DS598

Team 12

From Tweets to Trends

Predicting Stock Volumes Using X Sentiment

Spring 2025

Team Members

Yuchen Li

Syeda Aqeel

Hsiang Yu Huang

Introduction

- This project aims to predict stock price movements based on public sentiment expressed on X

Business Applications

- **Smart Trading**
  Leverages real-time sentiment to enhance algorithmic strategies, especially for fast-moving stocks like NVIDIA.

- **Early Risk Detection**
  Detects market panic signals through tweet tone shifts—supporting timely hedging.

- **Strategic Investment Insights**
  Identifies patterns in public mood and key influencers shaping investor sentiment.

- **Reputation Management**
  Helps NVIDIA refine communication by tracking sentiment dips tied to events (e.g., DeepSeek launch reaction).

- **Influencer Targeting for Marketing**
  Reveals which verified voices move the market—useful for brand partnerships or PR.

- **R&D and Academic Value**
  Offers a scalable blueprint for future applications across fintech and behavioral finance.

- **Real-Time, Low-Cost Forecasting**
  Automated insights without human bottlenecks—ideal for daily operational decision-making.

Why NVIDIA? Why Stock Volume

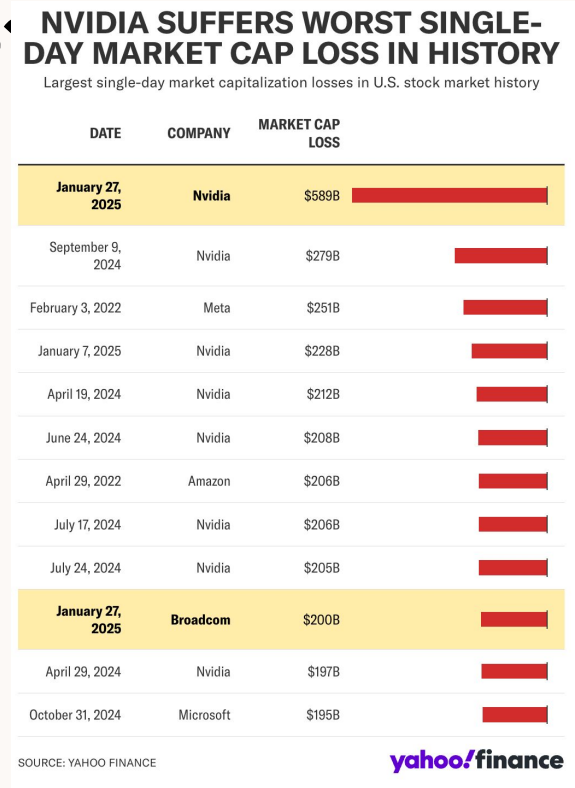
- Frequent Social Media Attention
  - DeepSeek's Launch made NVIDIA stock dropped

Why not stock price

- Non-stationary
- highly volatile nature

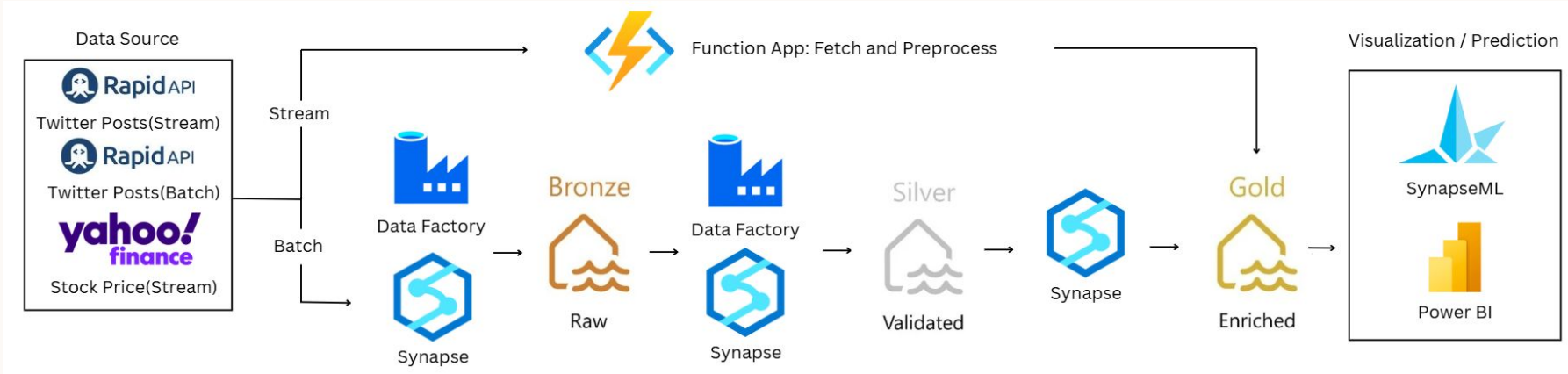
yahoo!finance

Nvidia stock plummets, loses record $589 billion as DeepSeek prompts questions over AI spending



Project Architecture

- Medallion Structure (Bronze → Silver → Gold)
  - Modular, Scalable Processing



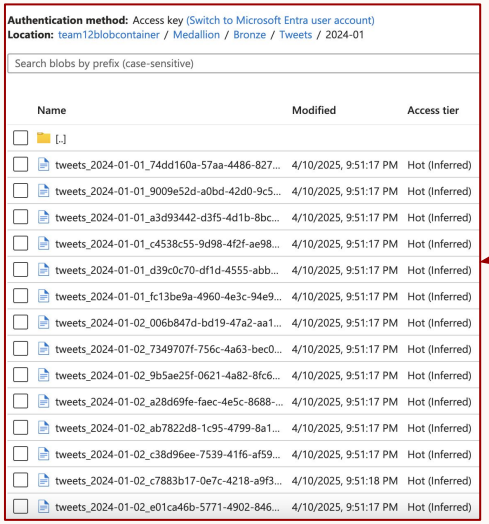
Data Sources

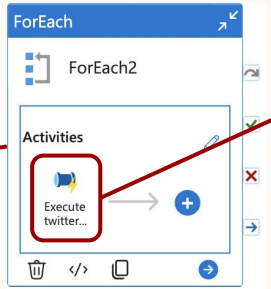
- Tweets via RapidAPI's Tweet V2 API (Batch + Streaming)
- Stock data via Yahoo Finance (*yfinance* in Synapse)

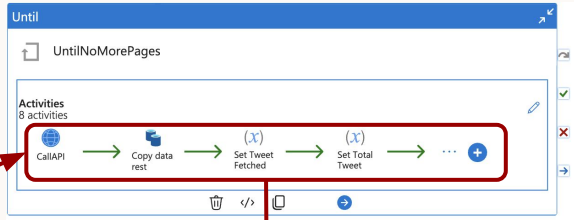
Batch Processing (Tweet)

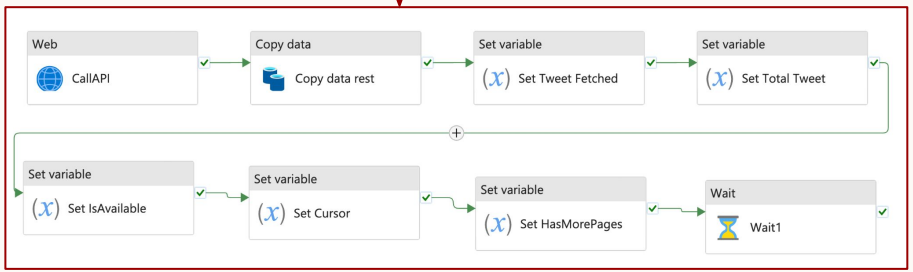
Bronze (via ADF)

