Introduction to Big Data

Jan 2025 Term – Graded Assignment 9

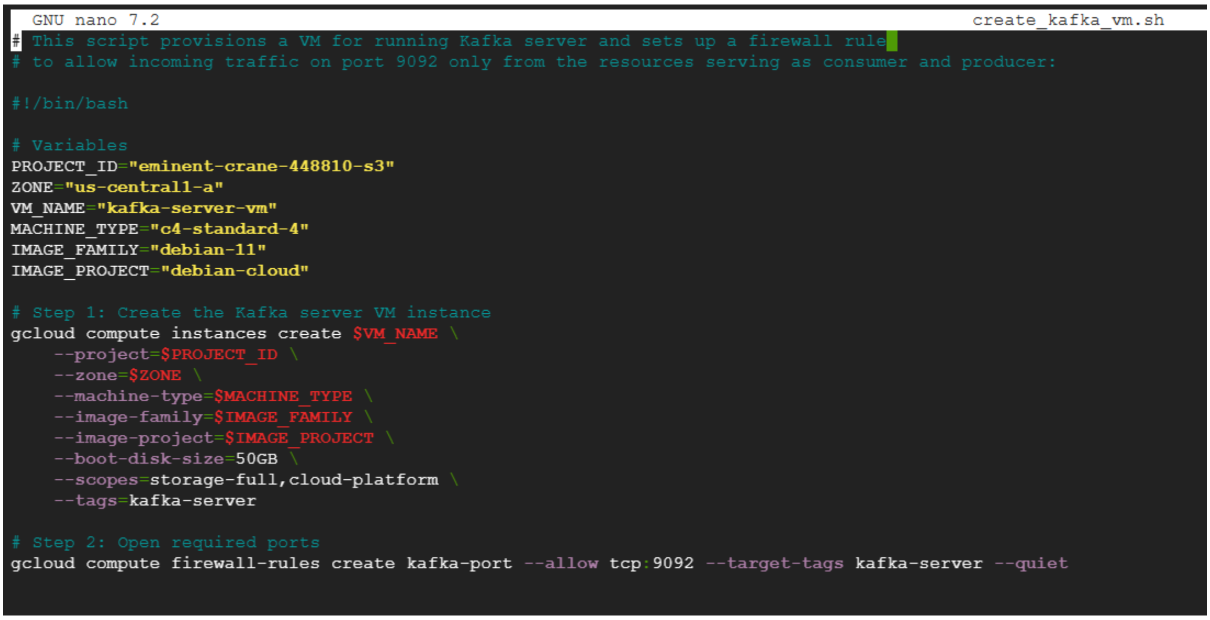
Name :- Satvik Chandrakar

Roll no :- 21f1000344

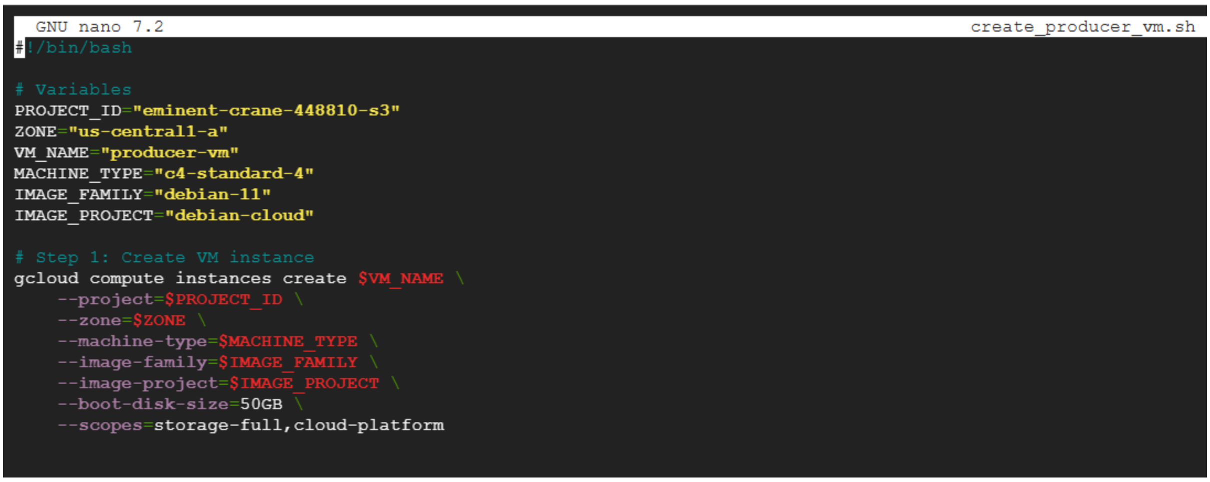
Task :- Convert the batch image classification use case given in the https://drive.google.com/file/d/1BufNhnDKvuLA0Vd59pdPK8bCbTg1JZu8/view?usp=sharing to a real-time execution model using Spark Streaming.

