

Stock Data Streaming And Analysis

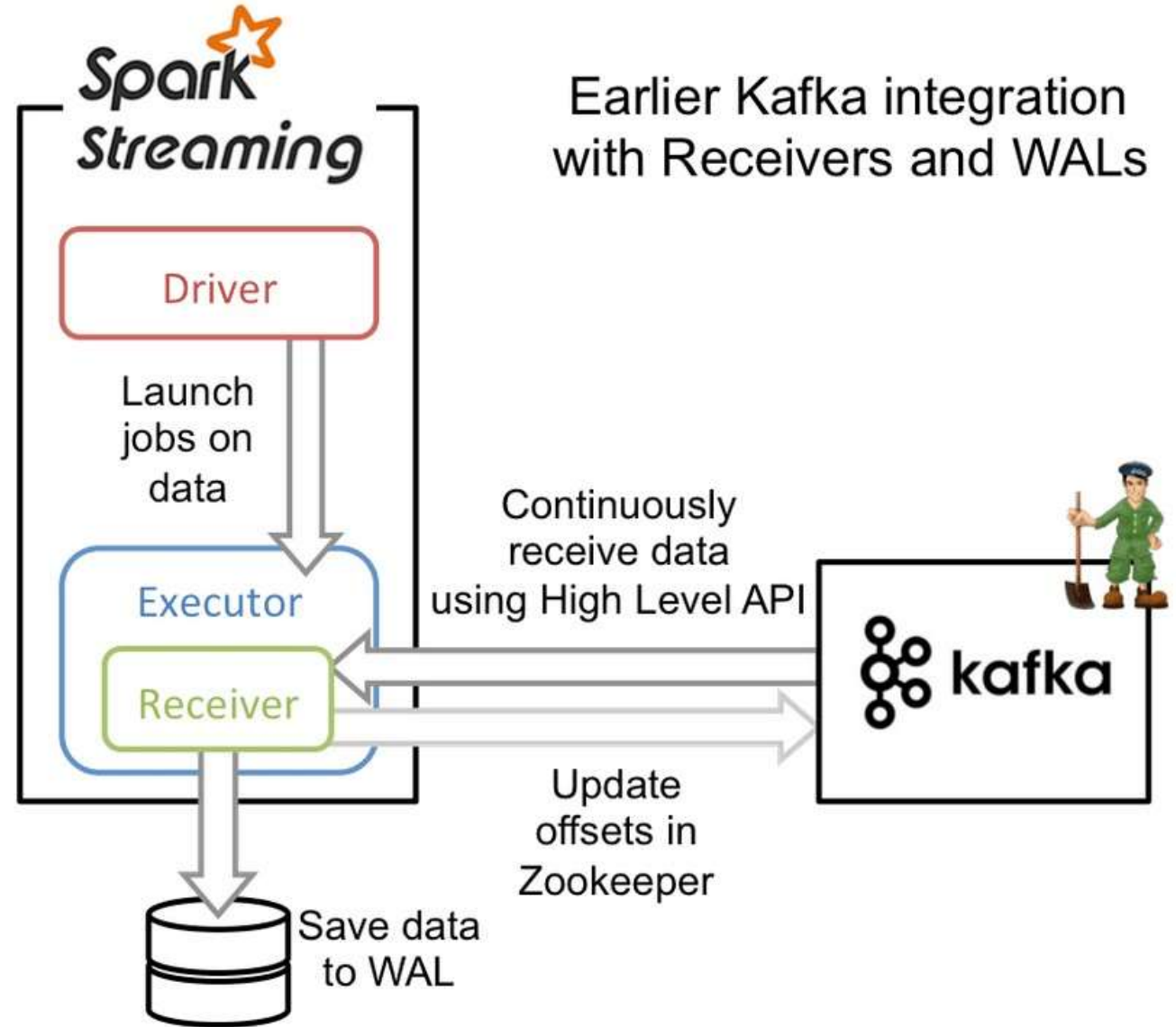
Introduction

- The stock market is a constellation of marketplaces where securities like stocks and bonds are bought and sold. Stock markets provide you with easy, transparent access to investment assets, and they help professional investors determine fair prices for public companies.
- Data analytics refers to the process of analyzing vast quantities of data to identify commonalities, insights, and trends. While data analytics tools can be used in many industries, including healthcare, politics, retail, banking, and government organizations, they're vital for competitiveness in modern financial markets.
- In finance, data analytics are crucial to interpret the ups and downs of capital markets. With great financial data science, traders and investment advisors earn the confidence to make informed decisions about buying, selling, or holding a particular security. It allows them to manage a portfolio based on short-, mid-, or long-term objectives.

Abstract

The main objective of the project is to visualize the analysis made from the live streaming of the stock data. The live data streaming from a stock analysis is initially used as the source data. Kafka streams the data from the Alpha Vintage website. The data stored in topics is then analyzed. For a variety of use scenarios, Jupyter notebook visualizes data analysis. The project has been improved with the aid of AWS services. The main objectives of the enhancement are to enhance data security and storage

System Architecture



Tools & Technologies

Streaming Tools

- Requests library
- Pandas datareader library
- API : Tiingo API, Alpha Vantage API
- Data Visualization Tool: Matplotlib, Seaborn