- Collect daily tweets with hashtag #$NVDA via RapidAPI
- Use bottom cursor for pagination across multiple calls
- ForEach Loop → Until Loop → Stores JSON output to Blob Storage (Bronze Layer)





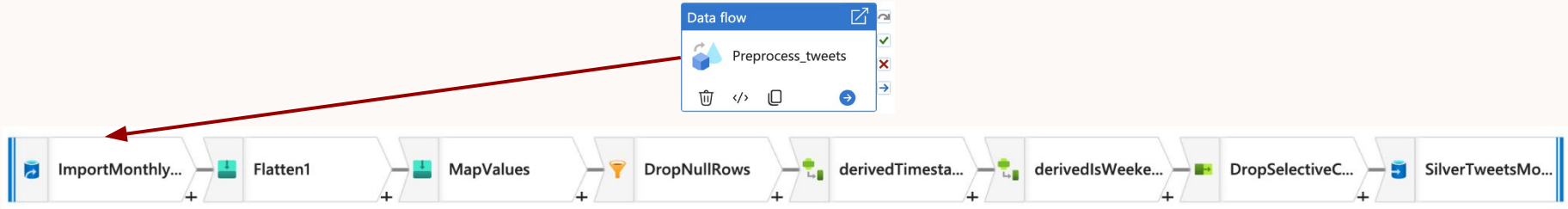




Batch Processing (Tweet)

Silver (via ADF Data Flow)

- Executed in monthly units
- Read raw JSON tweet files from Bronze Layer
- Steps Performed
  - Flatten for nested fields→ Map Values → Drop Null Rows → Derive New Columns → Drop Redundant Columns
- Output: Cleaned dataset (20 columns)
  - Store in Silver Layer



Stock Pipeline (Bronze)

Bronze (via Synapse)

- Collect daily stock information through python library finance
- Set Ticker as “NVDA”
- No data on non-market day

```
1  import yfinance as yf
2  import pandas as pd
3  from datetime import datetime, timedelta
4  # Pull data from Yahoo Finance
5  ticker = "NVDA"
6  start_date = "2024-01-01"
7  end_date = (datetime.strptime("2025-04-28", "%Y-%m-%d") + timedelta(days=1)).strftime('%Y-%m-%d')
8  df = yf.download(ticker, start=start_date, end=end_date)
```

✓ - Command executed in 250 ms on 7:03:32 PM, 5/02/25

Price	Date	Close	High	Low	Open	Volume
Ticker		NVDA	NVDA	NVDA	NVDA	NVDA
0	2024-01-02	48.149918	49.276493	47.577135	49.225514	411254000
1	2024-01-03	47.551144	48.165911	47.302237	47.467176	320896000
2	2024-01-04	47.979984	48.481795	47.490167	47.749068	306535000

Stock Pipeline (Silver)

Silver (via Synapse)

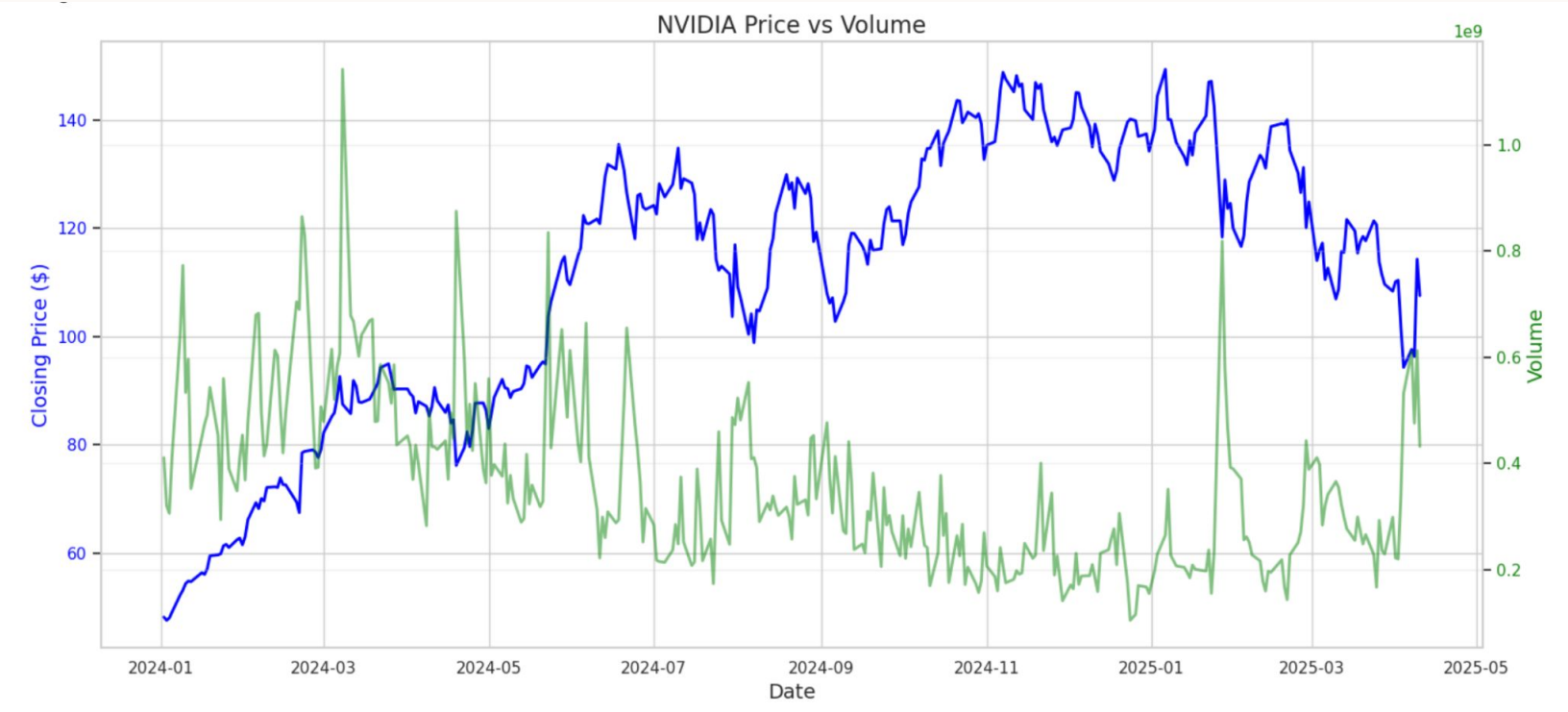
- Check and drop Duplicate rows
- Transform Date column to string time format, align with Tweet date format
- Save as Parquet format

```
Index(['Date', 'Close', 'High', 'Low', 'Open', 'Volume'], dtype='object')
```