My Approach :-

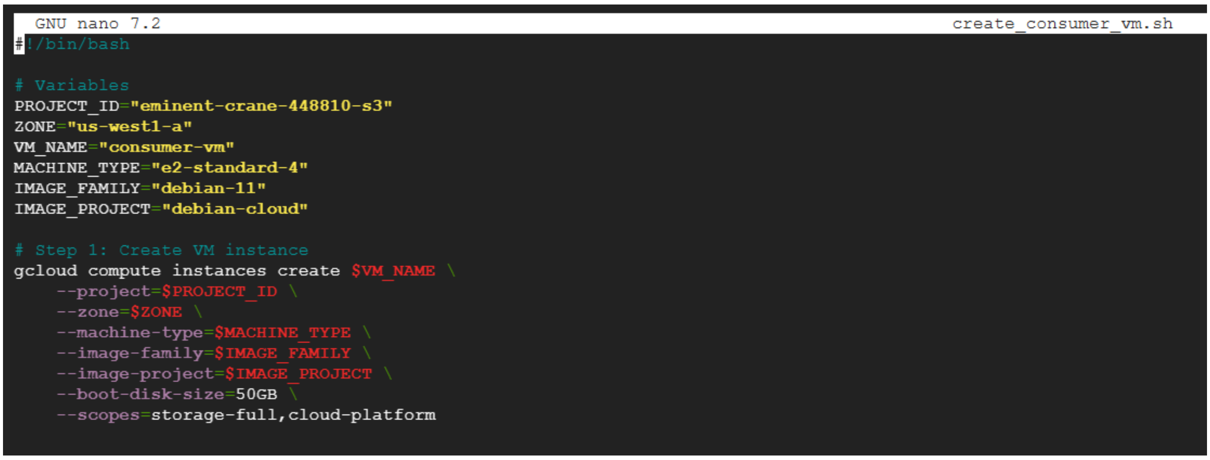
- Step 1 :- Opened the Google Cloud Shell
- Step 2 :- Run the following commands to authenticate:
  - *gcloud auth login*
  - *gcloud config set eminent-crane-448810-s3*
- Step 3 :- Provisioned three VMs for this assignment. One VM for serving as a kafka broker, one for serving as a producer and one for serving as a consumer. Producer downloads the dataset and writes one image data at a time to kafka topic. Consumer reads from the same kafka topic, processes the image data and predict its class using pretrained mobilenet_v2 from torchvision package.
  - create_kafka_vm.sh, create_producer_vm.sh and create_consumer_vm.sh were used for provisioning the VMs.



```
# 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
PROJECT_ID="eminent-crane-448810-s3"
ZONE="us-central1-a"
VM_NAME="kafka-server-vm"
MACHINE_TYPE="c4-standard-4"
IMAGE_FAMILY="debian-11"
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=50GB \
    --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
```