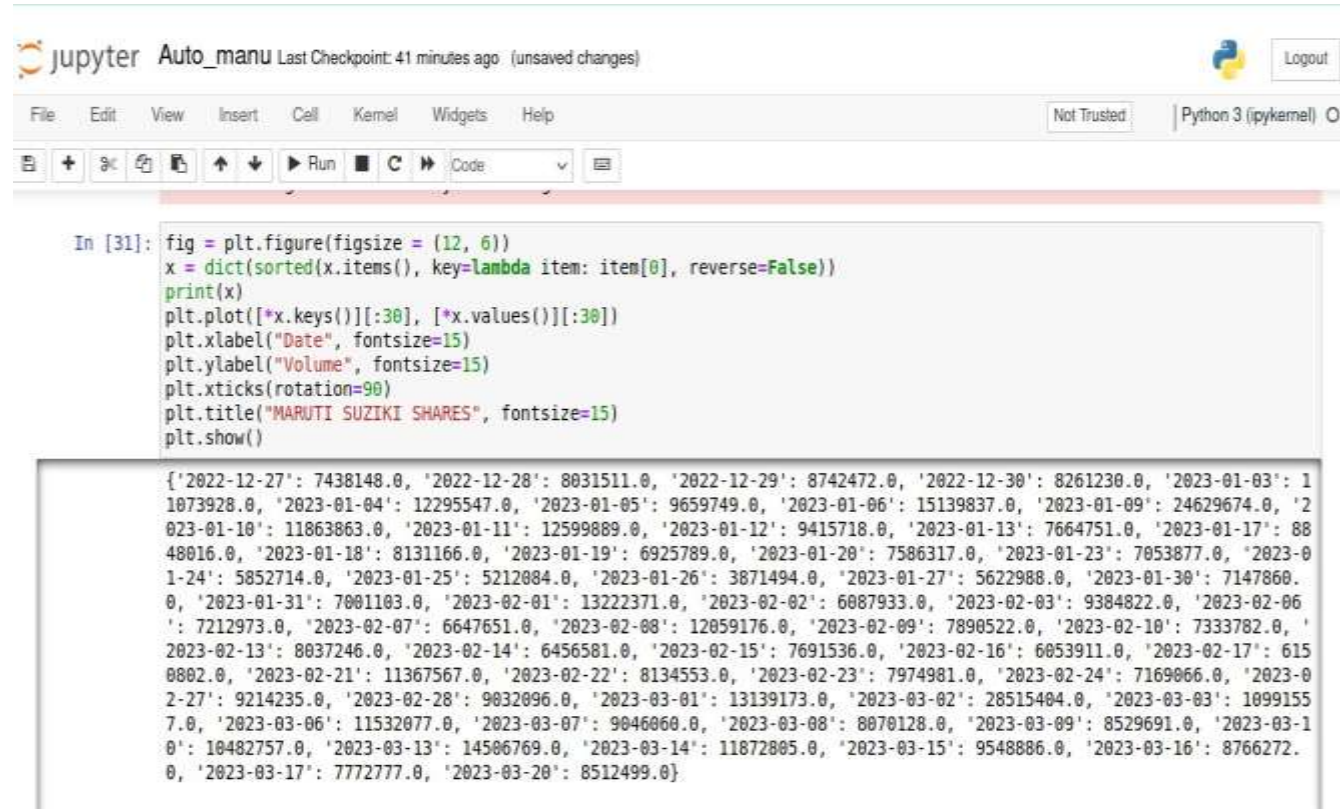
Technologies Used

- Spark
- Kafka
- Pyspark



Source Data

- The Live Streaming of data is extracted from AlphaVantage website.
- The data is converted into Json Format which contains Metadata and Time Series (Daily).
- Data Size is 2MB.



The screenshot shows a Jupyter Notebook interface with the title 'Auto_manu'. The top bar indicates 'Last Checkpoint: 41 minutes ago (unsaved changes)' and includes a 'Logout' button. The menu bar includes 'File', 'Edit', 'View', 'Insert', 'Cell', 'Kernel', 'Widgets', and 'Help'. The status bar shows 'Not Trusted' and 'Python 3 (ipykernel)'. The code cell contains the following Python code:

```
In [31]: fig = plt.figure(figsize = (12, 6))
x = dict(sorted(x.items(), key=lambda item: item[0], reverse=False))
print(x)
plt.plot([*x.keys()][:-30], [*x.values()][:-30])
plt.xlabel("Date", fontsize=15)
plt.ylabel("Volume", fontsize=15)
plt.xticks(rotation=90)
plt.title("MARUTI SUZUKI SHARES", fontsize=15)
plt.show()
```

The output of the code cell is a large JSON object representing a dictionary of dates and share volumes. The output is truncated in the middle of the image.

```
{'2022-12-27': 7438148.0, '2022-12-28': 8031511.0, '2022-12-29': 8742472.0, '2022-12-30': 8261230.0, '2023-01-03': 1
1073928.0, '2023-01-04': 12295547.0, '2023-01-05': 9659749.0, '2023-01-06': 15139837.0, '2023-01-09': 24629674.0, '2
023-01-10': 11863863.0, '2023-01-11': 12599889.0, '2023-01-12': 9415718.0, '2023-01-13': 7664751.0, '2023-01-17': 88
48016.0, '2023-01-18': 8131166.0, '2023-01-19': 6925789.0, '2023-01-20': 7586317.0, '2023-01-23': 7053877.0, '2023-0
1-24': 5852714.0, '2023-01-25': 5212884.0, '2023-01-26': 3871494.0, '2023-01-27': 5622988.0, '2023-01-30': 7147860.
0, '2023-01-31': 7001103.0, '2023-02-01': 13222371.0, '2023-02-02': 6087933.0, '2023-02-03': 9384822.0, '2023-02-06
': 7212973.0, '2023-02-07': 6647651.0, '2023-02-08': 12059176.0, '2023-02-09': 7890522.0, '2023-02-10': 7333782.0, '
2023-02-13': 8037246.0, '2023-02-14': 6456581.0, '2023-02-15': 7691536.0, '2023-02-16': 6053911.0, '2023-02-17': 615
0802.0, '2023-02-21': 11367567.0, '2023-02-22': 8134553.0, '2023-02-23': 7974981.0, '2023-02-24': 7169066.0, '2023-0
2-27': 9214235.0, '2023-02-28': 9032096.0, '2023-03-01': 13139173.0, '2023-03-02': 28515404.0, '2023-03-03': 1099155
7.0, '2023-03-06': 11532077.0, '2023-03-07': 9046060.0, '2023-03-08': 8070128.0, '2023-03-09': 8529691.0, '2023-03-1
0': 10482757.0, '2023-03-13': 14506769.0, '2023-03-14': 11872805.0, '2023-03-15': 9548886.0, '2023-03-16': 8766272.
0, '2023-03-17': 7772777.0, '2023-03-20': 8512499.0}
```


Source Code

```
2 from kafka import KafkaProducer
3 import requests
4 from json import dumps
5 import time
6
7 def on_message1(message):
8     producer1.send('m', message)
9     producer1.flush()
10
11 producer1 = KafkaProducer(value_serializer=lambda m: dumps(m).encode('utf-8'), bootstrap_servers=['localhost:9092'])
12
13
14 # url to collect data of Mahindra and Mahindra shares
15 url = 'https://www.alphavantage.co/query?function=TIME_SERIES_DAILY&symbol=M&M.BSE&apikey=36S8YNI4H3XTBL12&outputsize=full'
16 r = requests.get(url)
17 data=r.json()
18 #print(data)
19 #preprocessing
20 del(data['Meta Data'])
21 df=data['Time Series (Daily)']
22 #print(df)
23 jsonFile={}
24 j=0
25 for i in df.keys():
26     jsonFile[j]=df[i]
27     jsonFile[j]['0. date']=i
28     #print(i)
29     j+=1
30 print(jsonFile)
31 for i in jsonFile.keys():
32     jsonFile[i]['open']=jsonFile[i]['1. open']
33     jsonFile[i]['high']=jsonFile[i]['2. high']
34     jsonFile[i]['low']=jsonFile[i]['3. low']
35     jsonFile[i]['close']=jsonFile[i]['4. close']
36     jsonFile[i]['volume']=jsonFile[i]['5. volume']
37     jsonFile[i]['date']=jsonFile[i]['0. date']
38     del jsonFile[i]['1. open']
39     del jsonFile[i]['2. high']
40     del jsonFile[i]['3. low']
41     del jsonFile[i]['4. close']
42     del jsonFile[i]['5. volume']
```

```
14 from pyspark.sql import SparkSession
15 from pyspark.sql.functions import *
16 from pyspark.sql.types import *
17
18 spark = SparkSession \
19     .builder \
20     .appName("Manufacturing_companies") \
21     .getOrCreate()
22
23 raw_df = spark \
24     .readStream \
25     .format("kafka") \
26     .option("kafka.bootstrap.servers", "localhost:9092") \
27     .option("subscribe", "m") \
28     .option("startingOffsets", "latest") \
29     .load() \
30     .selectExpr("CAST(value AS STRING)")
31
32 # Created Schema for Structured Streaming
33
34 # timestamp,open,high,low,close,volume
35
36 schema = StructType(
37     [
38         StructField("open", StringType()),
39         StructField("high", StringType()),
40         StructField("low", StringType()),
41         StructField("close", StringType()),
42         StructField("volume", StringType()),
43         StructField("date", StringType())
44     ]
45 )
46 # Applied schema on data
47 data1= raw_df.select(from_json(raw_df.value, schema).alias("data"))
48
49 # Fetching required data for processing queries
50 output_df1 = data1.select(to_json(struct(col("data.volume"),col("data.date"))).alias("value"))
51
52 # Sending the data to kafka broker
53 query = output_df1.writeStream.format("kafka").option("kafka.bootstrap.servers", "localhost:9092").option("checkpointLocation", "/tmp/checkpoint1").option("topic", "m1").start()
```