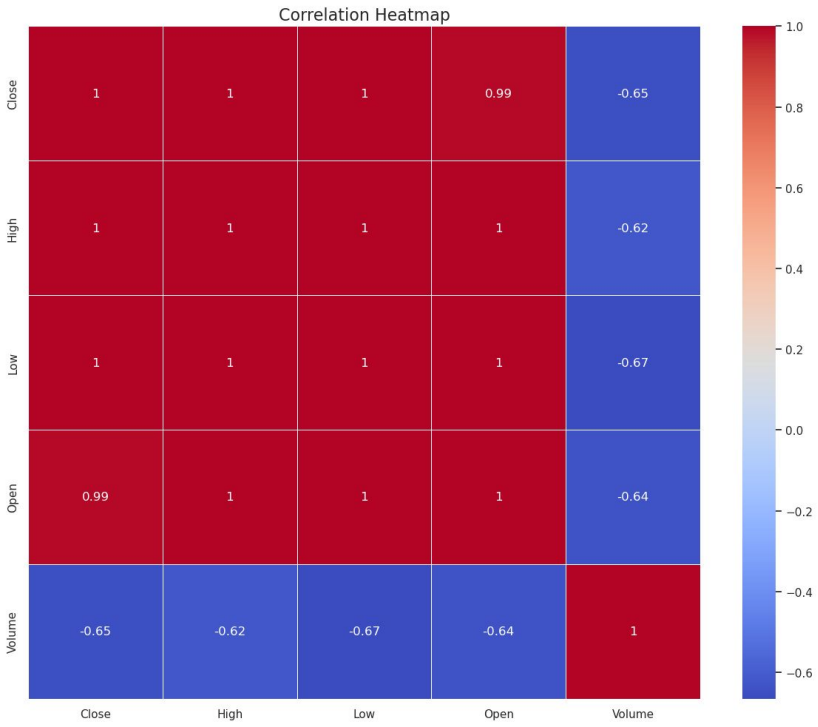
	Date	Close	High	Low	Open	Volume
0	2024-03-01 00:00:00	82.248108	82.269104	79.405174	79.969964	479135000
1	2024-03-04 00:00:00	85.205002	87.662079	83.687574	84.098415	615616000
2	2024-03-05 00:00:00	85.935760	86.068720	83.389600	85.241988	520639000
3	2024-03-06 00:00:00	88.670860	89.694526	87.001411	87.993089	582520000
4	2024-03-07 00:00:00	92.638550	92.736517	89.572556	90.128373	608119000

Exploratory Data Analysis

sharp drop in NVIDIA's stock price and surge in trading volume in early 2025 aligns with the market's reaction to DeepSeek

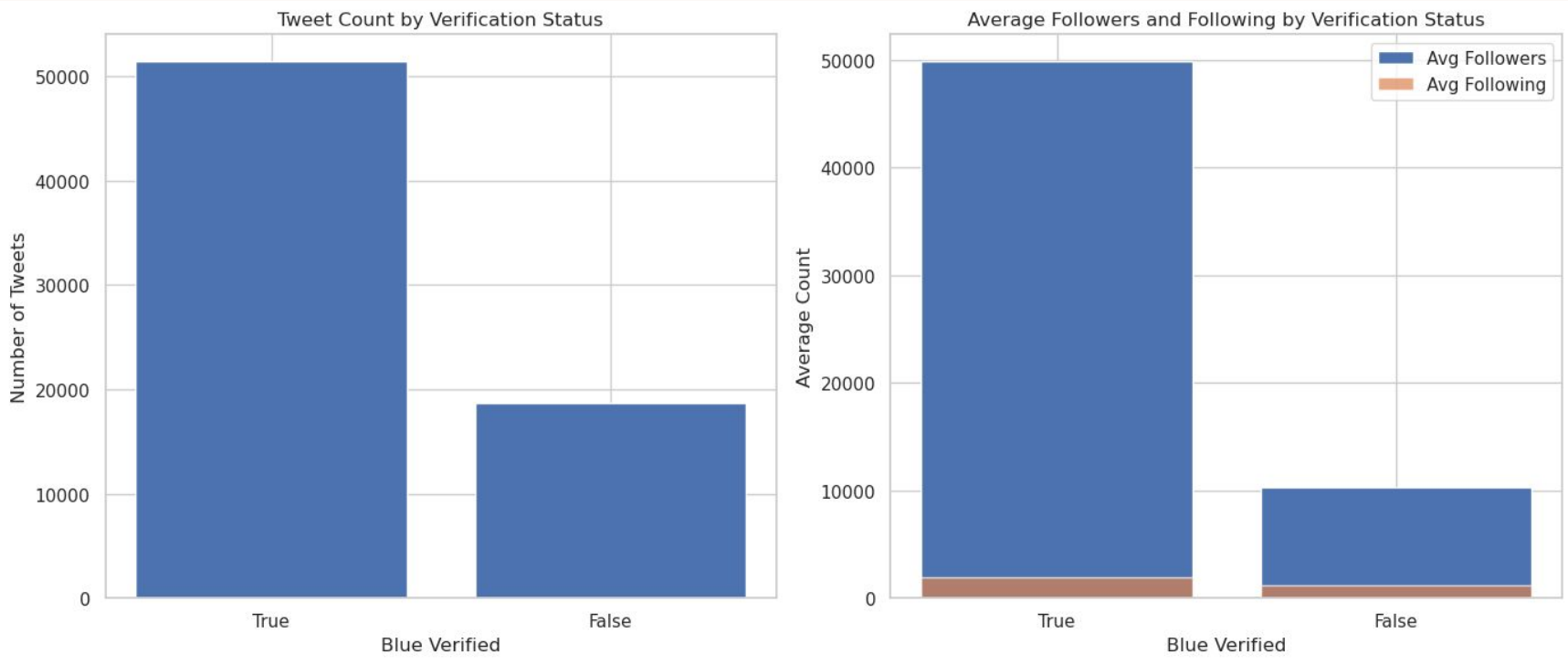


Correlation Between NVIDIA Stock Metrics

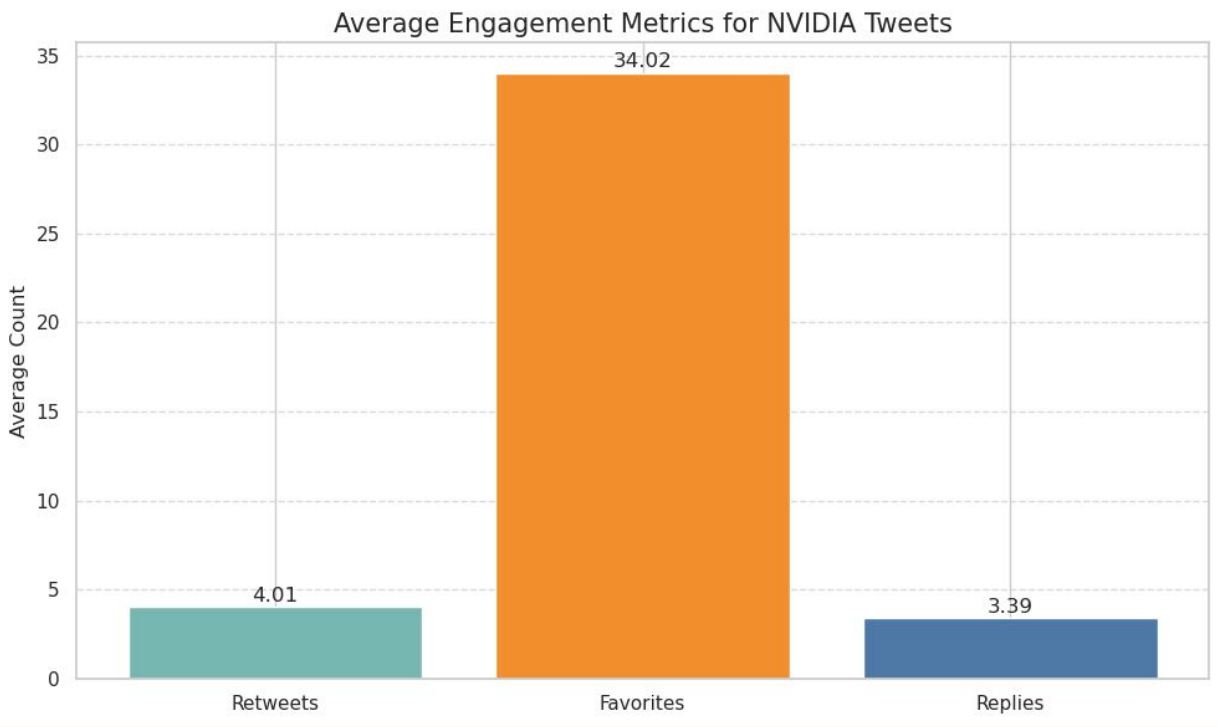



```
--- Summary of Stock Performance ---
Period: 2024-01-02 00:00:00 to 2025-04-10 00:00:00
Starting Price: $48.15
Ending Price: $107.57
Overall Price Change: 123.41%
Average Trading Volume: 360698947
Maximum Volume: 1142269000
```

