ANOMALY DETECTION FRAMEWORK

Overview

1. *Our Solution*
2. *Data Processing*
3. *Anomaly Detection*
4. *Data Visualization*

INTRODUCTION

- This project aims to develop a real-time anomaly detection framework to identify potential issues and irregularities within data compared to regular patterns.

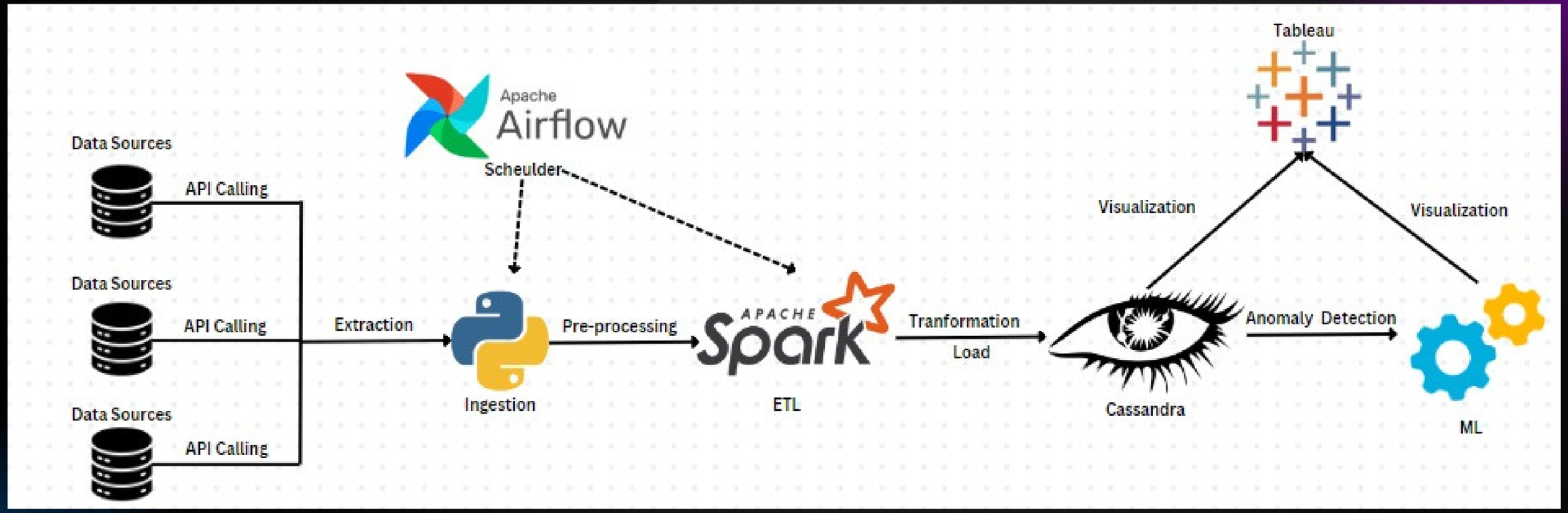
- By proactively flagging deviations, the system ensures data quality and facilitates early intervention for potential problems.

OUR SOLUTION 01

OUR SOLUTION

- *Data is loaded into an appropriate database system after being extracted, transformed, and loaded (ETL) from a variety of sources.*
- *Our system's foundation is Apache Spark, which allows for parallel distributed data processing to guarantee scalability and speed.*
- *We use the adtk package for anomaly detection, which uses past stock price trends to find anomalous patterns such outlier data points, spike levels, and volatility swings.*
- *Tableau is will be used to illustrate the findings of anomaly detection, giving users clear insights to quickly look into and fix abnormalities*
- *Airflow serves as a workflow management system, automating various stages of the data pipeline to ensure reliability and efficiency.*

SYSTEM ARCHITECTURE



DATA PIPELINE

Data Extraction	S1	S2	S3	S4	S5		
Transformation	S1	S2	S3	S4	S5	-	-
ML	-	S1	S2	S3	S4	S5	-
Load	-	-	S1	S2	S3	S4	S5
CYCLE	1	2	3	4	5	6	7