Zookeeper Initiation

```
ubuntu@ubuntu2004: ~/Downloads/kafka_2.11-2.0.0$ bin/zookeeper-server-start.sh config/zookeeper.properties
2023-04-04 22:18:37,648] INFO Reading configuration from: config/zookeeper.properties (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
2023-04-04 22:18:37,663] INFO autopurge.snapRetainCount set to 3 (org.apache.zookeeper.server.DataDirCleanupManager)
2023-04-04 22:18:37,664] INFO autopurge.purgeInterval set to 0 (org.apache.zookeeper.server.DataDirCleanupManager)
2023-04-04 22:18:37,664] INFO Purge task is not scheduled. (org.apache.zookeeper.server.DataDirCleanupManager)
2023-04-04 22:18:37,664] WARN Either no config or no quorum defined in config, running in standalone mode (org.apache.zookeeper.server.quorum.QuorumPeerMain)
2023-04-04 22:18:37,775] INFO Reading configuration from: config/zookeeper.properties (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
2023-04-04 22:18:37,776] INFO Starting server (org.apache.zookeeper.server.ZooKeeperServerMain)
2023-04-04 22:18:37,830] INFO Server environment:zookeeper.version=3.4.13-2d71af4dbe22557fda74f9a9b4309b15a7487f03, built on 06/29/2018 00:39 GMT (org.apache.zookeeper.server.ZooKeeperServer)
2023-04-04 22:18:37,831] INFO Server environment:host.name=ubuntu2004.linuxvmimages.local (org.apache.zookeeper.server.ZooKeeperServer)
2023-04-04 22:18:37,831] INFO Server environment:java.version=1.8.0_362 (org.apache.zookeeper.server.ZooKeeperServer)
2023-04-04 22:18:37,831] INFO Server environment:java.vendor=Private Build (org.apache.zookeeper.server.ZooKeeperServer)
2023-04-04 22:18:37,831] INFO Server environment:java.home=/usr/lib/jvm/java-8-openjdk-amd64/jre (org.apache.zookeeper.server.ZooKeeperServer)
2023-04-04 22:18:37,832] INFO Server environment:java.class.path=/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/activation-1.1.1.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/aopalliance-repackaged-2.5.0-b42.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/argparse4j-0.7.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/audience-annotations-0.5.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/commons-lang3-3.5.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/connect-api-2.0.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/connect-basic-auth-extension-2.0.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/connect-file-2.0.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/connect-json-2.0.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/connect-runtime-2.0.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/connect-transforms-2.0.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/guava-20.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/hk2-api-2.5.0-b42.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/hk2-locator-2.5.0-b42.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/hk2-utils-2.5.0-b42.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jackson-annotations-2.9.6.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jackson-core-2.9.6.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jackson-data-bind-2.9.6.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jackson-jaxrs-base-2.9.6.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jackson-jaxrs-json-provider-2.9.6.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jackson-module-jaxb-annotations-2.9.6.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/javassist-3.22.0-CR2.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/javax.annotation-api-1.2.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/javax.inject-1.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/javax.inject-2.5.0-b42.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/javax.servlet-api-3.1.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/javax.ws.rs-api-2.1.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jaxb-api-2.3.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jersey-client-2.27.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jersey-common-2.27.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jersey-container-servlet-2.27.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jersey-container-servlet-core-2.27.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jersey-hk2-2.27.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jersey-media-jaxb-2.27.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jersey-server-2.27.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jetty-client-9.4.11.v20180605.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jetty-continuation-9.4.11.v20180605.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jetty-io-9.4.11.v20180605.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jetty-security-9.4.11.v20180605.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jetty-server-9.4.11.v20180605.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jetty-servlet-9.4.11.v20180605.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jetty-servlets-9.4.11.v20180605.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jetty-util-9.4.11.v20180605.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jopt-simple-5.0.4.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/kafka_2.11-2.0.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/kafka-2.11-2.0.0-sources.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/kafka-clients-2.0.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/kafka-log4j-appender-2.0.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/kafka-streams-2.0.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/kafka-streams-examples-2.0.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/kafka-streams-scala-2.11-2.0.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/kafka-streams-test-utils-2.0.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/kafka-tools-2.0.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/log4j-1.2.17.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/lz4-java-1.4.1.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/maven-artifact-3.5.3.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/metrics-core-2.2.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/osgi-resource-locator-1.0.1.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/plexus-utils-3.1.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/reflections-0.9.11.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/rocksdbjni-5.7.3.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/scala-library-2.11.12.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/scala-logging-2.11-3.9.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/scala-reflect-2.11.12.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/slf4j-api-1.7.25.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/slf4j-log4j12-1.7.25.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/snappy-java-1.1.7.1.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/validation-api-1.0.Final.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/zookeeper-3.4.13.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/zookeeper-jmx-3.4.13.jar
```


