Stock Anomaly Detection System

Implementation Details :-

- Step 1 :- Opened the Google Cloud Shell
- Step 2 :- Run the following commands to authenticate:
    - *gcloud auth login*
    - *gcloud config set eminent-crane-448810-s3*
- Step 3 :- Provisioned three VMs for this assignment. One VM for serving as a kafka broker, one for serving as a producer and one for serving as a consumer. Producer processes the one stock file at a time and writes one trade data at a time to kafka topic. Consumer reads from the same kafka topic, processes the trade data and prints the output to console as well as to Pub/Sub topic (in case of A2 anomaly.
    - create_kafka_vm.sh, create_producer_vm.sh and create_consumer_vm.sh were used for provisioning the VMs.

```
  GNU nano 7.2                                                    create_kafka_vm.sh
# This script provisions a VM for running Kafka server and sets up a firewall rule
# to allow incoming traffic on port 9092 only from the resources serving as consumer and producer:

#!/bin/bash

# Variables
VM_NAME="kafka-server-vm"
PROJECT_ID="eminent-crane-448810-s3"
ZONE="us-central1-a"
MACHINE_TYPE="e2-standard-4"
IMAGE_FAMILY="debian-12"
IMAGE_PROJECT="debian-cloud"

# Step 1: Create the Kafka server VM instance
gcloud compute instances create $VM_NAME \
    --project=$PROJECT_ID \
    --zone=$ZONE \
    --machine-type=$MACHINE_TYPE \
    --image-family=$IMAGE_FAMILY \
    --image-project=$IMAGE_PROJECT \
    --boot-disk-size=200GB \
    --scopes=storage-full,cloud-platform \
    --tags=kafka-server

# Step 2: Open required ports
gcloud compute firewall-rules create kafka-port --allow tcp:9092 --target-tags kafka-server --quiet
```

```
chandrakarsatvik@cloudshell:~ (eminent-crane-448810-s3)$ ./create_kafka_vm.sh
Created [https://www.googleapis.com/compute/v1/projects/eminent-crane-448810-s3/zones/us-central1-a/instances/kafka-server-vm].
WARNING: Some requests generated warnings:
 - Disk size: '200 GB' is larger than image size: '10 GB'. You might need to resize the root repartition manually if the operating system does
oogle.com/compute/docs/disks/add-persistent-disk#resize_pd for details.

NAME: kafka-server-vm
ZONE: us-central1-a
MACHINE_TYPE: e2-standard-4
PREEMPTIBLE:
INTERNAL_IP: 10.128.0.38
EXTERNAL_IP: 34.28.179.159
STATUS: RUNNING
Creating firewall...working..Created [https://www.googleapis.com/compute/v1/projects/eminent-crane-448810-s3/global/firewalls/kafka-port].
Creating firewall...done.
NAME: kafka-port
NETWORK: default
DIRECTION: INGRESS
PRIORITY: 1000
ALLOW: tcp:9092
DENY:
DISABLED: False
```

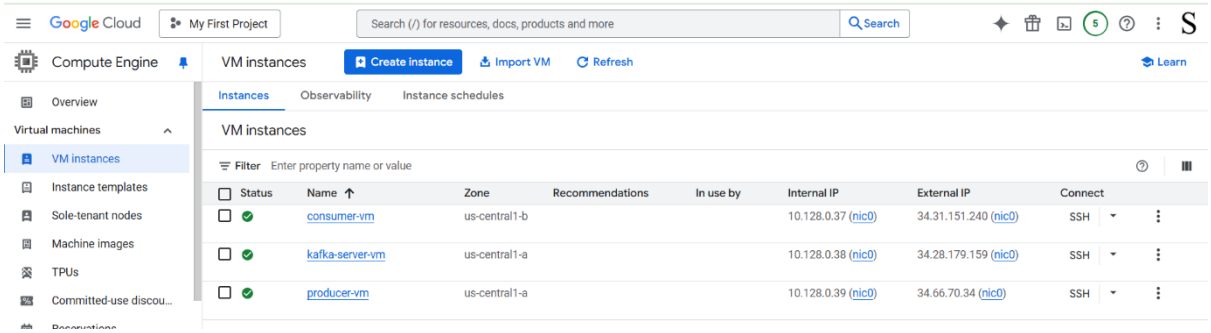
```
  GNU nano 7.2                                                    create_producer_vm.sh
#!/bin/bash

# Variables
PROJECT_ID="eminent-crane-448810-s3"
ZONE="us-central1-a"
VM_NAME="producer-vm"
MACHINE_TYPE="c4-standard-4"
IMAGE_FAMILY="debian-12"
IMAGE_PROJECT="debian-cloud"

# Step 1: Create VM instance
gcloud compute instances create $VM_NAME \
    --project=$PROJECT_ID \
    --zone=$ZONE \
    --machine-type=$MACHINE_TYPE \
    --image-family=$IMAGE_FAMILY \
    --image-project=$IMAGE_PROJECT \
    --boot-disk-size=200GB \
    --scopes=storage-full,cloud-platform
```

```
chandrakarsatvik@cloudshell:~ (eminent-crane-448810-s3)$ ./create_producer_vm.sh
Created [https://www.googleapis.com/compute/v1/projects/eminent-crane-448810-s3/zones/us-central1-a/instances/producer-vm].
WARNING: Some requests generated warnings:
 - Disk size: '200 GB' is larger than image size: '10 GB'. You might need to resize the root repartition manually if the operati
oogle.com/compute/docs/disks/add-persistent-disk#resize_pd for details.

NAME: producer-vm
ZONE: us-central1-a
MACHINE_TYPE: c4-standard-4
PREEMPTIBLE:
INTERNAL_IP: 10.128.0.39
EXTERNAL_IP: 34.66.70.34
STATUS: RUNNING
```

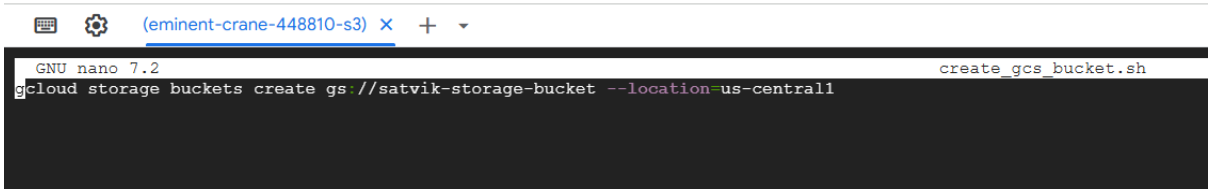
(eminent-crane-448810-s3)