02

DATA PROCESSING

DATA SOURCES

Yahoo Finance

Upstox

TradeFeeds

```
from pyspark.sql import SparkSession

spark = SparkSession.builder\
        .appName("StockPrices")\
        .getOrCreate()

stock_prices_df = spark.read\
                  .option("header", "true"
                  .option("inferSchema", "
                  .csv("hdfs:///stock_pric
```

APACHE SPARK AS ETL FRAMEWORK

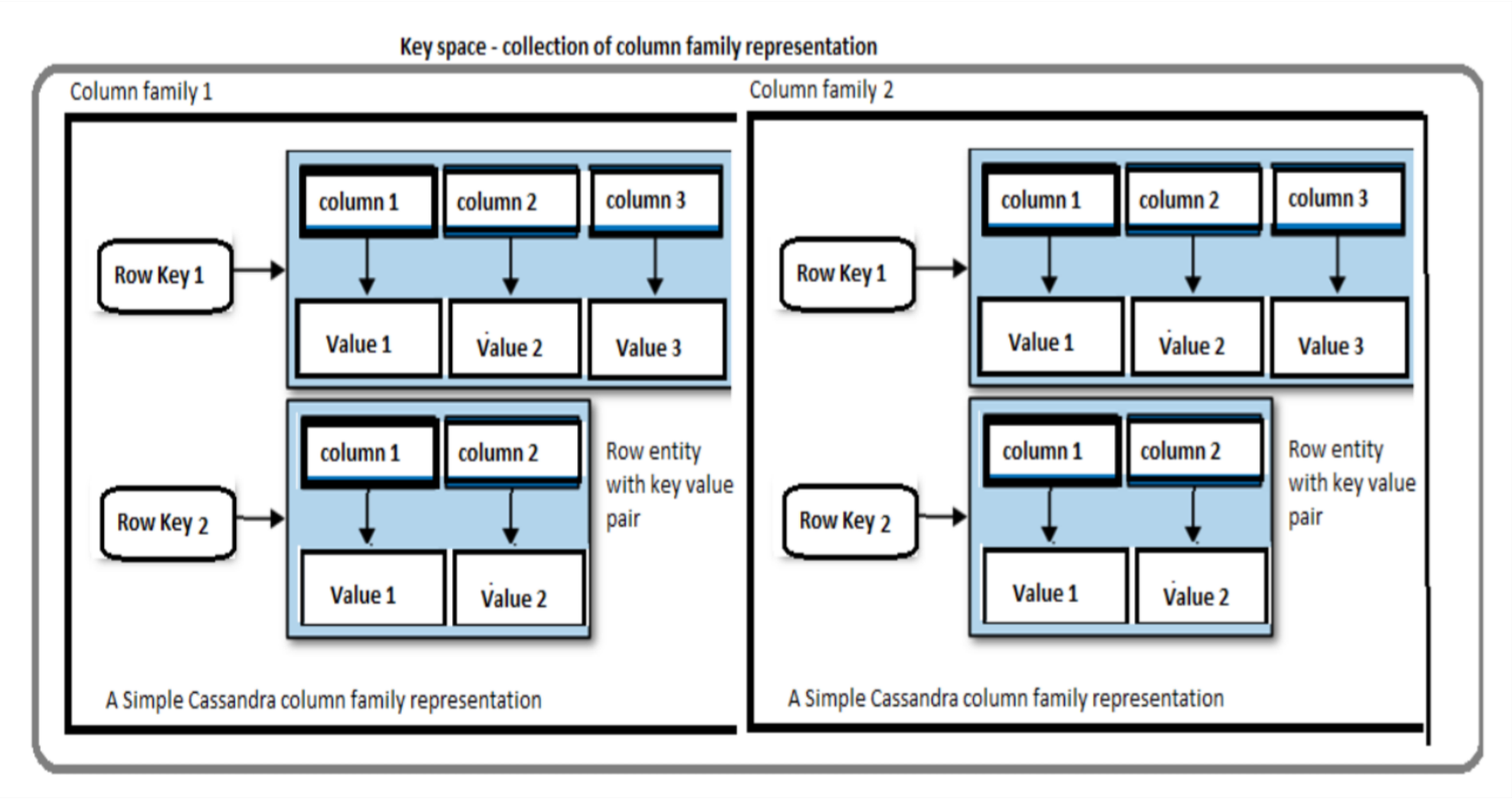
- *Apache Spark serves as the backbone of our system, enabling distributed processing of data in parallel.*
- *This facilitates efficient handling of large datasets, ensuring scalability and performance in anomaly detection tasks.*