Kafka Initiation

```
ubuntu@ubuntu2004: ~/Downloads/kafka_2.11-2.0.0
ubuntu@ubuntu2004: ~/Downloads/kafka_2.11-2.0.0
ubuntu@ubuntu2004:~/Downloads/kafka_2.11-2.0.0$ bin/kafka-server-start.sh config/server.properties
2023-04-04 22:19:35,170] INFO Registered kafka:type=kafka.Log4jController MBean (kafka.utils.Log4jControllerRegistration$)
2023-04-04 22:19:36,938] INFO starting (kafka.server.KafkaServer)
2023-04-04 22:19:36,945] INFO Connecting to zookeeper on localhost:2181 (kafka.server.KafkaServer)
2023-04-04 22:19:37,034] INFO [ZooKeeperClient] Initializing a new session to localhost:2181. (kafka.zookeeper.ZooKeeperClient)
2023-04-04 22:19:37,079] INFO Client environment:zookeeper.version=3.4.13-2d71af4dbe22557fda74f9a9b4309b15a7487f03, built on 06/29/2018 00:39 GMT (org.apache.zookeeper.ZooKeeper)
2023-04-04 22:19:37,081] INFO Client environment:host.name=ubuntu2004.linuxvmimages.local (org.apache.zookeeper.ZooKeeper)
2023-04-04 22:19:37,082] INFO Client environment:java.version=1.8.0_362 (org.apache.zookeeper.ZooKeeper)
2023-04-04 22:19:37,082] INFO Client environment:java.vendor=Private Build (org.apache.zookeeper.ZooKeeper)
2023-04-04 22:19:37,082] INFO Client environment:java.home=/usr/lib/jvm/java-8-openjdk-amd64/jre (org.apache.zookeeper.ZooKeeper)
2023-04-04 22:19:37,084] INFO Client environment:java.class.path=/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/activation-1.1.1.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/aopalliance-repackaged-2.5.0-b42.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/argparse4j-0.7.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/audience-annotations-0.5.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/commons-lang3-3.5.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/connect-api-2.0.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/connect-basic-auth-extension-2.0.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/connect-file-2.0.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/connect-json-2.0.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/connect-runtime-2.0.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/connect-transforms-2.0.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/guava-20.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/hk2-api-2.5.0-b42.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/hk2-locator-2.5.0-b42.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/hk2-utils-2.5.0-b42.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jackson-annotations-2.9.6.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jackson-core-2.9.6.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jackson-data-bind-2.9.6.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jackson-jaxrs-base-2.9.6.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jackson-jaxrs-json-provider-2.9.6.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jackson-module-jaxb-annotations-2.9.6.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/javax.annotation-api-1.2.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/javax.inject-1.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/javax.inject-2.5.0-b42.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/javax.servlet-api-3.1.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/java.ws.rs-api-2.1.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jaxb-api-2.3.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jersey-client-2.27.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jersey-common-2.27.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jersey-container-servlet-2.27.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jersey-container-servlet-core-2.27.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jersey-hk2-2.27.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jersey-media-jaxb-2.27.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jersey-server-2.27.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jetty-client-9.4.11.v20180605.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jetty-continuation-9.4.11.v20180605.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jetty-http-9.4.11.v20180605.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jetty-io-9.4.11.v20180605.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jetty-security-9.4.11.v20180605.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jetty-server-9.4.11.v20180605.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jetty-servlet-9.4.11.v20180605.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jetty-servlets-9.4.11.v20180605.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jetty-util-9.4.11.v20180605.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/jopt-simple-5.0.4.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/kafka_2.11-2.0.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/kafka_2.11-2.0.0-sources.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/kafka-clients-2.0.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/kafka-log4j-appender-2.0.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/kafka-streams-2.0.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/kafka-streams-examples-2.0.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/kafka-streams-scala_2.11-2.0.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/kafka-streams-test-utils-2.0.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/kafka-tools-2.0.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/log4j-1.2.17.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/lz4-java-1.4.1.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/maven-artifact-3.5.3.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/metrics-core-2.2.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/osgi-resource-locator-1.0.1.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/plexus-utils-3.1.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/reflections-0.9.11.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/rocksdbjni-5.7.3.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/scala-library-2.11.12.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/scala-logging_2.11-3.9.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/scala-reflect-2.11.12.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/slf4j-api-1.7.25.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/slf4j-log4j12-1.7.25.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/snappy-java-1.1.7.1.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/validation-api-1.1.0.Final.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/zkclient-0.10.0.jar:/home/ubuntu/Downloads/kafka_2.11-2.0.0/bin/../libs/zookeeper-3.4.13.jar (org.apache.zookeeper.ZooKeeper)
```


Topic Creation

```
ubuntu@ubuntu2004:~/Downloads/kafka_2.11-2.0.0$ bin/kafka-topics.sh --create --zookeeper localhost:2181 --topic mm --replication-factor 1 --partitions 1
Created topic "mm".
ubuntu@ubuntu2004:~/Downloads/kafka_2.11-2.0.0$ bin/kafka-topics.sh --create --zookeeper localhost:2181 --topic mm1 --replication-factor 1 --partitions 1
Created topic "mm1".
ubuntu@ubuntu2004:~/Downloads/kafka_2.11-2.0.0$
```

