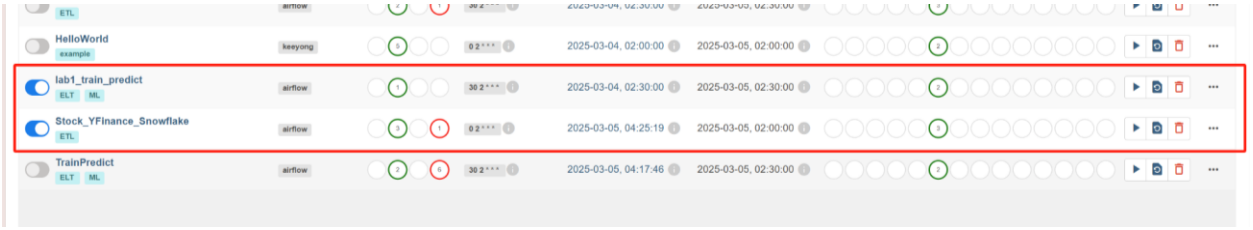


Figure 5: ETL Pipeline

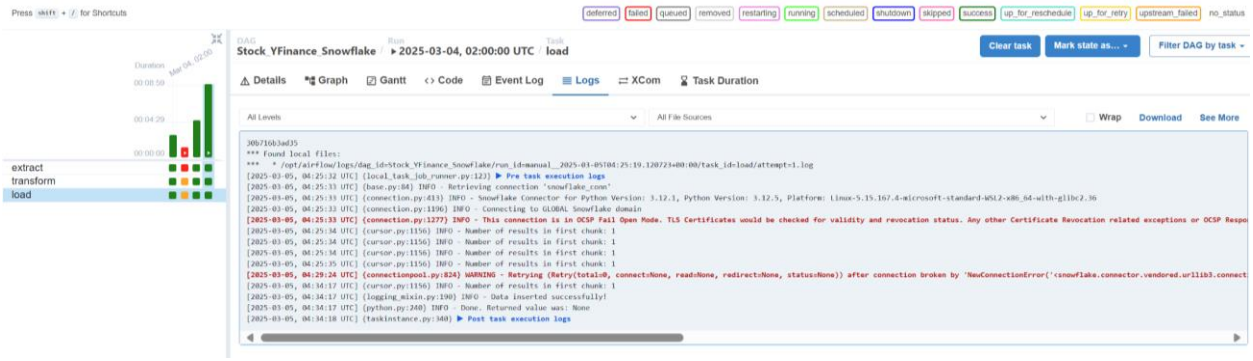
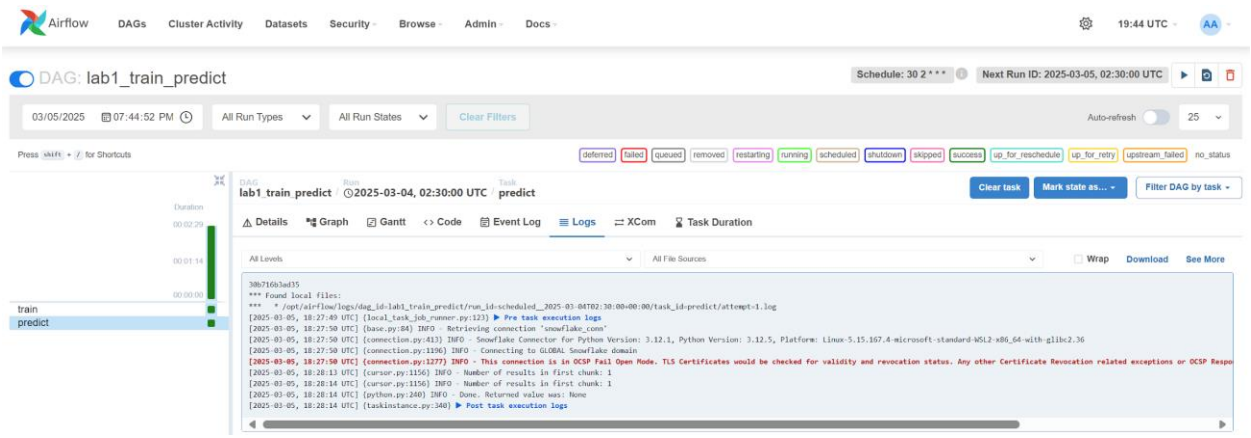


Figure 6: Forecasting Pipeline

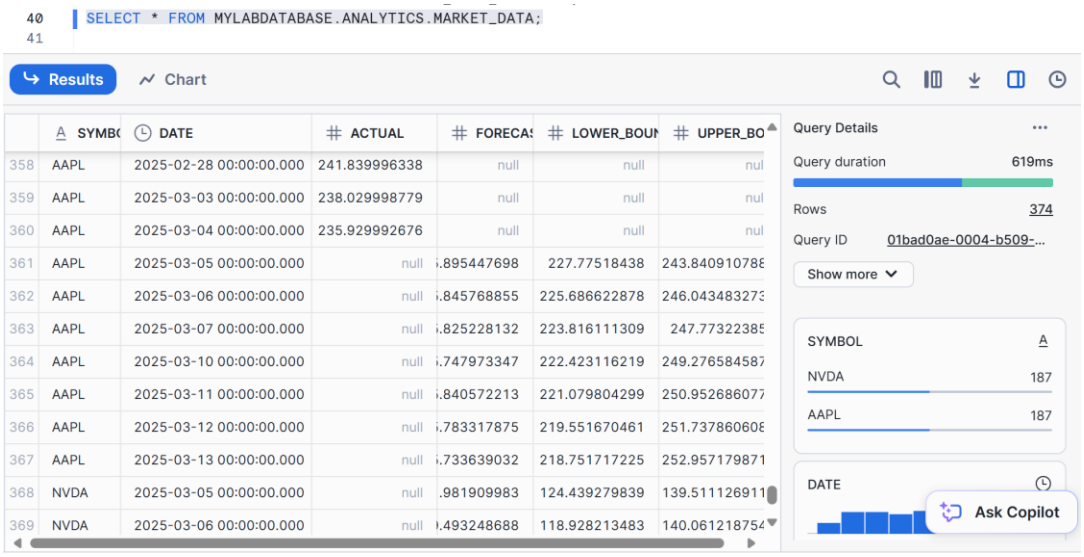


SQL Query

Refer to code.

VI. Results & Findings

Figure 7: Output in Snowflake



The predicted price for AAPL remains relatively stable, hovering around $235.7 to $235.8. The lower bound of the confidence interval gradually decreases from $227.78 to $218.75, indicating potential downside risk. NVDA's predicted price shows a gradual decline from $131.98 on March 5 to $122.63 on March 13.The lower bound drops significantly from $124.44 to $103.02, showing increasing downside risk.

References

[1] yahoo, “yfinance 0.2.54”, https://pypi.org/project/yfinance/
[2] K. Han, “Data226 Building a Stock Price Prediction Analytics using Snowflake & Airflow”, 2025
[3] K, Han, “country_capital_to_snowflakse_v2.py”, 2025
[4] K, Han, “sjsu-data226-SP25/week6/train_predict.py”,2025