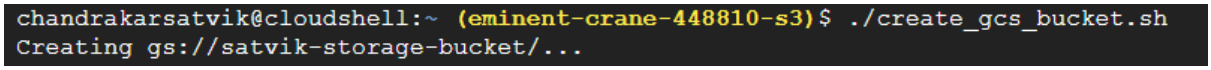
```
  GNU nano 7.2                                                    create_consumer_vm.sh
#!/bin/bash

# Variables
PROJECT_ID="eminent-crane-448810-s3"
ZONE="us-central1-b"
VM_NAME="consumer-vm"
MACHINE_TYPE="c4-standard-4"
IMAGE_FAMILY="debian-12"
IMAGE_PROJECT="debian-cloud"

# Step 1: Create VM instance
gcloud compute instances create $VM_NAME \
    --project=$PROJECT_ID \
    --zone=$ZONE \
    --machine-type=$MACHINE_TYPE \
    --image-family=$IMAGE_FAMILY \
    --image-project=$IMAGE_PROJECT \
    --boot-disk-size=200GB \
    --scopes=storage-full,cloud-platform
```

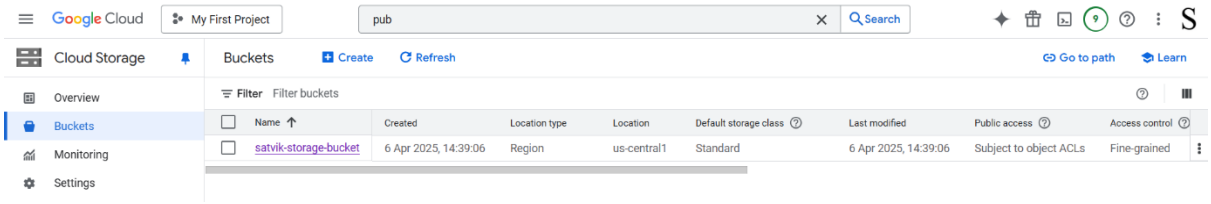
```
chandrakarsatvik@cloudshell:~ (eminent-crane-448810-s3)$ ./create_consumer_vm.sh
Created [https://www.googleapis.com/compute/v1/projects/eminent-crane-448810-s3/zones/us-central1-b/instances/consumer-vm].
WARNING: Some requests generated warnings:
 - Disk size: '200 GB' is larger than image size: '10 GB'. You might need to resize the root repartition manually if the operat
oogle.com/compute/docs/disks/add-persistent-disk#resize_pd for details.

NAME: consumer-vm
ZONE: us-central1-b
MACHINE_TYPE: c4-standard-4
PREEMPTIBLE:
INTERNAL_IP: 10.128.0.37
EXTERNAL_IP: 34.31.151.240
STATUS: RUNNING
```

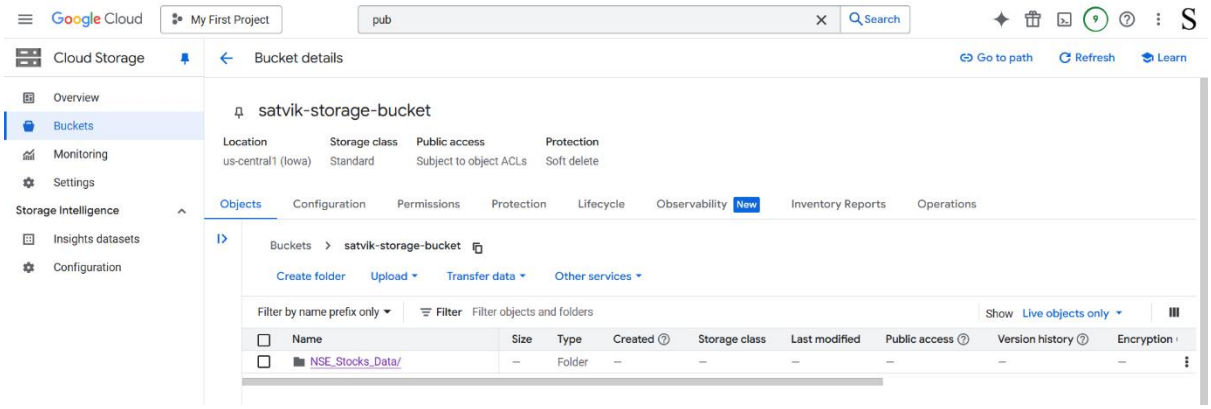


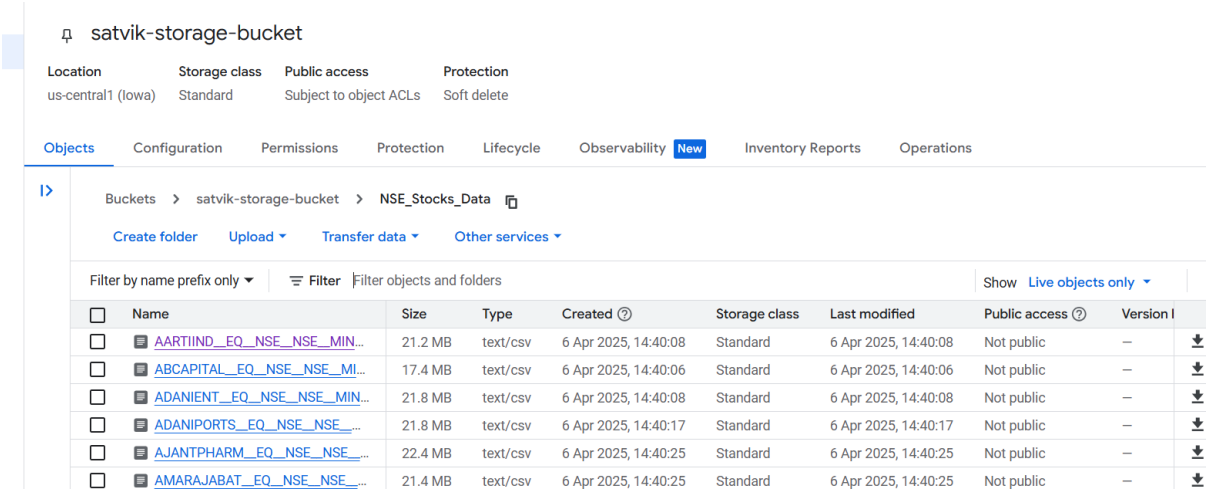
- Step 4 :- Create a GCS bucket and upload NSE_Stocks_Data folder to it using "Upload Folder" on the GCS Bucket console.