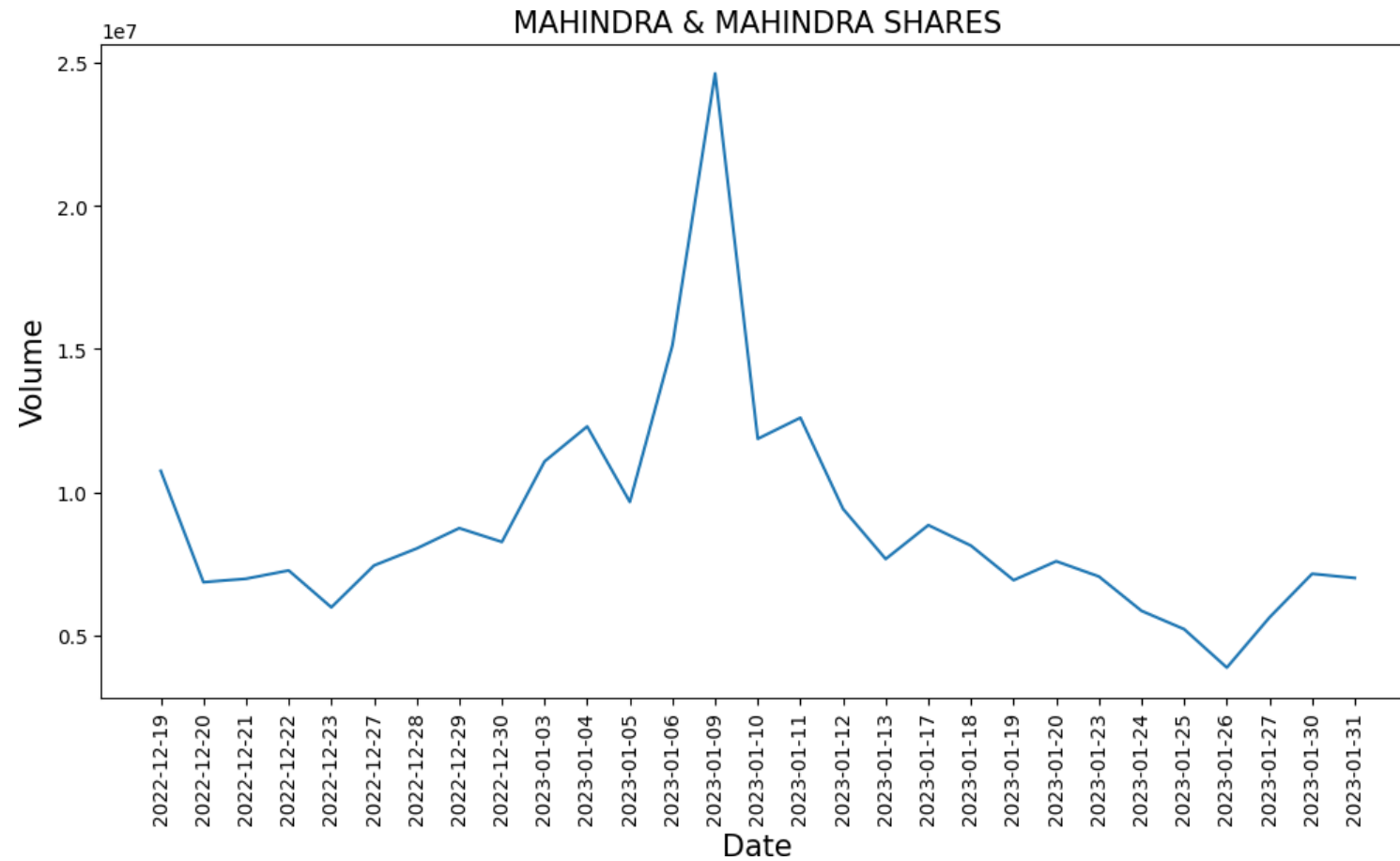
Preprocessed Data

```
{ '12612617' }, '2022-11-09': { '1. open': '19.5000', '2. high': '19.6000', '3. low': '18.7200', '4. close': '19.0900', 'low': '18.5400', 'close': '18.9200', 'volume': '14958689', 'date': '2023-04-04'}, 1: { 'open': '17.1100', 'high': '17.6400', 'low': '16.9800', 'close': '17.8000', 'high': '18.0000', 'low': '16.9600', 'close': '16.9800', 'volume': '10343927', 'date': '2023-03-29'}, 2: { 'open': '17.3100', 'volume': '14749419', 'date': '2023-03-29'}, 5: { 'open': '17.2200', 'high': '17.9300', 'low': '16.9400', 'close': '16.9400', 'volume': '10945793', 'date': '2023-03-24'}, 8: { 'open': '18.1950', 'high': '18.9200', 'low': '17.9800', 'close': '18.0000', 'volume': '10211574', 'date': '2023-03-21'}, 11: { 'open': '19.0100', 'date': '2023-03-20'}, 12: { 'open': '18.8200', 'high': '18.9600', 'low': '18.5404', 'close': '18.8200', 'low': '18.7300', 'close': '18.7500', 'volume': '8766272', 'date': '2023-03-16'}, 14: { 'open': '19.0350', 'low': '18.2800', 'close': '18.9400', 'volume': '9548886', 'date': '2023-03-15'}, 15: { 'open': '18.9900', 'high': '19.2300', 'low': '18.4500', 'close': '19.0000', 'high': '19.2100', 'low': '18.1852', 'close': '18.6500', 'volume': '14506769', 'date': '2023-03-10'}, 18: { 'open': '20.9400', 'high': '21.2050', 'low': '20.7600', 'close': '21.0400', 'volume': '8070128', 'date': '2023-03-07'}, 21: { 'open': '22.8700', 'high': '23.0200', 'low': '22.5700', 'close': '22.6600', 'volume': '9046060', 'date': '2023-03-06'}, 22: { 'open': '22.9300', 'high': '23.0200', 'low': '22.5700', 'close': '22.6600', 'volume': '28515404', 'date': '2023-03-02'}, 24: { 'open': '20.9173', 'date': '2023-03-01'}, 25: { 'open': '20.5100', 'high': '20.8000', 'low': '20.3400', 'close': '20.4100', 'volume': '9214235', 'date': '2023-02-27'}, 27: { 'open': '20.9600', 'high': '21.1200', 'low': '20.5600', 'close': '20.8400', 'volume': '8134553', 'date': '2023-02-21'}, 31: { 'open': '22.3400', 'high': '22.4100', 'low': '20.8200', 'volume': '11367567', 'date': '2023-02-21'}, 32: { 'open': '22.3800', 'high': '22.7300', 'low': '22.2600', 'close': '22.2900', 'volume': '6053911', 'date': '2023-02-15'}, 34: { 'open': '22.2900', 'high': '22.7800', 'volume': '7691536', 'date': '2023-02-14'}, 35: { 'open': '22.0800', 'high': '22.3700', 'low': '21.8100', 'close': '22.3600', 'volume': '7333782', 'date': '2023-02-10'}, 37: { 'open': '22.5300', 'date': '2023-02-09'}, 38: { 'open': '23.6900', 'high': '23.6950', 'low': '22.1200', 'close': '22.1200', 'high': '24.0300', 'low': '23.4000', 'close': '23.9900', 'volume': '6647651', 'date': '2023-02-07'}, 40: { 'open': '24.3000', 'high': '25.1200', 'low': '24.2650', 'close': '24.4100', 'volume': '6087933', 'date': '2023-02-06'}, 41: { 'open': '24.7700', 'low': '23.9700', 'close': '24.4100', 'volume': '13222371', 'date': '2023-02-01'}, 44: { 'open': '23.4300', 'high': '23.7700', 'low': '23.2000', 'high': '23.6650', 'low': '23.0400', 'close': '23.3200', 'volume': '7147860', 'date': '2023-01-27'}, 47: { 'open': '23.6500', 'high': '23.6500', 'low': '23.0800', 'volume': '5622988', 'date': '2023-01-26'}, 48: { 'open': '23.0200', 'high': '23.4800', 'low': '22.6810', 'close': '23.3800', 'volume': '5852714', 'date': '2023-01-24'}, 50: { 'open': '22.9300', 'date': '2023-01-23'}, 51: { 'open': '22.3900', 'high': '23.1099', 'low': '22.2600', 'close': '22.7600', 'low': '22.9600', 'low': '22.3000', 'close': '22.7800', 'volume': '6925789', 'date': '2023-01-19'}, 53: { 'open': '22.8300', 'high': '23.0470', 'low': '22.6900', 'close': '23.0100', 'volume': '7664751', 'date': '2023-01-18'}, 54: { 'open': '20.9300', 'high': '22.2550', 'low': '20.4700', 'high': '20.8900', 'low': '20.3200', 'close': '20.8600', 'volume': '11863863', 'date': '2023-01-12'}, 57: { 'open': '20.4700', 'high': '20.8900', 'low': '20.3200', 'close': '20.8600', 'volume': '11863863', 'date': '2023-01-12'}, 60: { 'open': '21.7400', 'high': '22.7400', 'low': '21.7400', 'close': '22.7400', 'volume': '12612617', 'date': '2022-11-09' }
```