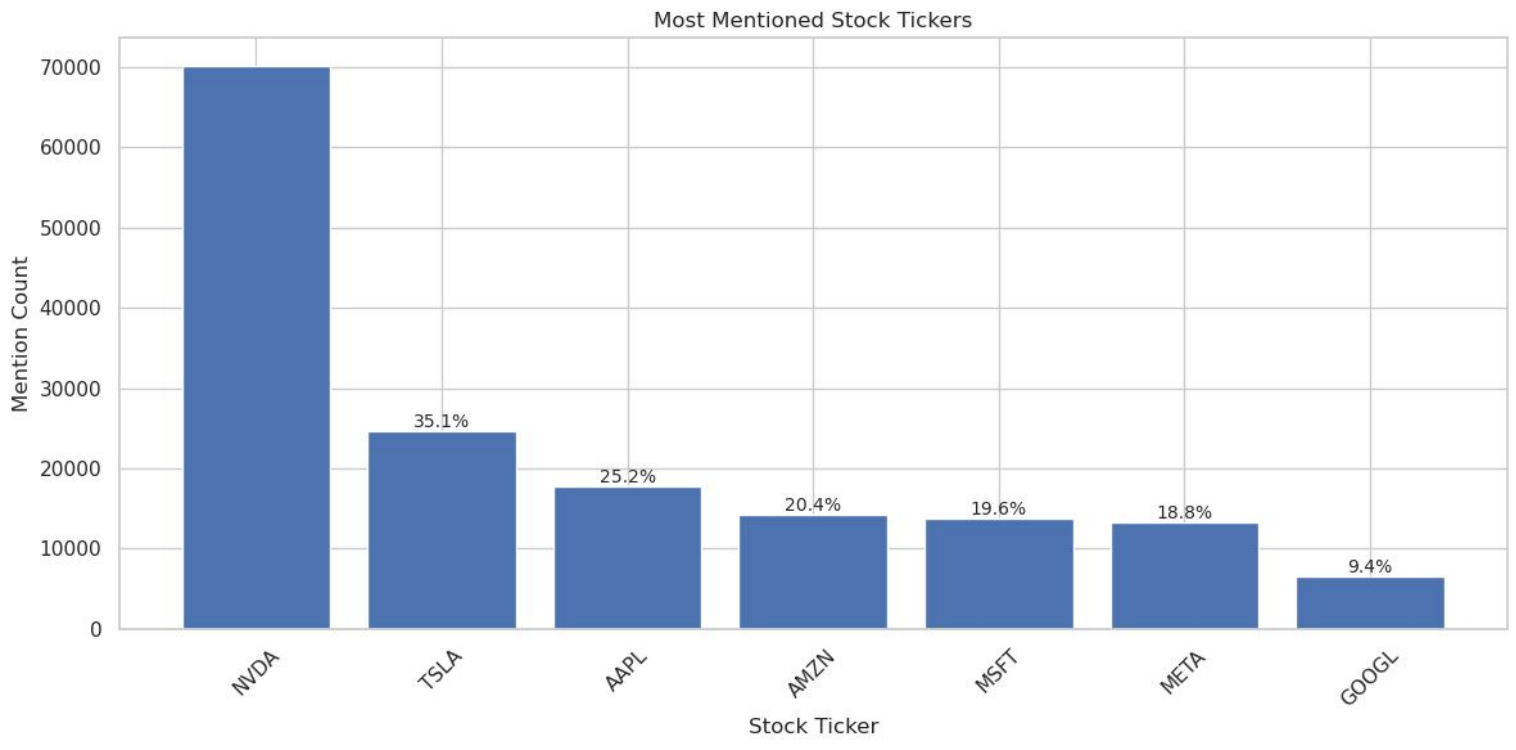
Engagement and Reach of Blue-Verified Users Tweeting About #NVIDIA



Favorites Drive Engagement on #NVIDIA Tweets, Outpacing Retweets and Replies



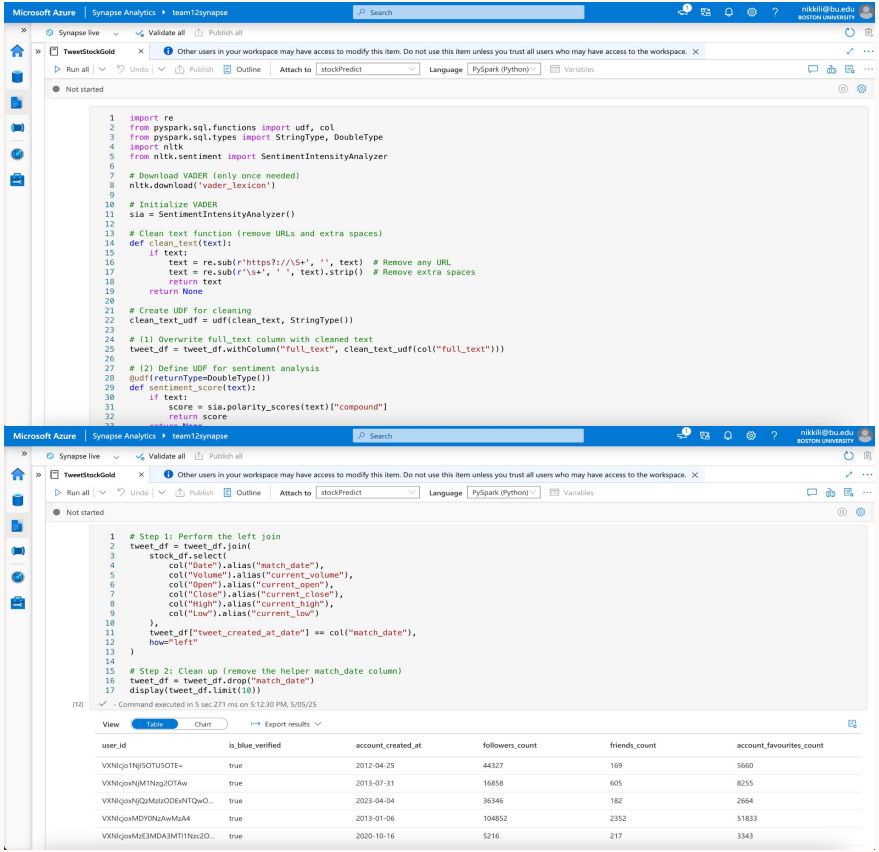
Most Co-Mentioned Stock Tickers in #NVIDIA Tweets



Gold Merge (Tweet + Stock)