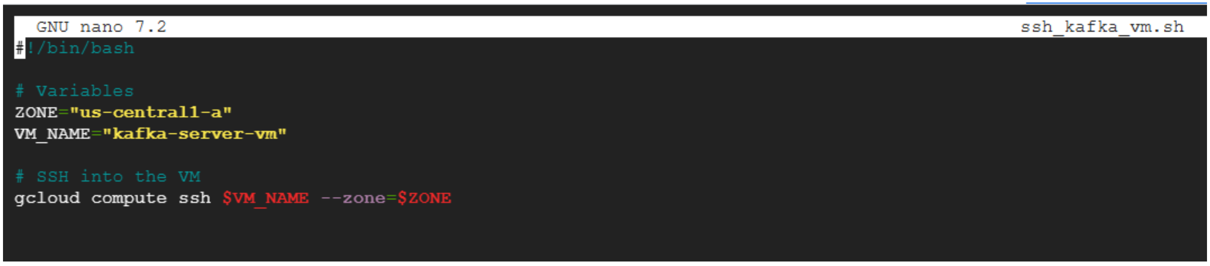


```
  GNU nano 7.2                                                                                create_producer_vm.sh
#!/bin/bash

# Variables
PROJECT_ID="eminent-crane-448810-s3"
ZONE="us-central1-a"
VM_NAME="producer-vm"
MACHINE_TYPE="c4-standard-4"
IMAGE_FAMILY="debian-11"
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=50GB \
    --scopes=storage-full,cloud-platform
```

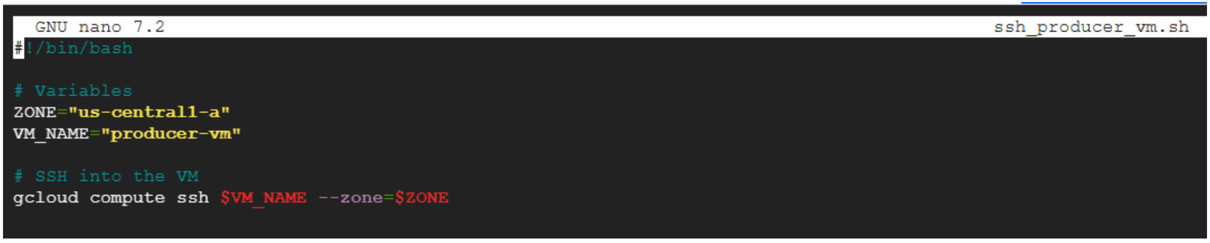


```
  GNU nano 7.2                                                                              create_consumer_vm.sh
#!/bin/bash

# Variables
PROJECT_ID="eminent-crane-448810-s3"
ZONE="us-west1-a"
VM_NAME="consumer-vm"
MACHINE_TYPE="e2-standard-4"
IMAGE_FAMILY="debian-11"
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=50GB \
    --scopes=storage-full,cloud-platform
```

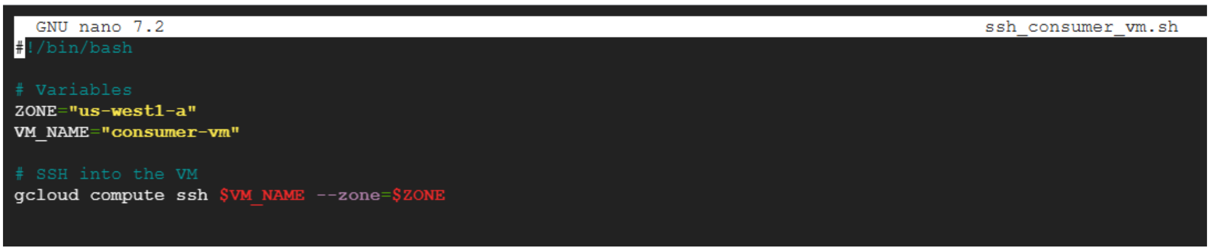
- Step 4 :- Created ssh_kafka_vm.sh, ssh_producer_vm.sh and ssh_consumer_vm.sh to SSH into the VMs.