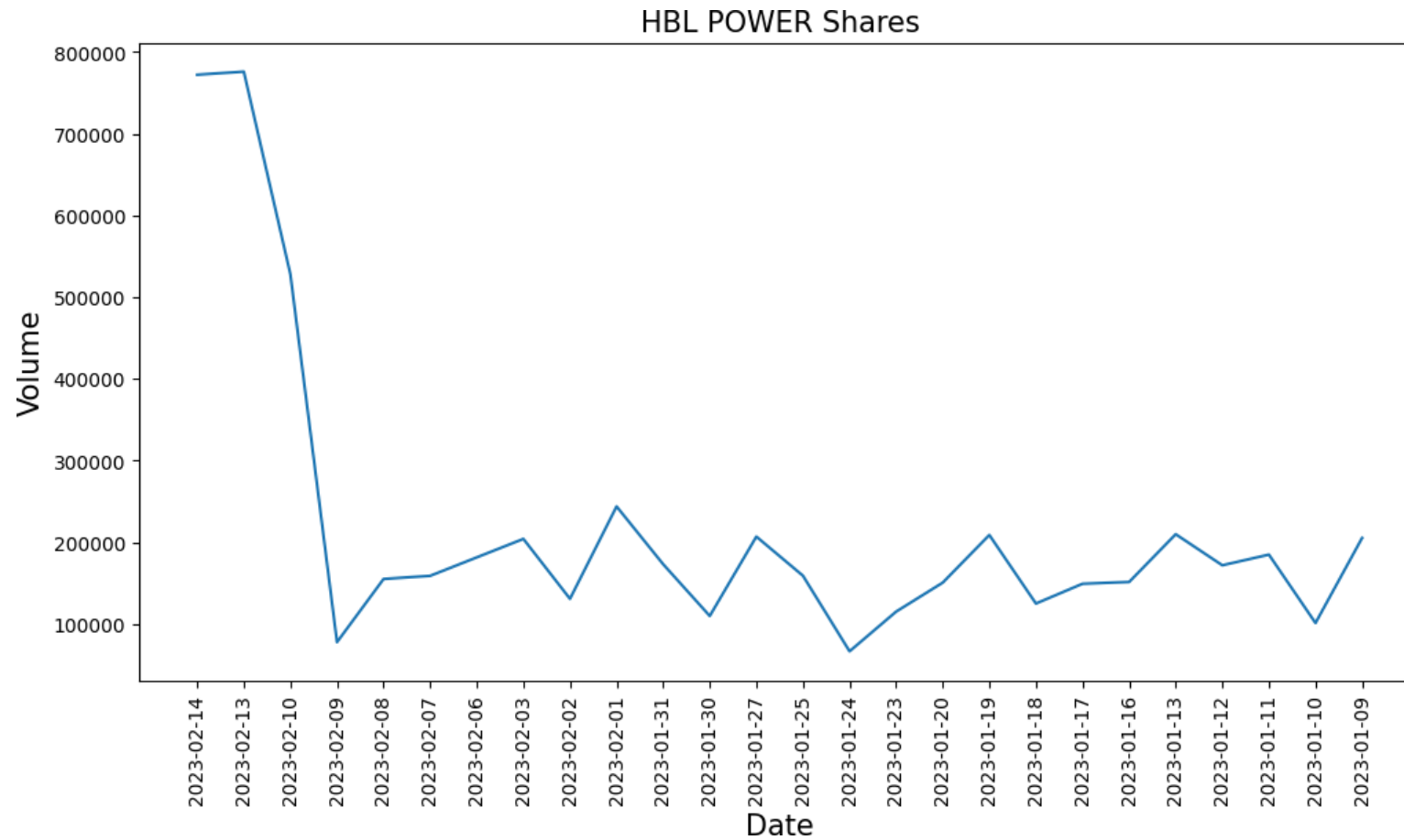

Spark - Submit

```
ubuntu@ubuntu2004:~/Downloads/spark-3.1.2-bin-hadoop3.2/bin$ spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.12:3.1.2 /home/ubuntu/Downloads/Mahendra/mm1.py
23/04/04 22:34:32 WARN Utils: Your hostname, ubuntu2004 resolves to a loopback address: 127.0.1.1; using 192.168.86.42 instead (on interface ens33)
23/04/04 22:34:32 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
:: loading settings :: url = jar:file:/home/ubuntu/Downloads/spark-3.1.2-bin-hadoop3.2/jars/ivy-2.4.0.jar!/org/apache/ivy/core/settings/ivysettings.xml
Ivy Default Cache set to: /home/ubuntu/.ivy2/cache
The jars for the packages stored in: /home/ubuntu/.ivy2/jars
org.apache.spark#spark-sql-kafka-0-10_2.12 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-77eeae84-0ce2-4520-93f6-0617be7a116c;1.0
   confs: [default]
   found org.apache.spark#spark-sql-kafka-0-10_2.12;3.1.2 in central
   found org.apache.spark#spark-token-provider-kafka-0-10_2.12;3.1.2 in central
   found org.apache.kafka#kafka-clients;2.6.0 in central
   found com.github.luben#zstd-jni;1.4.8-1 in central
   found org.lz4#lz4-java;1.7.1 in central
   found org.xerial.snappy#snappy-java;1.1.8.2 in central
   found org.slf4j#slf4j-api;1.7.30 in central
   found org.spark-project.spark#unused;1.0.0 in central
   found org.apache.commons#commons-pool2;2.6.2 in central
:: resolution report :: resolve 2772ms :: artifacts dl 200ms
:: modules in use:
com.github.luben#zstd-jni;1.4.8-1 from central in [default]
org.apache.commons#commons-pool2;2.6.2 from central in [default]
org.apache.kafka#kafka-clients;2.6.0 from central in [default]
org.apache.spark#spark-sql-kafka-0-10_2.12;3.1.2 from central in [default]
org.apache.spark#spark-token-provider-kafka-0-10_2.12;3.1.2 from central in [default]
org.lz4#lz4-java;1.7.1 from central in [default]
org.slf4j#slf4j-api;1.7.30 from central in [default]
org.spark-project.spark#unused;1.0.0 from central in [default]
org.xerial.snappy#snappy-java;1.1.8.2 from central in [default]
-----
|               |          modules          |      artifacts      |
|   conf   | number | search | dwnlded | evicted | || number | dwnlded |
|-----|-----|-----|-----|-----|
|   default |    9   |    0   |    0   |    0   | ||    9   |    0   |
|-----|-----|-----|-----|-----|
:: retrieving :: org.apache.spark#spark-submit-parent-77eeae84-0ce2-4520-93f6-0617be7a116c
   confs: [default]
   0 artifacts copied, 9 already retrieved (0kB/264ms)
23/04/04 22:34:42 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
python3: can't open file '/home/ubuntu/Downloads/Mahendra/mm1.py': [Errno 2] No such file or directory
log4j:WARN No appenders could be found for logger (org.apache.spark.util.ShutdownHookManager).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
ubuntu@ubuntu2004:~/Downloads/spark-3.1.2-bin-hadoop3.2/bin$
```