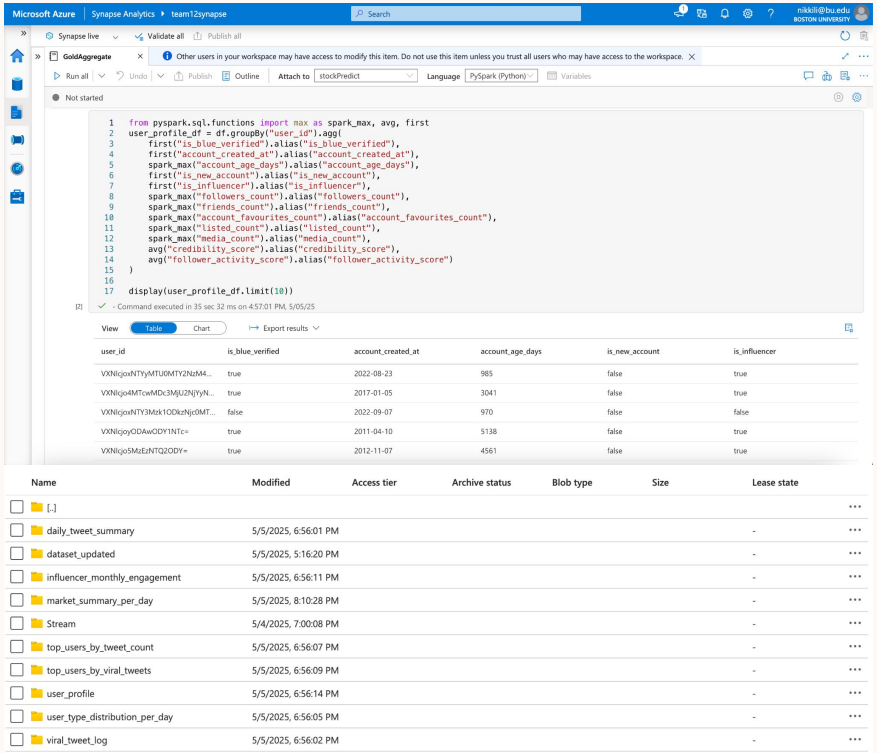
Gold Layer (via Synapse Spark)

- Perform Sentiment Analysis with VADER
- Feature Engineering
  - Engagement Scope
  - User Metrics
  - Flags
- Join & Output
  - Merge two datasets on date
  - Add next_available_volume (target for ML)
  - Output to Gold/dataset_updated/



Gold Aggregate for Visualization

- **User Profile Table:** Captures metrics like follower count, credibility, influencer flags
- **Daily Tweet Summary:** Avg. sentiment, engagement totals, viral count by date
- **Market Summary per Day:** Merges tweet stats with daily stock performance
- **Viral Tweet Log:** Stores all high-impact viral tweets
- **User Type Distribution:** Verified vs. Unverified, Influencer vs. General, New vs. Established
- **Top Users by Tweet Count / Viral Tweets:** Highlights most active and influential contributors
- **Influencer Monthly Engagement: Tracks** influencers' impact over time

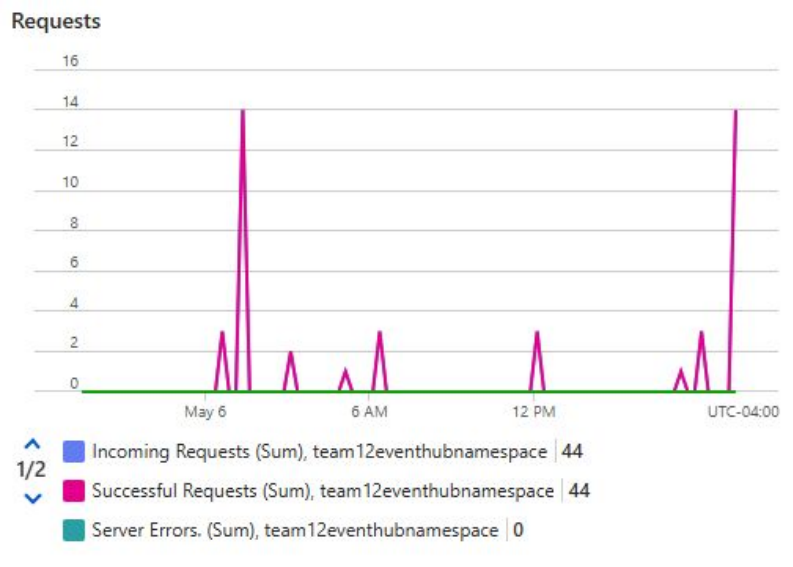


Stream Data

- Triggers Every day at 5am
- Process directly to Gold Layer Schema
- Save to Blob Storage
- Skip non-market day data

Why not Event Hub?

- Limited Message size of 1MB



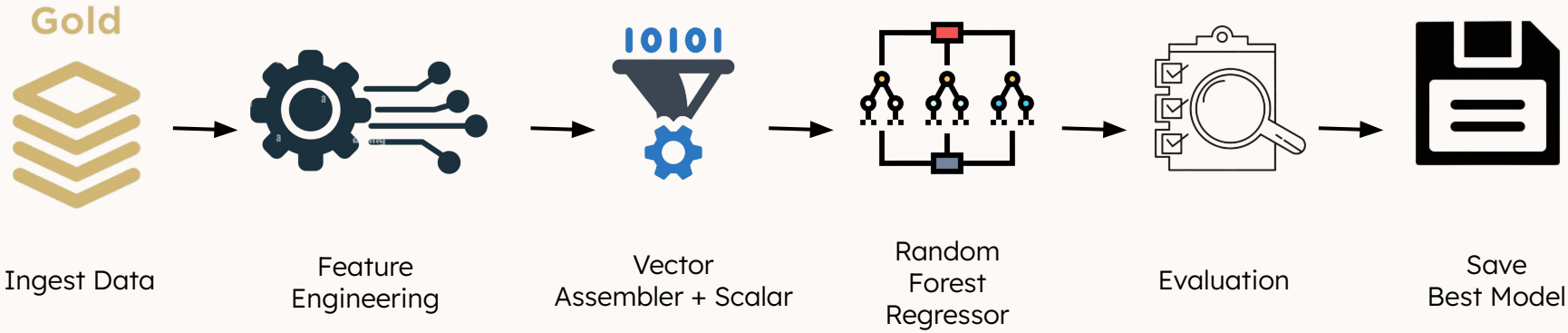

```
@app.timer_trigger(
    schedule="0 0 9 * * *",
    arg_name="myTimer",
    run_on_startup=False,
    use_monitor=False
)
```

```
def save_to_blob(data: dict, date: datetime) -> bool:
    """Save data to Azure Blob Storage"""
    try:
        date = date.strftime('%Y-%m-%d')
        blob_path = f"Medallion/Gold/Stream/{date}.json"
        data_json = json.dumps(data, indent=4)
        blob_client = blob_service_client.get_blob_client(container=container, blob=blob_path)
        blob_client.upload_blob(data_json, overwrite=True)
        return True
```