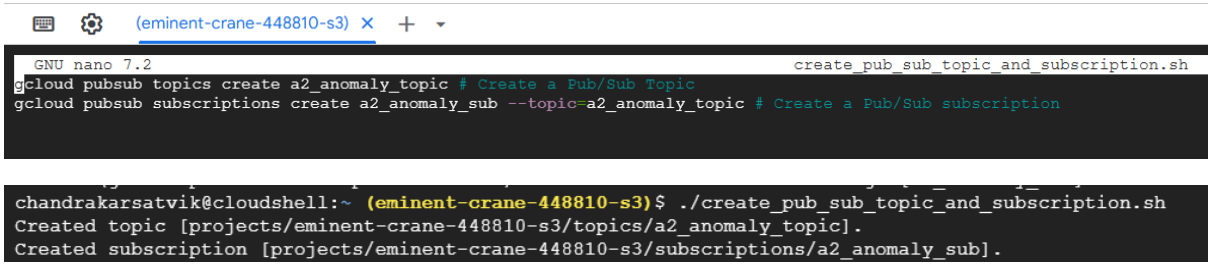


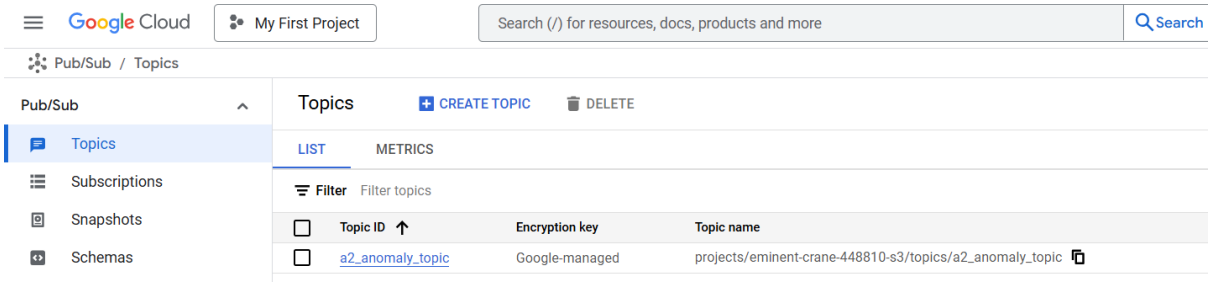


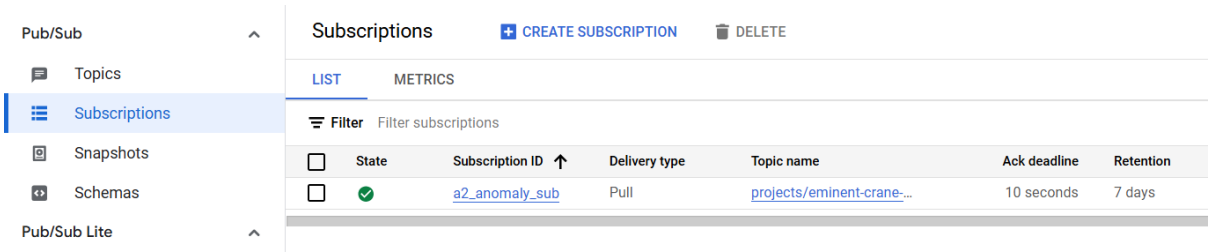




- Step 5 :- Created a Pub/Sub Topic and Pub/Sub Subscription.







- Step 6 :- SSH into the kafka-vm, installed the dependencies, started the zookeeper & kafka server and created the topic.

SSH-in-browser UPLOAD FILE DOWNLOAD FILE

```
  GNU nano 7.2                                    install_dependencies.sh
# To be executed inside VM
# Install Java
sudo apt update
sudo apt install default-jdk -y
java -version
# Download & Extract Kafka
wget https://downloads.apache.org/kafka/3.7.2/kafka_2.13-3.7.2.tgz
tar -xvzf kafka_2.13-3.7.2.tgz
mv kafka_2.13-3.7.2 kafka

sudo apt update && sudo apt install netcat -y
```

SSH-in-browser UPLOAD FILE DOWNLOAD FILE

```
  GNU nano 7.2                                    start_zookeeper.sh
# To be executed inside kafka_vm
# Start Zookeeper
cd kafka
bin/zookeeper-server-start.sh config/zookeeper.properties
```