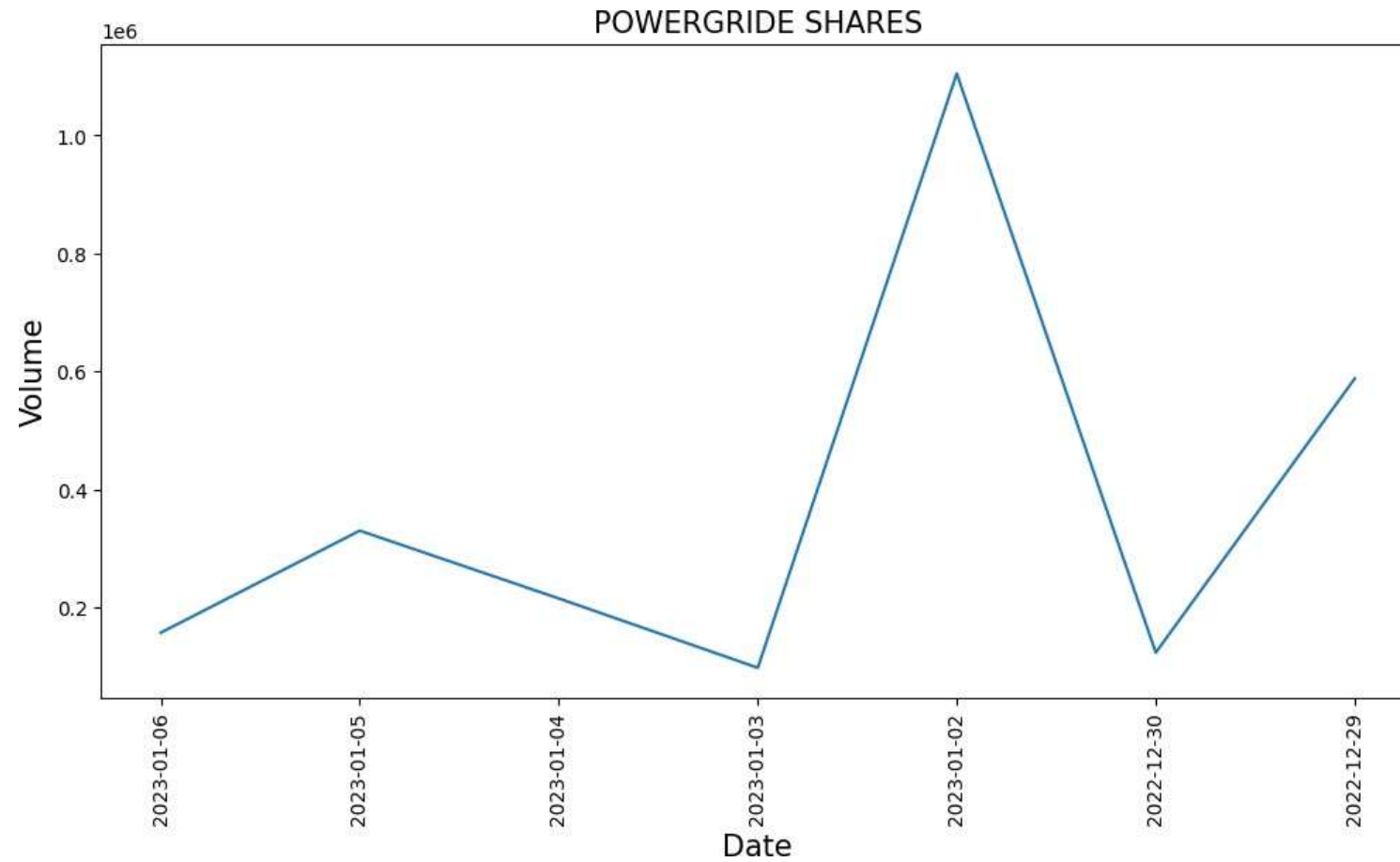
Mahendra_Auto_Manufactures



HBL_Power_Battery_Company



Powergrid_Charging_Point

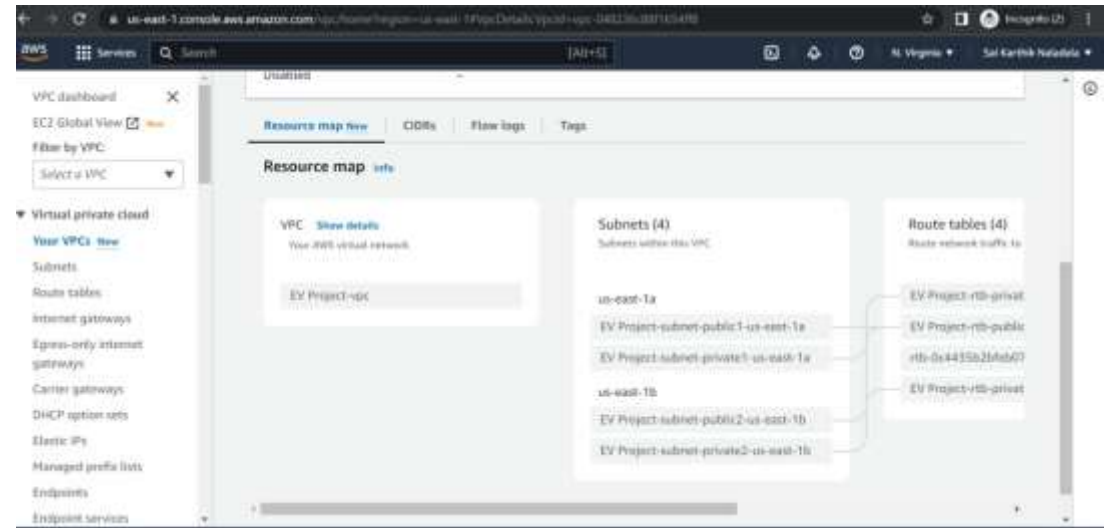


AWS Services



VPC Creation

- Firstly we create a separate VPC for our project with minimum two public and private subnets, though we are deploying our application private subnet but to connect to internet we need public subnets as well.
- We have created my VPC 'EV Project-vpc' and now as we need our private subnet to connect to internet for installing required libraries in our EC2 instance we have to create NAT gateway



Creation of NAT Gateway

NAT gateway settings

Name - optional
Create a tag with a key of Name and a value that you specify.
my-nat-gateway-01
The name can be up to 255 characters long.

Subnet
Select a subnet in which to create the NAT gateway.
subnet-0a01a07a16e066eb2 (EV Project-subnet-public1-us-east-1a)

Connectivity type
Select a connectivity type for the NAT gateway.
☒ Public
☐ Private

Elastic IP allocation ID [info](#)
Assign an Elastic IP address to the NAT gateway.
eipalloc-07130749c36d57d3a [Allocate Elastic IP](#)

[Additional settings](#) [info](#)

VPC > Route tables > rtb-04e4cf13d1ae0592f > Edit routes

Edit routes

Destination	Target	Status	Propagated
pl-63a5400a	vpc-01ff2ca6a37b16fc2	Active	No
10.0.0.0/16	local	Active	No
0.0.0.0/0	nat-098fdebeeb1409ceb	Active	No

[Add route](#) [Remove](#)

[Cancel](#) [Preview](#) [Save changes](#)

Launching MSK Cluster

- Created an Amazon MSK (Amazon Managed Streaming for Apache Kafka), which will stream the data through Kafka. To optimize cost we have launched only two brokers which will be running in each private subnet.



Kinesis Data Firehose

► Amazon Kinesis Data Firehose: How it works

Choose source and destination

Specify the source and the destination for your delivery stream. You cannot change the source and destination of your delivery stream once it has been created.

Source [Info](#)

Direct PUT ▼

Destination [Info](#)

Amazon S3 ▼

Delivery stream name

Delivery stream name

PUT-S3-xhwjd

Convert record format [Info](#)

Data in Apache Parquet or Apache ORC format is typically more efficient to query than JSON. Kinesis Data Firehose can convert your JSON-formatted source records using a schema from a table defined in [AWS Glue](#). For records that aren't in JSON format, create a Lambda function that converts them to JSON in the Transform source records with AWS Lambda section above.

☒ Enable record format conversion

Output format

☒ Apache Parquet

☐ Apache ORC

Schema for source records

Kinesis Data Firehose references table definitions stored in AWS Glue. Choose an AWS Glue table to specify a schema for your source records.

AWS Glue region

US East (N. Virginia) ▼

AWS Glue database

ev-database ▼ [Refresh](#)



Use the same S3 bucket and prefix

You can analyze the data delivered by this delivery stream by querying the AWS Glue table you specified

Lambda

Producer Lambda:

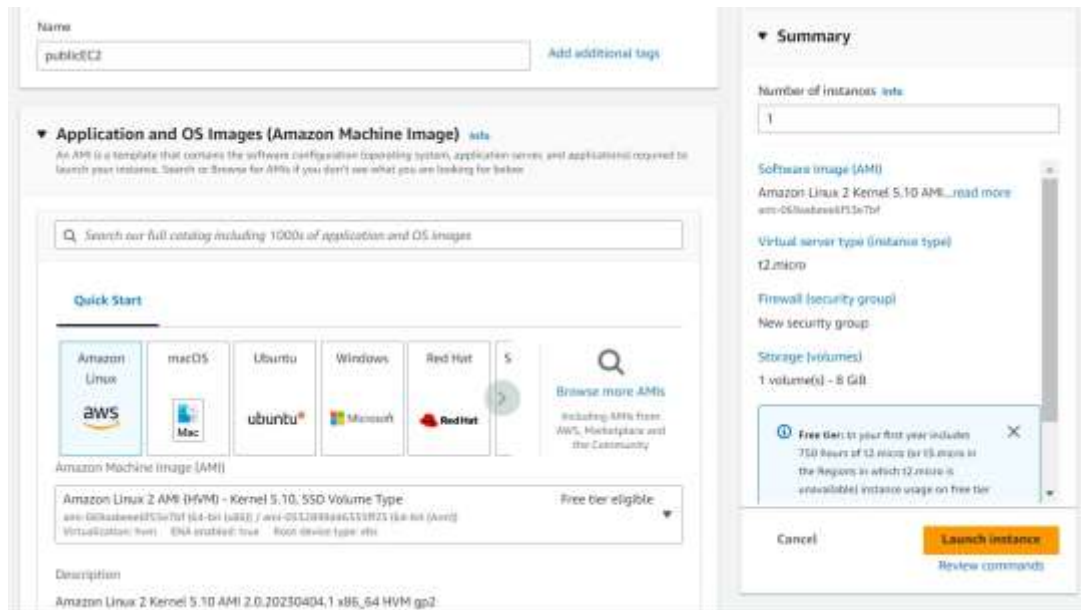
- This lambda function will get the data from API using requests, preprocess the data and publishes the data to Kafka producer.