Machine Learning Pipeline

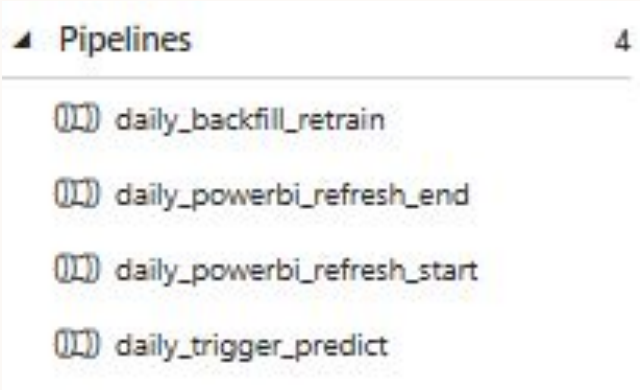
Spark MLlib

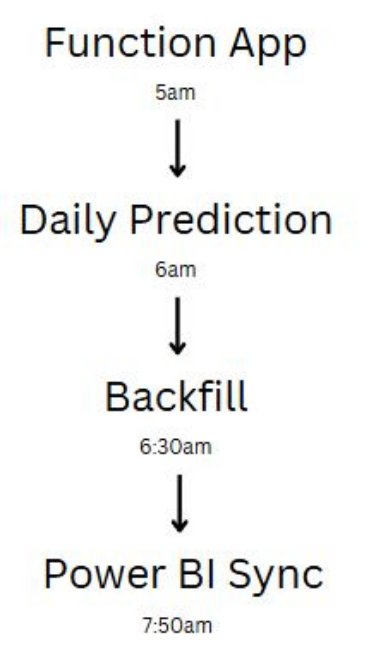
Goal: Predict next-available NVIDIA stock volume from tweet sentiment



Daily Trigger Routine for Automatic Prediction

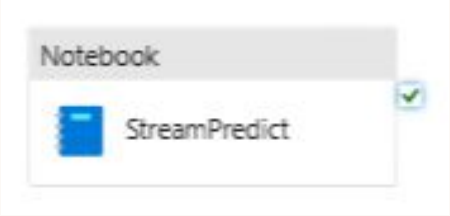
- Automated prediction pipeline for streaming data
- Starting from 5am everyday





Daily Prediction Workflow

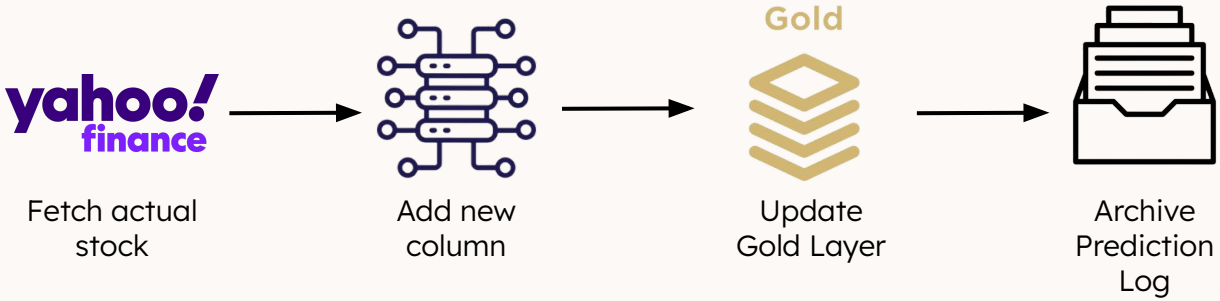
- **Input**: :Loads streaming data from Blob Storage.
- **Preprocessing**: Transformation pipeline used for prediction
- **Prediction**: Loads best model and predicts for the end-of-day stock volume
- **Output**: Write to the “Prediction/Streaming” folder.

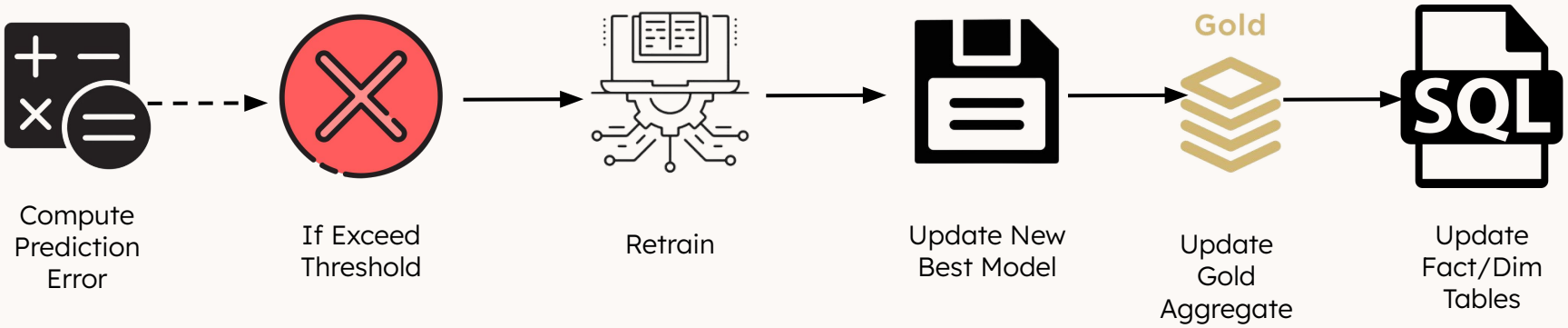


daily trigger	Schedule trigger	5/6/2025, 7:00:00 AM	Succeeded		
daily_trigger_predict	5/6/2025, 7:00:02 AM	5/6/2025, 7:06:32 AM	6m 31s	daily trigger	Succeeded

2025-05-05	5/6/2025, 7:05:59 AM

Backfill & Retraining Logic





daily backfill retrain	Schedule trigger	5/6/2025, 6:30:01 PM		Succeeded	
daily_backfill_retrain	5/6/2025, 6:30:01 PM	5/6/2025, 6:54:46 PM	24m 45s	daily backfill retrain	Succeeded

SQL Integration & Pool Management

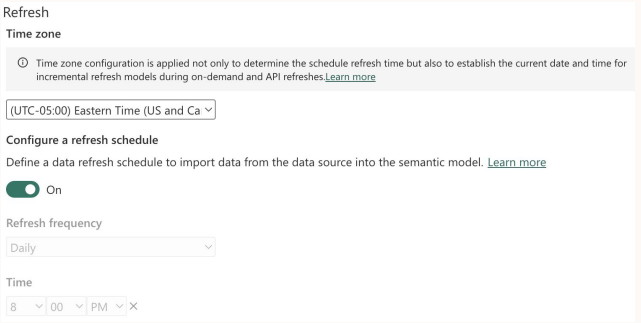
External Tables: Aggregated Gold Layer data

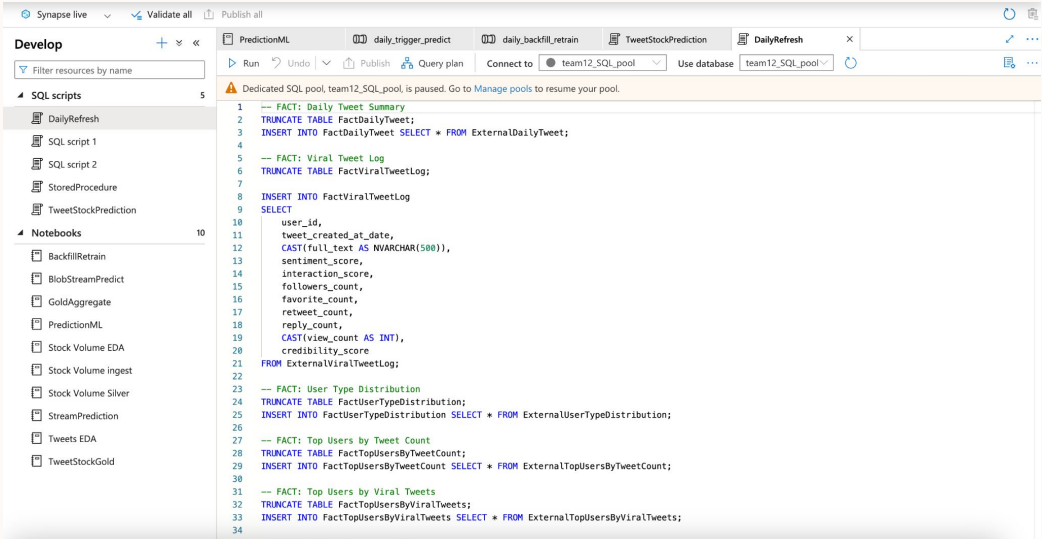
Fact & Dimension Tables: Built on top of external tables for PowerBI