PRE-PROCESSING & FEATURE ENGINEERING

Date	Open	High	Low	Close	Adj Close	Volume
1992-06-26	0.328125	0.347656	0.320313	0.335938	0.267131	224358400
1992-06-29	0.339844	0.367188	0.332031	0.359375	0.285767	58732800
1992-06-30	0.367188	0.371094	0.343750	0.347656	0.276449	34777600
1992-07-01	0.351563	0.359375	0.339844	0.355469	0.282661	18316800
1992-07-02	0.359375	0.359375	0.347656	0.355469	0.282661	13996800
...	...	...	...	...	...	...
2023-06-23	99.650002	99.730003	97.519997	98.339996	98.339996	18765000
2023-06-26	98.339996	98.769997	97.480003	98.230003	98.230003	6069000
2023-06-27	98.389999	99.059998	97.730003	98.720001	98.720001	5034100
2023-06-28	98.639999	98.639999	97.300003	98.610001	98.610001	6581900
2023-06-29	98.610001	98.830002	97.980003	98.680000	98.680000	5111200

- *Before feeding data into the anomaly detection engine, pre-processing steps are undertaken.*
- *This involves feature engineering to create relevant features from the existing data, enhancing the detection accuracy of anomalies.*

CASSANDRA DATABASE

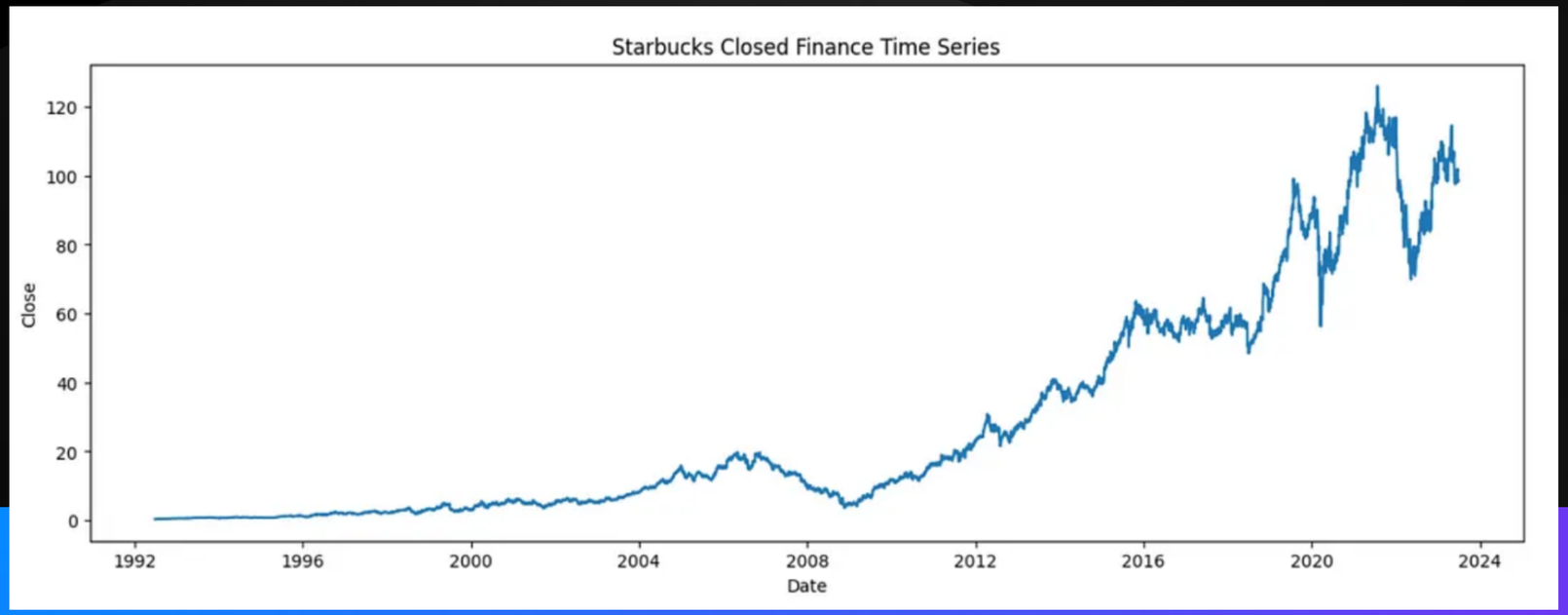


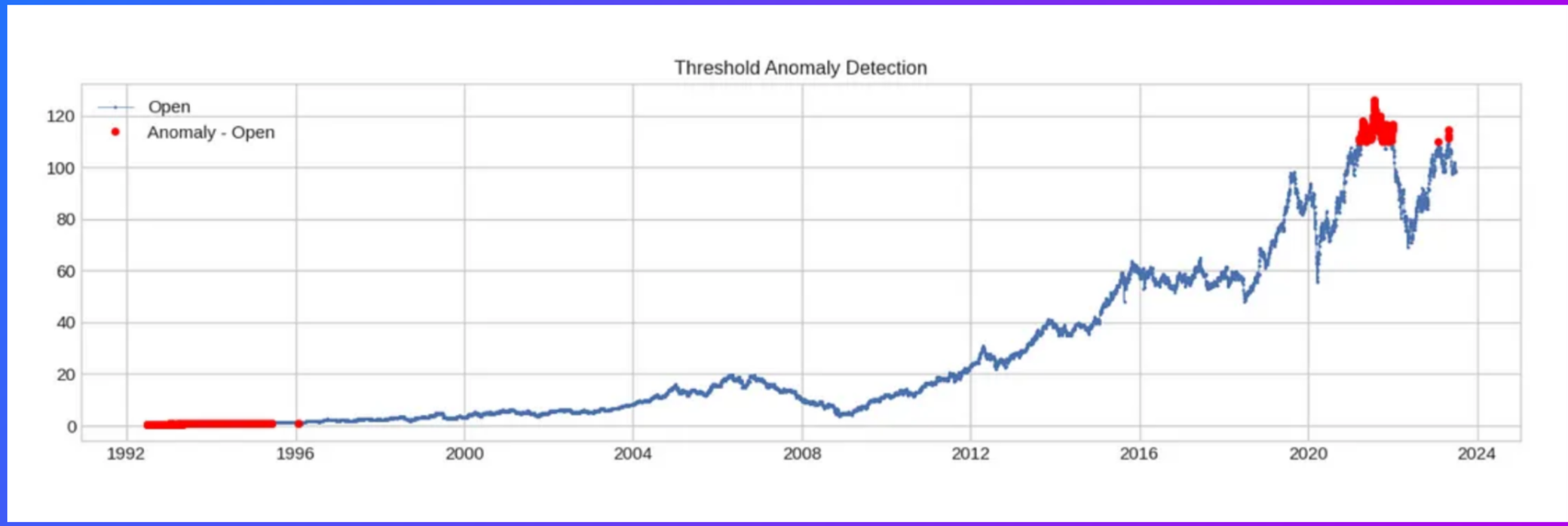