Consumer Lambda:

- This lambda function will consume the data from Kafka topic and pushes the data into Kinesis Firehose.

EC2 Instances

- We need an EC2 instance to create a Kafka topic for which we need to install Java and Kafka in linux machine. So we are creating two EC2 instance each in public and private subnet of our VPC.



Installing Kafka on EC2 Instance

- Now we have to send SHH into private EC2 instance from public EC2 instance using PuTTY, PuTTYgen and WinSCP and install Java and Kafka using the below commands.

```
ec2-user@ip-10-0-140-144:~  
login as: ec2-user  
Authenticating with public key "publicEC2Key"  
  
  _ | _ | _ )  
  _ | ( _ | /  Amazon Linux 2 AMI  
  _ | \ _ | _ |  
  
https://aws.amazon.com/amazon-linux-2/  
[ec2-user@ip-10-0-7-180 ~]$ ls  
publicEC2Key.pem  
[ec2-user@ip-10-0-7-180 ~]$ chmod 400 publicEC2Key.pem  
[ec2-user@ip-10-0-7-180 ~]$ ssh -i "publicEC2Key.pem" ec2-user@10.0.140.144  
The authenticity of host '10.0.140.144 (10.0.140.144)' can't be established.  
ECDSA key fingerprint is SHA256:y6YuQnt3sFh23VDHaCFouhyAG04+5lMJJaYK/HZSNfZg.  
ECDSA key fingerprint is MD5:94:d7:fe:74:b2:d5:d8:a2:ea:62:df:5f:f8:53:c6:b4.  
Are you sure you want to continue connecting (yes/no)? yes  
Warning: Permanently added '10.0.140.144' (ECDSA) to the list of known hosts.  
  
  _ | _ | _ )  
  _ | ( _ | /  Amazon Linux 2 AMI  
  _ | \ _ | _ |  
  
https://aws.amazon.com/amazon-linux-2/  
[ec2-user@ip-10-0-140-144 ~]$
```

S3

[Amazon S3](#) > [Buckets](#) > [evproject](#) > [2023/](#) > [04/](#) > [16/](#) > [09/](#)

09/

Copy S3 URI

Objects

Properties

Objects (1)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Refresh

Copy S3 URI

Copy URL

Download

Open

Delete


Actions

Create folder

Upload

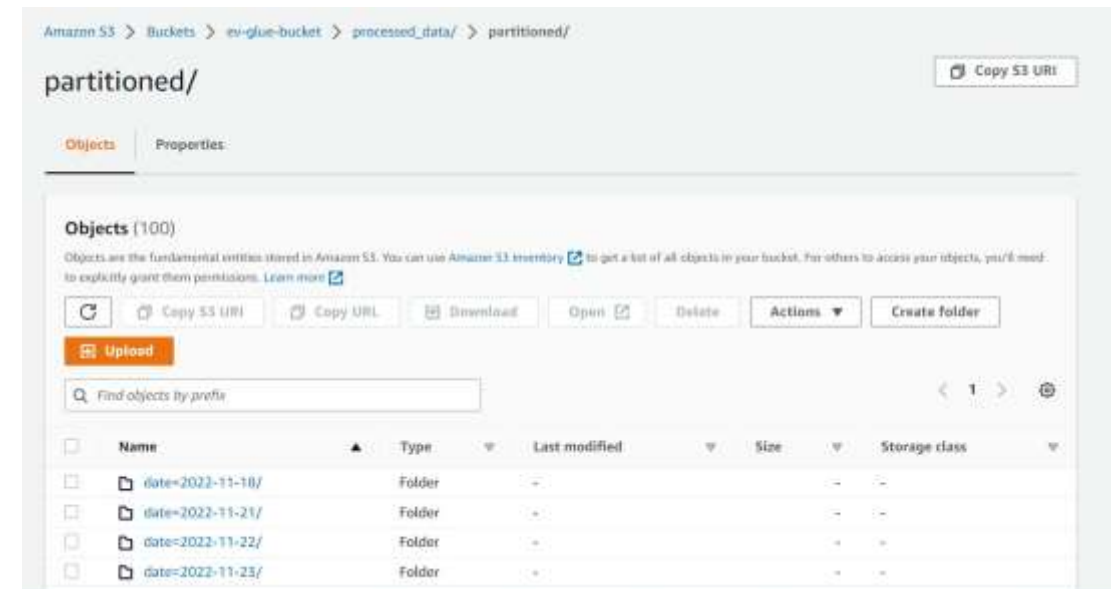
Find objects by prefix

< 1 > ⚙

<input checked="" type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input checked="" type="checkbox"/>	 PUT-S3-xlwJd-1-2023-04-16-09-45-09-ac6e0caa-6199-43b9-8f54-4c461ff069a1.parquet	parquet	April 16, 2023, 15:16:20 (UTC+05:30)	4.3 KB	Standard

Glue

- As the data is successfully placed in s3, now we will create Glue job which will read the files from the s3 directory, process by extracting few columns and place the results in different s3 bucket and also create the Athena table in the mentioned database and stores the data.
- Processed data is successfully stored in given s3 location with date as partitioned column



Athena

- Query - select * from processed_data limit 10;

Query results

Query stats

Completed

Time in queue: 190 ms Run time: 893 ms Data scanned: 6.41 KB

Results (10)

Copy

Download results

Search rows

< 1 > ⌕

#	opening_share	highest_share_value	lowest_share_value	closing_share	total_volume	date
1	23.050000	23.660000	22.970000	23.610000	14259641.000000	2022-11-22
2	18.510000	18.680000	18.250000	18.420000	10720963.000000	2023-04-11
3	18.820000	18.960000	18.540400	18.880000	7772777.000000	2023-03-17
4	23.330000	23.750000	23.310000	23.650000	4503564.000000	2022-11-25
5	16.900000	17.110000	16.570000	16.940000	10945793.000000	2023-03-24
6	17.230000	17.270000	16.760000	17.100000	10311461.000000	2023-03-27
7	22.290000	22.660000	22.050000	22.440000	6456581.000000	2023-02-14
8	20.620000	20.890400	20.220000	20.420000	6975270.000000	2022-12-21
9	20.940000	21.000000	20.340000	20.410000	9214235.000000	2023-02-27
10	20.440000	20.570000	19.925000	20.120000	10743638.000000	2022-12-19

Conclusion

The examination of the stock data is displayed using real-time stock data from various businesses. The AWS Services were utilized for the project's improvement. The enhancement overcame the issues with storage and security. The analyzed data from various companies can be compared from the visualization. The visualization helps the customer to choose the profitable stock from the analyzed data. Ultimately, the consumer might anticipate profitable results from the stock data analysis.

References

- 1 <https://www.digitalocean.com/community/tutorials/how-to-set-up-jupyternotebook-with-python-3-on-ubuntu-20-04-and-connect-via-ssh-tunneling>
- 2 <https://sparkbyexamples.com/kafka/apache-kafka-cluster-setup/>
- 3 <https://www.alphavantage.co/documentation/>



Thank U!