SSH-in-browser UPLOAD FILE

```
Linux kafka-server-vm 6.1.0-31-cloud-amd64 #1 SMP PREEMPT_DYNAMIC Debian 6.1.128-1 (2025-02-07) x86_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Sun Apr  6 11:09:33 2025 from 35.235.244.34
chandrakarsatvik@kafka-server-vm:~$ ls
create_kafka_topic.sh  install_dependencies.sh  kafka  kafka_2.13-3.7.2.tgz  start_kafka.sh  start_zookeeper.sh
chandrakarsatvik@kafka-server-vm:~$ ./start_zookeeper.sh
[2025-04-06 11:22:41,883] INFO Reading configuration from: config/zookeeper.properties (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2025-04-06 11:22:41,885] WARN config/zookeeper.properties is relative. Prepend ./ to indicate that you're sure! (org.apache.zookeeper.server.quorum.QuorumPeerCo
[2025-04-06 11:22:41,889] INFO clientPortAddress is 0.0.0.0:2181 (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2025-04-06 11:22:41,889] INFO secureClientPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2025-04-06 11:22:41,889] INFO observerMasterPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2025-04-06 11:22:41,889] INFO metricsProvider.className is org.apache.zookeeper.metrics.impl.DefaultMetricsProvider (org.apache.zookeeper.server.quorum.QuorumPe
[2025-04-06 11:22:41,892] INFO autopurge.snapRetainCount set to 3 (org.apache.zookeeper.server.DatadirCleanupManager)
[2025-04-06 11:22:41,892] INFO autopurge.purgeInterval set to 0 (org.apache.zookeeper.server.DatadirCleanupManager)
[2025-04-06 11:22:41,892] INFO Purge task is not scheduled. (org.apache.zookeeper.server.DatadirCleanupManager)
[2025-04-06 11:22:41,892] WARN Either no config or no quorum defined in config, running in standalone mode (org.apache.zookeeper.server.quorum.QuorumPeerMain)
[2025-04-06 11:22:41,894] INFO Log4j 1.2 jmx support not found; jmx disabled. (org.apache.zookeeper.jmx.ManagedUtil)
[2025-04-06 11:22:41,895] INFO Reading configuration from: config/zookeeper.properties (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2025-04-06 11:22:41,895] WARN config/zookeeper.properties is relative. Prepend ./ to indicate that you're sure! (org.apache.zookeeper.server.quorum.QuorumPeerCo
[2025-04-06 11:22:41,896] INFO clientPortAddress is 0.0.0.0:2181 (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2025-04-06 11:22:41,896] INFO secureClientPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2025-04-06 11:22:41,896] INFO observerMasterPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2025-04-06 11:22:41,896] INFO metricsProvider.className is org.apache.zookeeper.metrics.impl.DefaultMetricsProvider (org.apache.zookeeper.server.quorum.QuorumPe
[2025-04-06 11:22:41,896] INFO Starting server (org.apache.zookeeper.server.ZooKeeperServerMain)
[2025-04-06 11:22:41,914] INFO ServerMetrics initialized with provider org.apache.zookeeper.metrics.impl.DefaultMetricsProvider@17c1bced (org.apache.zookeeper.se
[2025-04-06 11:22:41,919] INFO ACL digest algorithm is: SHA1 (org.apache.zookeeper.server.auth.DigestAuthenticationProvider)
[2025-04-06 11:22:41,919] INFO zookeeper.DigestAuthenticationProvider.enabled = true (org.apache.zookeeper.server.auth.DigestAuthenticationProvider)
[2025-04-06 11:22:41,923] INFO zookeeper.snapshot.trust.empty : false (org.apache.zookeeper.server.persistence.FileTxnSnapLog)
[2025-04-06 11:22:41,935] INFO  (org.apache.zookeeper.server.ZooKeeperServer)
[2025-04-06 11:22:41,935] INFO   ____          _                                          (org.apache.zookeeper.server.ZooKeeperServer)
[2025-04-06 11:22:41,935] INFO  |___  /                 | |                                 (org.apache.zookeeper.server.ZooKeeperServer)
[2025-04-06 11:22:41,935] INFO     / /   ___    ___    | |  __   ___   ___   _ __    ___   _ __  (org.apache.zookeeper.server.ZooKeeperServer)
[2025-04-06 11:22:41,935] INFO    / /   / _ \   / _ \   | | / /  / _ \  / _ \ | '_ \  / _ \ | '__| (org.apache.zookeeper.server.ZooKeeperServer)
[2025-04-06 11:22:41,935] INFO   / /__  | (_) | | (_) |  |   <  |  __/ |  __/ | |_) | |  __/ | |     (org.apache.zookeeper.server.ZooKeeperServer)
[2025-04-06 11:22:41,935] INFO  /_____|  \___/   \___/  |_|\_\  \___|  \___| | .__/   \___| |_| (org.apache.zookeeper.server.ZooKeeperServer)
[2025-04-06 11:22:41,935] INFO                                               | |                  (org.apache.zookeeper.server.ZooKeeperServer)
[2025-04-06 11:22:41,936] INFO                                               |_|                  (org.apache.zookeeper.server.ZooKeeperServer)
[2025-04-06 11:22:41,936] INFO  (org.apache.zookeeper.server.ZooKeeperServer)
[2025-04-06 11:22:41,938] INFO Server environment:zookeeper.version=3.8.4-9316c2a7a97e1666d8f4593f34dd6fc36ecc436c, built on 2024-02-12 22:16 UTC (org.apache.zoo
[2025-04-06 11:22:41,938] INFO Server environment:host.name=kafka-server-vm.us-central1-a.c.eminent-crane-448810-s3.internal (org.apache.zookeeper.server.ZooKeep
```

Zookeeper started in the first terminal of kafka-vm

SSH-in-browser

UPLOAD FIL

```
GNU nano 7.2                                                    start_kafka.sh
# To be executed inside kafka_vm
# Start Kafka

# Navigate to the Kafka directory
cd kafka

KAFKA_EXTERNAL_IP=$(curl -s ifconfig.me)

# Update (or append) the listeners configuration. This ensures that Kafka binds to all interfaces and advertises the correct external address.
if grep -q "^listeners=" config/server.properties; then
    sed -i 's|^listeners=.*|listeners=PLAINTEXT://0.0.0.0:9092|' config/server.properties
else
    echo "listeners=PLAINTEXT://0.0.0.0:9092" >> config/server.properties
fi

# Update (or append) the advertised.listeners configuration. This ensures that Kafka binds to all interfaces and advertises the correct external address.
if grep -q "^advertised.listeners=" config/server.properties; then
    sed -i "s|^advertised.listeners=.*|advertised.listeners=PLAINTEXT://$KAFKA_EXTERNAL_IP:9092|" config/server.properties
else
    echo "advertised.listeners=PLAINTEXT://$KAFKA_EXTERNAL_IP:9092" >> config/server.properties
fi

bin/kafka-server-start.sh config/server.properties
```

SSH-in-browser

UPLOAD FILE DOWNLOAD FILE

```
Linux kafka-server-vm 6.1.0-31-cloud-amd64 #1 SMP PREEMPT_DYNAMIC Debian 6.1.128-1 (2025-02-07) x86_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Sun Apr  6 11:22:13 2025 from 35.235.244.32
chandrakarsatvik@kafka-server-vm:~$ ls
create_kafka_topic.sh  install_dependencies.sh  kafka  kafka_2.13-3.7.2.tgz  start_kafka.sh  start_zookeeper.sh
chandrakarsatvik@kafka-server-vm:~$ ./start_kafka.sh
[2025-04-06 11:23:15,066] INFO Registered kafka:type=kafka.Log4jController MBean (kafka.utils.Log4jControllerRegistration$)
[2025-04-06 11:23:15,385] INFO Setting -D jdk.tls.rejectClientInitiatedRenegotiation=true to disable client-initiated TLS renegotiation (org.apache.zookeeper.common.X509Util)
[2025-04-06 11:23:15,495] INFO Registered signal handlers for TERM, INT, HUP (org.apache.kafka.common.utils.LoggingSignalHandler)
[2025-04-06 11:23:15,497] INFO starting (kafka.server.KafkaServer)
[2025-04-06 11:23:15,498] INFO Connecting to zookeeper on localhost:2181 (kafka.server.KafkaServer)
[2025-04-06 11:23:15,516] INFO [ZooKeeperClient Kafka server] Initializing a new session to localhost:2181. (kafka.zookeeper.ZooKeeperClient)
[2025-04-06 11:23:15,521] INFO Client environment:zookeeper.version=3.8.4-9316c2a7a97e1666d8f4593f34dd6fc36ecc436c, built on 2024-02-12 22:16 UTC (org.apache.zookeeper.ZooKeeper)
[2025-04-06 11:23:15,521] INFO Client environment:host.name=kafka-server-vm.us-central1-a.c.eminent-crane-448810-s3.internal (org.apache.zookeeper.ZooKeeper)
[2025-04-06 11:23:15,521] INFO Client environment:java.version=17.0.14 (org.apache.zookeeper.ZooKeeper)
[2025-04-06 11:23:15,521] INFO Client environment:java.vendor=Debian (org.apache.zookeeper.ZooKeeper)
[2025-04-06 11:23:15,521] INFO Client environment:java.home=/usr/lib/jvm/java-17-openjdk-amd64 (org.apache.zookeeper.ZooKeeper)
```