Stored Procedure - DailyRefresh:

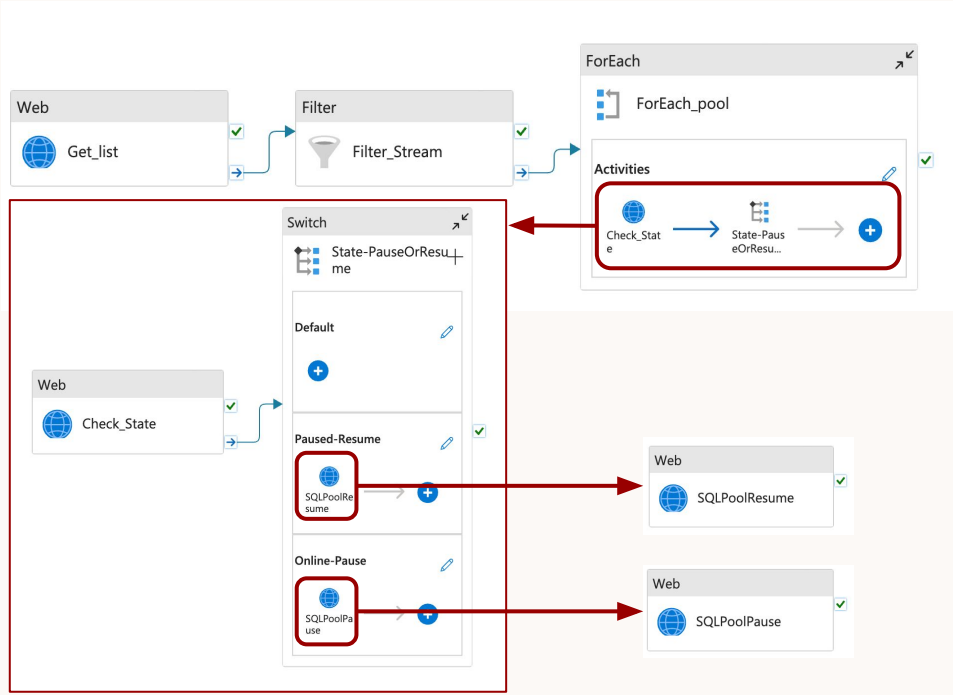
- Truncates & Refreshes Fact Tables
- Cast data types as needed
- Applies transformations