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

# Variables
ZONE="us-central1-a"
VM_NAME="kafka-server-vm"

# SSH into the VM
gcloud compute ssh $VM_NAME --zone=$ZONE
```

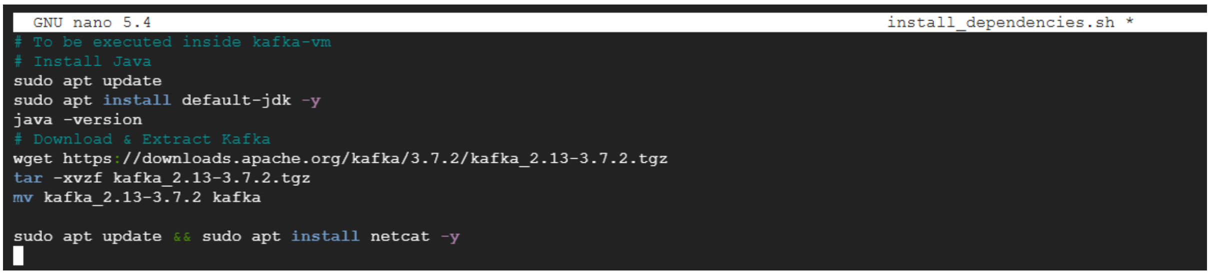


```
  GNU nano 7.2                                                                                 ssh_producer_vm.sh
#!/bin/bash