Cassandra's Data Storage Model

ANOMALY DETECTION

03

ANOMALY DETECTION

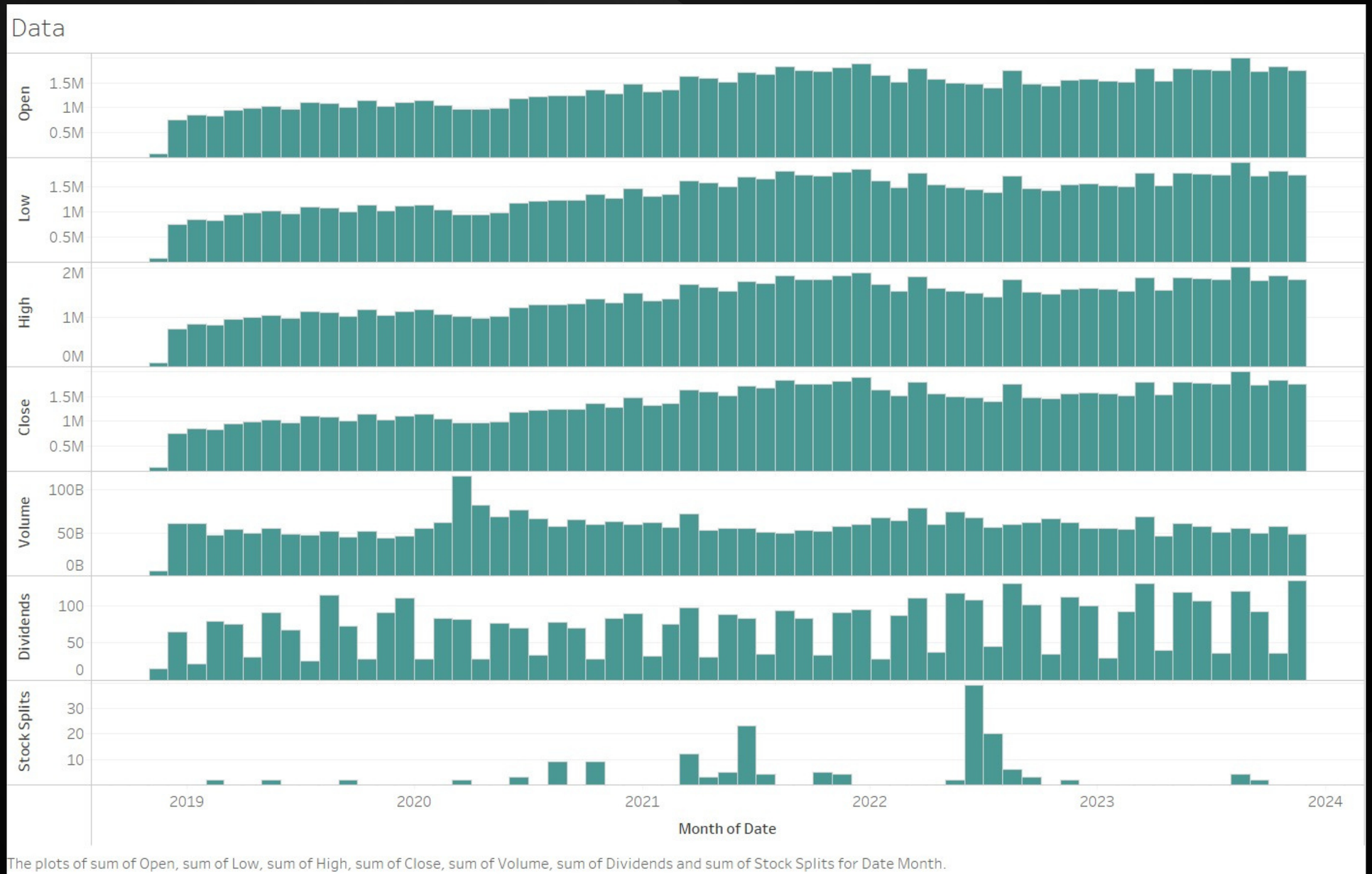




04

DATA VISUALISATION

VISUALISATION USING



AUTOMATION WITH *AIRFLOW*

- *Airflow serves as a workflow management system, automating the various stages of the data pipeline.*
- *From scheduling data extraction to anomaly detection, Airflow streamlines the entire process, ensuring reliability and efficiency in anomaly detection workflows*

Our team

PARTH PETKAR

ANSHU PARIHAR

PARTH KALANI

LEEVAN HERALD