Kafka server started in the second terminal of kafka-vm

SSH-in-browser

UPLOAD FILE

```
GNU nano 7.2                                                    create_kafka_topic.sh
# Create a Kafka Topic

# --partitions 1 --replication-factor 1

#!/bin/bash

# Variables
KAFKA_BROKER="$(curl -s ifconfig.me):9092" # Kafka broker's IP
TOPIC_NAME=$1                              # First argument: Topic name
PARTITIONS=${2:-1}                         # Second argument (default = 1 partition)
REPLICATION_FACTOR=${3:-1}                 # Third argument (default = 1 replica)

# Validate inputs
if [ -z "$TOPIC_NAME" ]; then
    echo "Usage: $0 <topic-name> [partitions] [replication-factor]"
    exit 1
fi

# Check if Kafka is reachable
echo "Checking Kafka broker at $KAFKA_BROKER..."
nc -zv $(echo $KAFKA_BROKER | cut -d':' -f1) 9092
if [ $? -ne 0 ]; then
    echo "Error: Unable to reach Kafka broker at $KAFKA_BROKER. Check broker status and network settings."
    exit 1
fi

# Create Kafka topic
echo "Creating Kafka topic: $TOPIC_NAME with $PARTITIONS partitions and $REPLICATION_FACTOR replication factor..."
cd kafka
bin/kafka-topics.sh --create --topic "$TOPIC_NAME" --bootstrap-server "$KAFKA_BROKER" --partitions "$PARTITIONS" --replication-factor "$REPLICATION_FACTOR"

# Verify topic creation
echo "Verifying topic creation..."
bin/kafka-topics.sh --list --bootstrap-server "$KAFKA_BROKER" | grep "$TOPIC_NAME"
if [ $? -eq 0 ]; then
    echo "Kafka topic '$TOPIC_NAME' created successfully!"
else
    echo "Error: Failed to create Kafka topic '$TOPIC_NAME'. Check Kafka logs."
    exit 1
fi
```

Created the kafka topic named stock-input-data. create_kafka_topic.sh was executed on the third terminal of the kafka-vm.

```
chandrakarsatvik@kafka-server-vm:~$ ./create_kafka_topic.sh stock-input-data 1 1
Checking Kafka broker at 34.28.179.159:9092...
Connection to 34.28.179.159 9092 port [tcp/*] succeeded!
Creating Kafka topic: stock-input-data with 1 partitions and 1 replication factor...
Created topic stock-input-data.
Verifying topic creation...
stock-input-data
Kafka topic 'stock-input-data' created successfully!
chandrakarsatvik@kafka-server-vm:~$
```

- Step 6 :- SSH into the producer-vm, installed the dependencies, downloaded the NSE_Stocks_Data from the GCS storage bucket, created the producer.py file to processes the one stock file at a time and writes one trade data at a time to kafka topic.

SSH-in-browser

```
GNU nano 7.2                                                    install_dependencies.sh
# To be executed inside producer_vm
# Install Java
sudo apt update
sudo apt install default-jdk -y
java -version
# Download & Extract Spark
wget https://downloads.apache.org/spark/spark-3.5.5/spark-3.5.5-bin-hadoop3.tgz
tar -xvzf spark-3.5.5-bin-hadoop3.tgz
mv spark-3.5.5-bin-hadoop3 spark
sudo apt update && sudo apt install -y google-cloud-sdk python3 python3-pip scala
pip3 install google-cloud-storage kafka-python pyspark
```

SSH-in-browser

```
GNU nano 7.2                                                    download_data_gcs_bucket.sh
gsutil cp -r gs://satvik-storage-bucket/NSE_Stocks_Data /home/chandrakarsatvik/
```

```
(venv) chandrakarsatvik@producer-vm:~$ ./download_data_gcs_bucket.sh
Copying gs://satvik-storage-bucket/NSE_Stocks_Data/AARTIIND__EQ__NSE__NSE__MINUTE.csv...
Copying gs://satvik-storage-bucket/NSE_Stocks_Data/ABCAPITAL__EQ__NSE__NSE__MINUTE.csv...
Copying gs://satvik-storage-bucket/NSE_Stocks_Data/ADANIENT__EQ__NSE__NSE__MINUTE.csv...
Copying gs://satvik-storage-bucket/NSE_Stocks_Data/ADANIPORTS__EQ__NSE__NSE__MINUTE.csv...
```