Runs **Daily** to sync Power BI dashboards



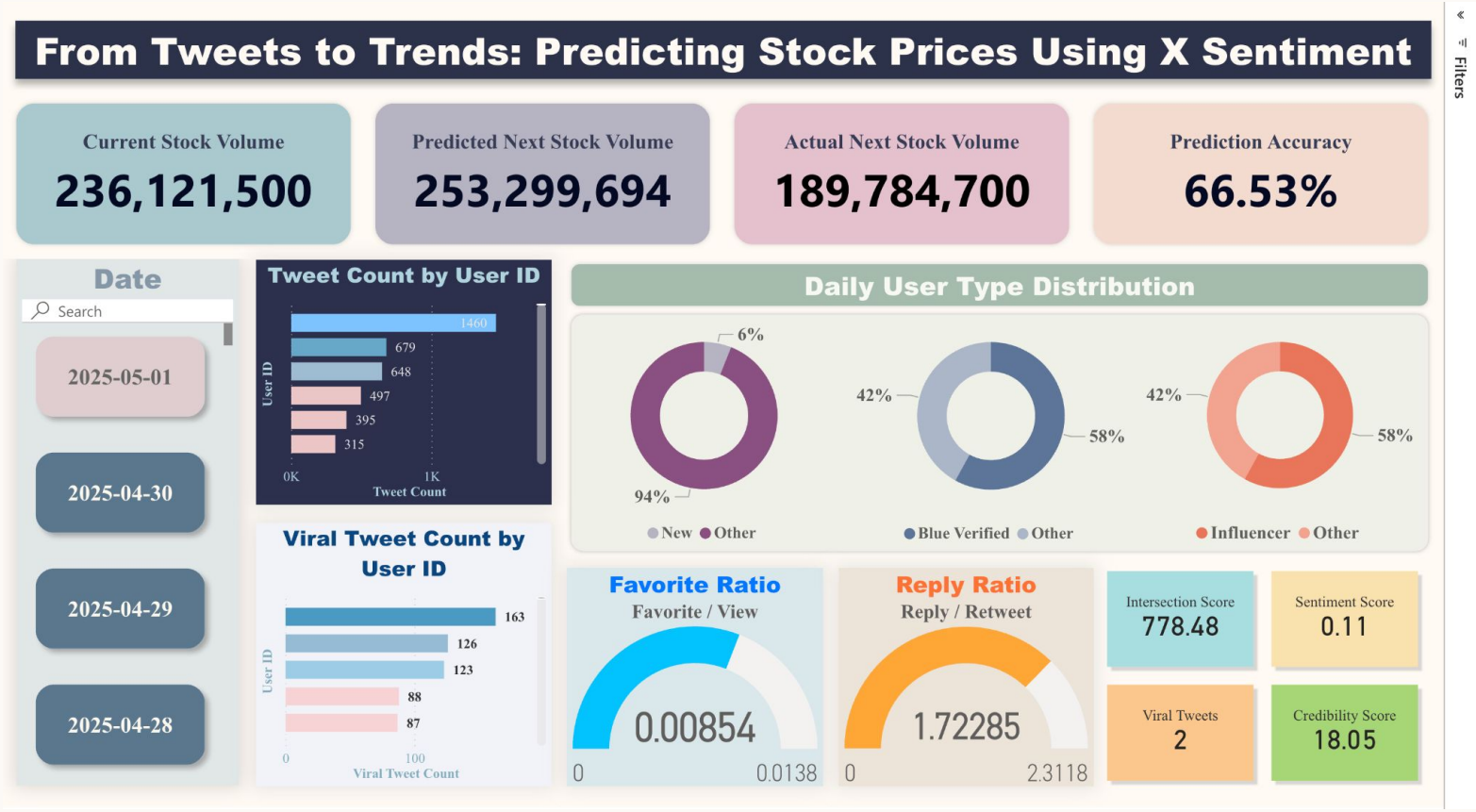


SQL Pool Management

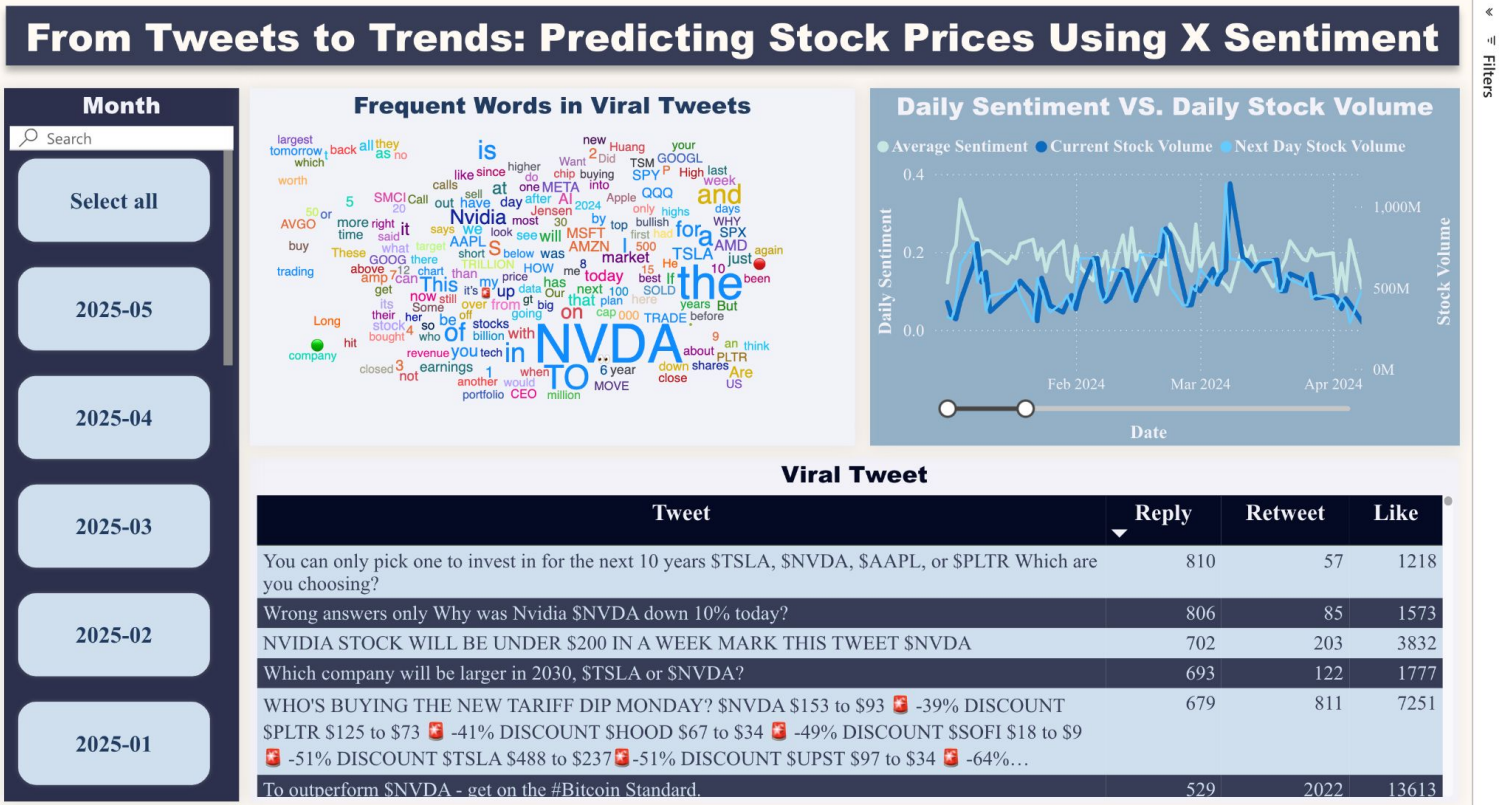


daily_powerbi_refresh_end	5/6/2025, 6:53:00 PM	5/6/2025, 6:54:45 PM	1m 45s	0ef642e6-9f10-452b...	Succeeded
daily_powerbi_refresh_start	5/6/2025, 6:39:20 PM	5/6/2025, 6:42:31 PM	3m 11s	47380450-2ea0-41b...	Succeeded

PowerBI Dashboard



Power BI Dashboard

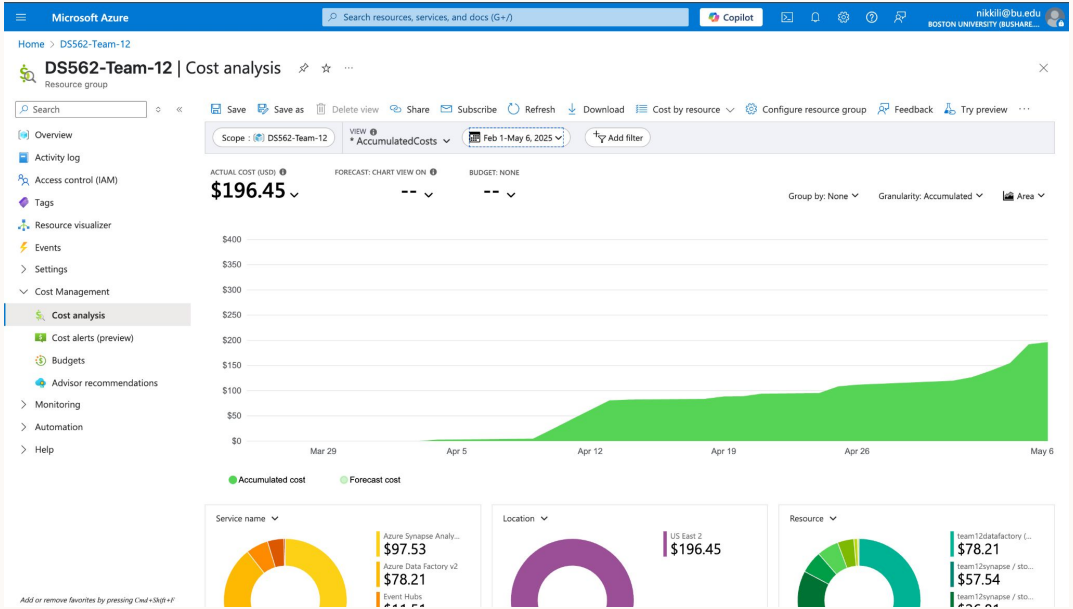


Viral Tweet

Tweet	Reply	Retweet	Like
You can only pick one to invest in for the next 10 years $TSLA, $NVDA, $AAPL, or $PLTR Which are you choosing?	810	57	1218
Wrong answers only Why was Nvidia $NVDA down 10% today?	806	85	1573
NVIDIA STOCK WILL BE UNDER $200 IN A WEEK MARK THIS TWEET $NVDA	702	203	3832
Which company will be larger in 2030, $TSLA or $NVDA?	693	122	1777
WHO'S BUYING THE NEW TARIFF DIP MONDAY? $NVDA $153 to $93 ⛽ -39% DISCOUNT $PLTR $125 to $73 ⛽ -41% DISCOUNT $HOOD $67 to $34 ⛽ -49% DISCOUNT $SOFI $18 to $9 ⛽ -51% DISCOUNT $TSLA $488 to $237⛽-51% DISCOUNT $UPST $97 to $34 ⛽ -64%…	679	811	7251
To outperform $NVDA - get on the #Bitcoin Standard.	529	2022	13613

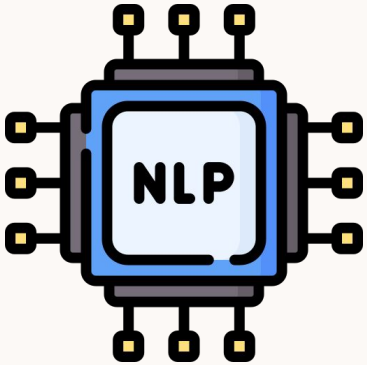
Challenges & Optimizations

- Data skew in modeling
- SQL pool cost management
- Feedback loop tuning
- Efficient API pagination



Future Work

- Weekend Tweet Utilization
- Complete Tweet Coverage
- Multi-stock analysis
- Anomaly detection
- Advanced NLP models



GameStop

TESLA

AMD

amc

Thank You!