SSH-in-browser

UPLOAD FILE DOWNLOAD FILE

```
  GNU nano 7.2                                                    producer.py
from pyspark.sql import SparkSession
from pyspark.sql.functions import *
import time
import os
from kafka import KafkaProducer
import json

# 🔧CONFIGURATION
KAFKA_TOPIC = "stock-input-data"
KAFKA_BOOTSTRAP_SERVERS = "34.28.179.159:9092"
CSV_ROOT_FOLDER = "./NSE_Stocks_Data"  # Root folder with nested CSV files

# 🚀Create Spark Session
spark = SparkSession.builder \
    .appName("RecursiveCSVToKafkaStreamer") \
    .getOrCreate()

producer = KafkaProducer(bootstrap_servers=KAFKA_BOOTSTRAP_SERVERS, value_serializer=lambda v: json.dumps(v).encode('utf-8'))


# 🔁Recursively find all .csv files
for root, dirs, files in os.walk(CSV_ROOT_FOLDER):
    for file in files:
        if file.endswith(".csv"):
            full_path = os.path.join(root, file)
            print(f"\n📂Processing: {full_path}")

            # Derive stock_id from filename (before first '')
            stock_id = file.split("__")[0]

            # Load CSV
            df = spark.read.option("header", True).csv(full_path)
            print(f"🔍Raw rows in {file}: {df.count()}")

            # 🧹Preprocess: add stock_id, cast types
            df = df.select(
                lit(stock_id).alias("stock_id"),
                to_timestamp("timestamp").alias("timestamp"),
                col("close").cast("double").alias("close_price"),
                col("volume").cast("long")
            )

            # Filter bad rows
            df = df.filter("stock_id IS NOT NULL AND timestamp IS NOT NULL AND close_price IS NOT NULL AND volume IS NOT NULL")

            # Sort the data
            df = df.orderBy(col("timestamp"))
```

```
(venv) chandrakarsatvik@producer-vm:~$ python3 producer.py
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
25/04/06 12:17:42 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

📂Processing: ./NSE_Stocks_Data/BAJAJFINSV__EQ__NSE__NSE__MINUTE.csv
🔍Raw rows in BAJAJFINSV__EQ__NSE__NSE__MINUTE.csv: 370546
✅Cleaned rows in BAJAJFINSV__EQ__NSE__NSE__MINUTE.csv: 370408
✅Sent: Row(stock_id='BAJAJFINSV', timestamp=datetime.datetime(2017, 1, 2, 3, 45), close_price=2901.4, volume=347)
✅Sent: Row(stock_id='BAJAJFINSV', timestamp=datetime.datetime(2017, 1, 2, 3, 46), close_price=2899.25, volume=1419)
✅Sent: Row(stock_id='BAJAJFINSV', timestamp=datetime.datetime(2017, 1, 2, 3, 47), close_price=2868.3, volume=642)
✅Sent: Row(stock_id='BAJAJFINSV', timestamp=datetime.datetime(2017, 1, 2, 3, 48), close_price=2885.0, volume=712)
✅Sent: Row(stock_id='BAJAJFINSV', timestamp=datetime.datetime(2017, 1, 2, 3, 49), close_price=2890.0, volume=637)
✅Sent: Row(stock_id='BAJAJFINSV', timestamp=datetime.datetime(2017, 1, 2, 3, 50), close_price=2892.7, volume=591)
✅Sent: Row(stock_id='BAJAJFINSV', timestamp=datetime.datetime(2017, 1, 2, 3, 51), close_price=2880.2, volume=404)
✅Sent: Row(stock_id='BAJAJFINSV', timestamp=datetime.datetime(2017, 1, 2, 3, 52), close_price=2876.5, volume=201)
✅Sent: Row(stock_id='BAJAJFINSV', timestamp=datetime.datetime(2017, 1, 2, 3, 53), close_price=2884.65, volume=402)
✅Sent: Row(stock_id='BAJAJFINSV', timestamp=datetime.datetime(2017, 1, 2, 3, 54), close_price=2896.65, volume=208)
✅Sent: Row(stock_id='BAJAJFINSV', timestamp=datetime.datetime(2017, 1, 2, 3, 55), close_price=2900.0, volume=1200)
✅Sent: Row(stock_id='BAJAJFINSV', timestamp=datetime.datetime(2017, 1, 2, 3, 56), close_price=2903.35, volume=535)
✅Sent: Row(stock_id='BAJAJFINSV', timestamp=datetime.datetime(2017, 1, 2, 3, 57), close_price=2899.95, volume=908)
✅Sent: Row(stock_id='BAJAJFINSV', timestamp=datetime.datetime(2017, 1, 2, 3, 58), close_price=2899.95, volume=183)
✅Sent: Row(stock_id='BAJAJFINSV', timestamp=datetime.datetime(2017, 1, 2, 3, 59), close_price=2896.0, volume=197)
✅Sent: Row(stock_id='BAJAJFINSV', timestamp=datetime.datetime(2017, 1, 2, 4, 0), close_price=2899.95, volume=350)
✅Sent: Row(stock_id='BAJAJFINSV', timestamp=datetime.datetime(2017, 1, 2, 4, 1), close_price=2895.1, volume=284)
```

- Step 7 :- SSH into the consumer-vm, installed the dependencies, created the consumer.py file to read the data from the kafka topic, processes the trade data to detect anomaly and prints the output to console as well as to Pub/Sub topic (in case of A2 anomaly)

SSH-in-browser

```
GNU nano 7.2                                                    install_dependencies.sh *
# To be executed inside VM
# Install Java
sudo apt update
sudo apt install default-jdk -y
java -version
# Download & Extract Kafka
wget https://downloads.apache.org/kafka/3.7.2/kafka_2.13-3.7.2.tgz
tar -xvzf kafka_2.13-3.7.2.tgz
mv kafka_2.13-3.7.2 kafka
# Download & Extract Spark
wget https://downloads.apache.org/spark/spark-3.5.5/spark-3.5.5-bin-hadoop3.tgz
tar -xvzf spark-3.5.5-bin-hadoop3.tgz
mv spark-3.5.5-bin-hadoop3 spark
echo 'Kafka & Spark setup completed!'
# Packages
sudo apt update && sudo apt install -y google-cloud-sdk python3 python3-pip scala
pip3 install google-cloud-storage kafka-python pyspark pandas
```

SSH-in-browser

UPLOAD FILE

```
GNU nano 7.2                                                    consumer.py
from pyspark.sql import SparkSession
from pyspark.sql.functions import *
from pyspark.sql.types import StructType, StringType, DoubleType, LongType, TimestampType
from google.cloud import pubsub_v1

# Initialize Spark session with the necessary configuration
spark = SparkSession.builder \
    .appName("StockAnomalyDetection") \
    .config("spark.sql.streaming.statefulOperator.allowMultiple", "true") \
    .getOrCreate()

KAFKA_TOPIC = "stock-input-data"
KAFKA_BOOTSTRAP_SERVERS = "34.28.179.159:9092"
PROJECT_ID = "eminent-crane-448810-s3"
TOPIC_NAME = "a2_anomaly_topic"
publisher = pubsub_v1.PublisherClient()
topic_path = publisher.topic_path(PROJECT_ID, TOPIC_NAME)

# Define schema
schema = StructType() \
    .add("stock_id", StringType()) \
    .add("timestamp", TimestampType()) \
    .add("close_price", DoubleType()) \
    .add("volume", LongType())

# Kafka stream
df_raw = spark.readStream \
    .format("kafka") \
    .option("kafka.bootstrap.servers", KAFKA_BOOTSTRAP_SERVERS) \
    .option("subscribe", KAFKA_TOPIC) \
    .option("startingOffsets", "latest") \
    .load()

# Parse JSON and timestamps
df = df_raw.selectExpr("CAST(value AS STRING)") \
    .select(from_json(col("value"), schema).alias("data")) \
    .select("data.*") \
    .withColumn("timestamp", to_timestamp("timestamp")) \
    .withWatermark("timestamp", "10 minutes")  # Single watermark that works for both analyses
```

SSH-in-browser

```
GNU nano 7.2                                                    run_consumer.sh
export PYSPARK_SUBMIT_ARGS="--packages org.apache.spark:spark-sql-kafka-0-10_2.12:3.5.5 pyspark-shell"
python3 consumer.py
```

SSH-in-browser

UPLOAD FILE DOWNLOAD FILE

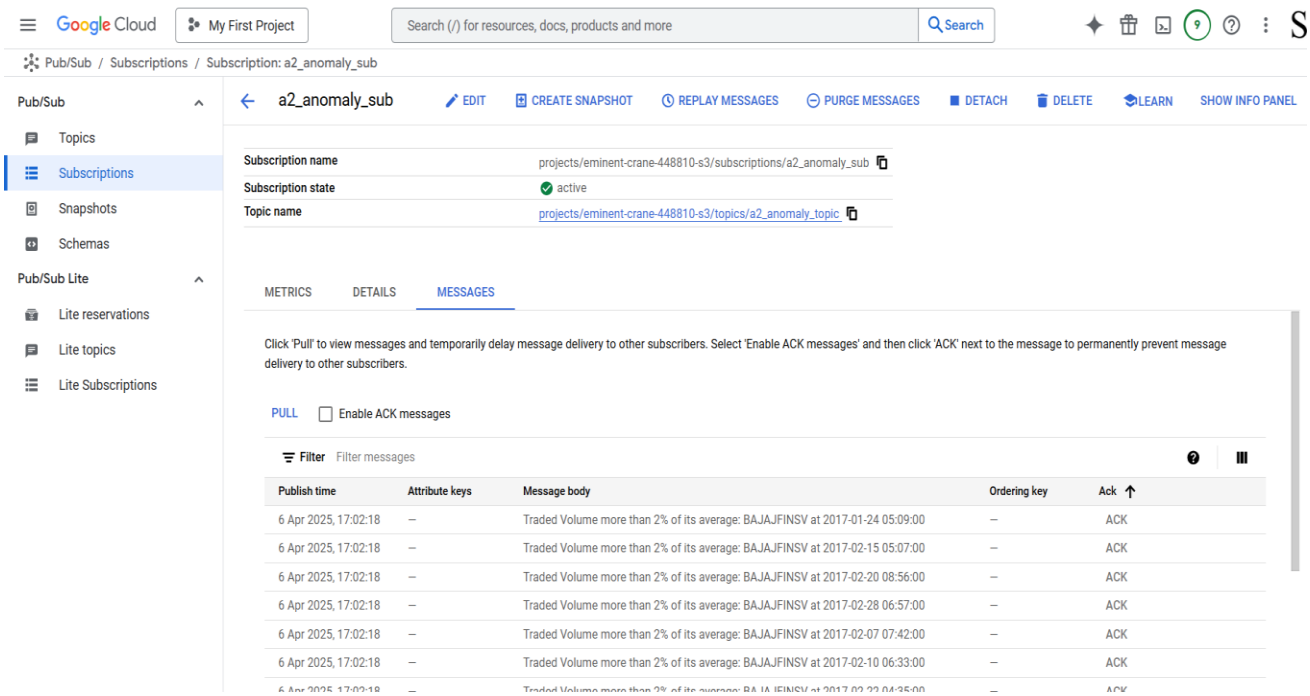
```
consumer.py  install dependencies.sh  kafka  kafka 2.13-3.7.2.tgz  run consumer.sh  spark  spark-3.5.5-bin-hadoop3.tgz  venv
(venv) chandrakarsatvik@consumer-vm:~$ nano install_dependencies.sh
(venv) chandrakarsatvik@consumer-vm:~$ nano consumer.py
(venv) chandrakarsatvik@consumer-vm:~$ nano run_consumer.sh
(venv) chandrakarsatvik@consumer-vm:~$ ./run_consumer.sh
:: loading settings :: url = jar:file:/home/chandrakarsatvik/venv/lib/python3.11/site-packages/pyspark/jars/ivy-2.5.1.jar!/org/apache/ivy/core/settings/ivysettings.xml
Ivy Default Cache set to: /home/chandrakarsatvik/.ivy2/cache
The jars for the packages stored in: /home/chandrakarsatvik/.ivy2/jars
org.apache.spark#spark-sql-kafka-0-10_2.12 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-ecd051b2-3c28-486b-af6a-8d1c4de3c59c;1.0
        confs: [default]
        found org.apache.spark#spark-sql-kafka-0-10_2.12;3.5.5 in central
        found org.apache.spark#spark-token-provider-kafka-0-10_2.12;3.5.5 in central
        found org.apache.kafka#kafka-clients;3.4.1 in central
        found org.lz4#lz4-java;1.8.0 in central
        found org.xerial.snappy#snappy-java;1.1.10.5 in central
        found org.slf4j#slf4j-api;2.0.7 in central
        found org.apache.hadoop#hadoop-client-runtime;3.3.4 in central
        found org.apache.hadoop#hadoop-client-api;3.3.4 in central
        found commons-logging#commons-logging;1.1.3 in central
        found com.google.code.findbugs#jsr305;3.0.0 in central
        found org.apache.commons#commons-pool2;2.11.1 in central
:: resolution report :: resolve 321ms :: artifacts dl 12ms
        :: modules in use:
        com.google.code.findbugs#jsr305;3.0.0 from central in [default]
        commons-logging#commons-logging;1.1.3 from central in [default]
        org.apache.commons#commons-pool2;2.11.1 from central in [default]
        org.apache.hadoop#hadoop-client-api;3.3.4 from central in [default]
        org.apache.hadoop#hadoop-client-runtime;3.3.4 from central in [default]
        org.apache.kafka#kafka-clients;3.4.1 from central in [default]
        org.apache.spark#spark-sql-kafka-0-10_2.12;3.5.5 from central in [default]
        org.apache.spark#spark-token-provider-kafka-0-10_2.12;3.5.5 from central in [default]
        org.lz4#lz4-java;1.8.0 from central in [default]
        org.slf4j#slf4j-api;2.0.7 from central in [default]
        org.xerial.snappy#snappy-java;1.1.10.5 from central in [default]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   11  |   0   |   0   |   0   ||   11  |   0   |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent-ecd051b2-3c28-486b-af6a-8d1c4de3c59c
        confs: [default]
        0 artifacts copied, 11 already retrieved (0kB/6ms)
```

SSH-in-browser

UPLOAD FILE

```
25/04/06 11:31:02 WARN AdminClientConfig: These configurations '[key.deserializer, value.deserializer, enable.auto.commit, max.poll.records, auto.offset.reset]
et.
-------------------------------------------
Batch: 0
-------------------------------------------
+--------+---------+-----------+------+------------+
|stock_id|timestamp|close_price|volume|anomaly_type|
+--------+---------+-----------+------+------------+
+--------+---------+-----------+------+------------+

-------------------------------------------
Batch: 1
-------------------------------------------
+--------+---------+-----------+------+------------+
|stock_id|timestamp|close_price|volume|anomaly_type|
+--------+---------+-----------+------+------------+
+--------+---------+-----------+------+------------+

-------------------------------------------
Batch: 2
-------------------------------------------
+----------+-------------------+-----------+------+------------+ (35 + 4) / 400]
|stock_id  |timestamp          |close_price|volume|anomaly_type|
+----------+-------------------+-----------+------+------------+
|BAJAJFINSV|2017-01-02 03:46:00|2899.25    |1419  |A2          |
+----------+-------------------+-----------+------+------------+

-------------------------------------------
Batch: 3
-------------------------------------------
+----------+-------------------+-----------+------+------------+
|stock_id  |timestamp          |close_price|volume|anomaly_type|
+----------+-------------------+-----------+------+------------+
|BAJAJFINSV|2017-01-04 09:40:00|2967.0     |772   |A2          |
|BAJAJFINSV|2017-02-08 09:31:00|3648.0     |1010  |A2          |
|BAJAJFINSV|2017-01-03 06:23:00|2966.25    |477   |A2          |
|BAJAJFINSV|2017-01-02 09:15:00|2886.0     |102   |A2          |
|BAJAJFINSV|2017-01-06 09:18:00|2993.9     |1005  |A2          |
|BAJAJFINSV|2017-01-06 07:53:00|2997.0     |16    |A2          |
|BAJAJFINSV|2017-02-01 09:59:00|3328.0     |2418  |A2          |
|BAJAJFINSV|2017-02-03 06:46:00|3422.0     |1001  |A2          |
|BAJAJFINSV|2017-01-03 08:37:00|2993.45    |5765  |A2          |
|BAJAJFINSV|2017-01-24 07:44:00|3122.1     |44    |A2          |
|BAJAJFINSV|2017-01-02 06:51:00|2882.0     |52    |A2          |
```

```
-------------------------------------------
Batch: 3
-------------------------------------------
+----------+-------------------+------------+------+------------+
|stock_id  |timestamp          |close_price|volume|anomaly_type|
+----------+-------------------+------------+------+------------+
|BAJAJFINSV|2017-01-04 09:40:00|2967.0      |772   |A2          |
|BAJAJFINSV|2017-02-08 09:31:00|3648.0      |1010  |A2          |
|BAJAJFINSV|2017-01-03 06:23:00|2966.25     |477   |A2          |
|BAJAJFINSV|2017-01-02 09:15:00|2886.0      |102   |A2          |
|BAJAJFINSV|2017-01-06 09:18:00|2993.9      |1005  |A2          |
|BAJAJFINSV|2017-01-06 07:53:00|2997.0      |16    |A2          |
|BAJAJFINSV|2017-02-01 09:59:00|3328.0      |2418  |A2          |
|BAJAJFINSV|2017-02-03 06:46:00|3422.0      |1001  |A2          |
|BAJAJFINSV|2017-01-03 08:37:00|2993.45     |5765  |A2          |
|BAJAJFINSV|2017-01-24 07:44:00|3122.1      |44    |A2          |
|BAJAJFINSV|2017-01-02 06:51:00|2882.0      |52    |A2          |
|BAJAJFINSV|2017-02-01 09:53:00|3330.6      |1337  |A2          |
|BAJAJFINSV|2017-01-30 09:59:00|3230.0      |3727  |A2          |
|BAJAJFINSV|2017-02-06 04:45:00|3490.0      |252   |A2          |
|BAJAJFINSV|2017-01-25 09:48:00|3169.0      |1498  |A2          |
|BAJAJFINSV|2017-02-07 08:28:00|3690.0      |338   |A2          |
|BAJAJFINSV|2017-01-16 06:50:00|3018.5      |19    |A2          |
|BAJAJFINSV|2017-01-03 08:35:00|2967.05     |75    |A2          |
|BAJAJFINSV|2017-01-23 05:04:00|3047.3      |415   |A2          |
|BAJAJFINSV|2017-02-06 08:08:00|3576.95     |867   |A2          |
+----------+-------------------+------------+------+------------+
only showing top 20 rows

-------------------------------------------
Batch: 4
-------------------------------------------
+----------+-------------------+------------+------+------------+
|stock_id  |timestamp          |close_price|volume|anomaly_type|
+----------+-------------------+------------+------+------------+
|BAJAJFINSV|2017-02-15 07:27:00|3630.0      |2488  |A2          |
|BAJAJFINSV|2017-03-09 09:12:00|3811.3      |81    |A2          |
|BAJAJFINSV|2017-03-24 04:43:00|4184.0      |128   |A2          |
|BAJAJFINSV|2017-03-27 09:33:00|4080.05     |484   |A2          |
|BAJAJFINSV|2017-03-08 06:25:00|3780.95     |138   |A2          |
|BAJAJFINSV|2017-02-10 06:42:00|3628.0      |359   |A2          |
```



METRICS DETAILS MESSAGES

PULL ☐ Enable ACK messages

Filter Filter messages

Publish time	Attribute keys	Message body	Ordering key	Ack
6 Apr 2025, 17:02:18	–	Traded Volume more than 2% of its average: BAJAJFINSV at 2017-01-24 05:09:00	–	Deadline exceeded
6 Apr 2025, 17:02:18	–	Traded Volume more than 2% of its average: BAJAJFINSV at 2017-02-15 05:07:00	–	Deadline exceeded
6 Apr 2025, 17:02:18	–	Traded Volume more than 2% of its average: BAJAJFINSV at 2017-02-20 08:56:00	–	Deadline exceeded
6 Apr 2025, 17:02:18	–	Traded Volume more than 2% of its average: BAJAJFINSV at 2017-02-28 06:57:00	–	Deadline exceeded
6 Apr 2025, 17:02:18	–	Traded Volume more than 2% of its average: BAJAJFINSV at 2017-02-07 07:42:00	–	Deadline exceeded
6 Apr 2025, 17:02:18	–	Traded Volume more than 2% of its average: BAJAJFINSV at 2017-02-10 06:33:00	–	Deadline exceeded
6 Apr 2025, 17:02:18	–	Traded Volume more than 2% of its average: BAJAJFINSV at 2017-02-22 04:35:00	–	Deadline exceeded
6 Apr 2025, 17:02:18	–	Traded Volume more than 2% of its average: BAJAJFINSV at 2017-01-30 08:54:00	–	Deadline exceeded
6 Apr 2025, 17:02:18	–	Traded Volume more than 2% of its average: BAJAJFINSV at 2017-02-28 04:06:00	–	Deadline exceeded
6 Apr 2025, 17:02:18	–	Traded Volume more than 2% of its average: BAJAJFINSV at 2017-02-16 05:07:00	–	Deadline exceeded
6 Apr 2025, 17:02:18	–	Traded Volume more than 2% of its average: BAJAJFINSV at 2017-02-17 04:24:00	–	Deadline exceeded
6 Apr 2025, 17:02:18	–	Traded Volume more than 2% of its average: BAJAJFINSV at 2017-01-27 04:23:00	–	Deadline exceeded
6 Apr 2025, 17:02:18	–	Traded Volume more than 2% of its average: BAJAJFINSV at 2017-02-10 04:47:00	–	Deadline exceeded

----------X----------