# Variables
ZONE="us-central1-a"
VM_NAME="producer-vm"

# SSH into the VM
gcloud compute ssh $VM_NAME --zone=$ZONE
```

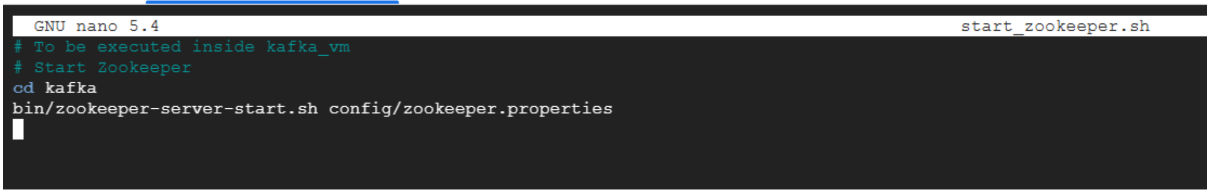


```
#!/bin/bash

# Variables
ZONE="us-west1-a"
VM_NAME="consumer-vm"

# SSH into the VM
gcloud compute ssh $VM_NAME --zone=$ZONE
```

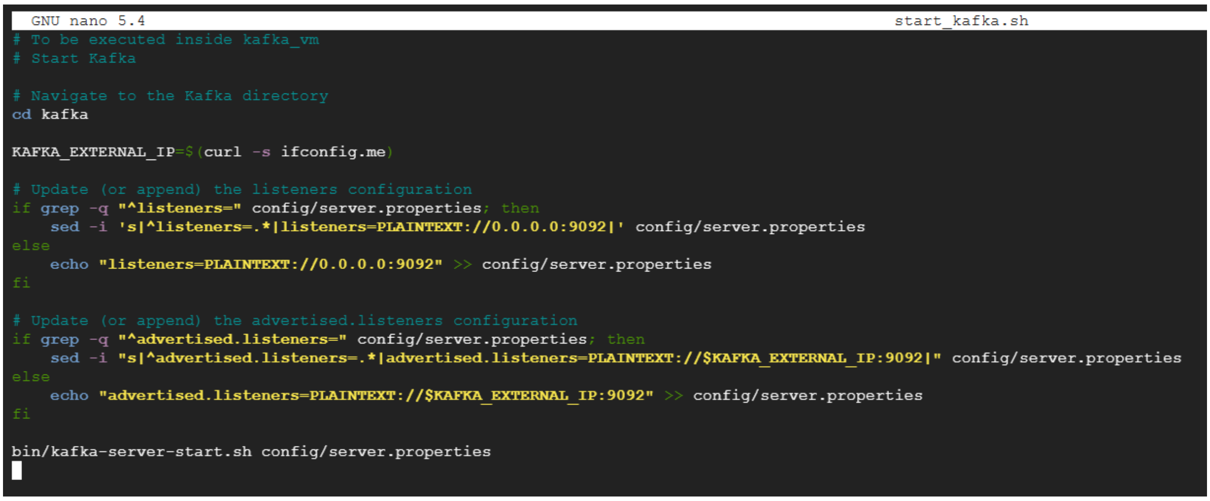
- Step 5 :- SSH into the kafka-vm, installed the dependencies, started the zookeeper & kafka server and created the topic.



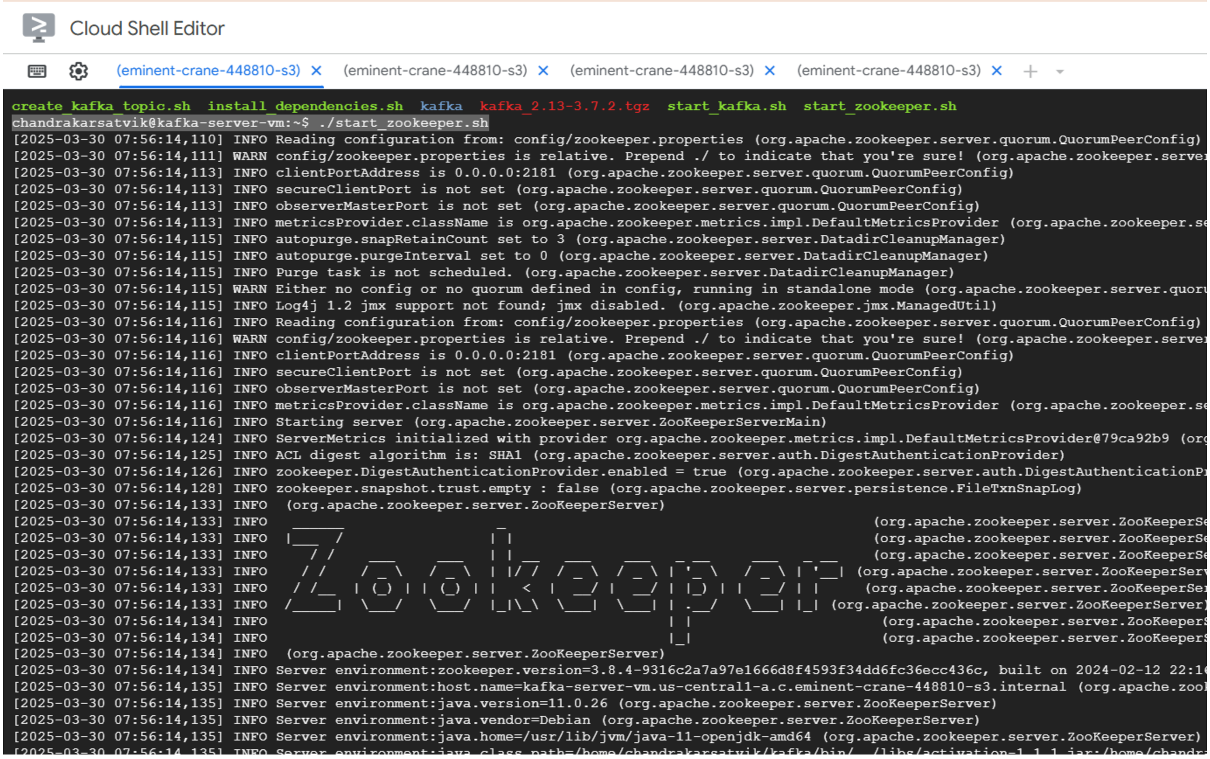
```
# To be executed inside kafka-vm
# Install Java
sudo apt update
sudo apt install default-jdk -y
java -version
# Download & Extract Kafka
wget https://downloads.apache.org/kafka/3.7.2/kafka_2.13-3.7.2.tgz
tar -xvzf kafka_2.13-3.7.2.tgz
mv kafka_2.13-3.7.2 kafka

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



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

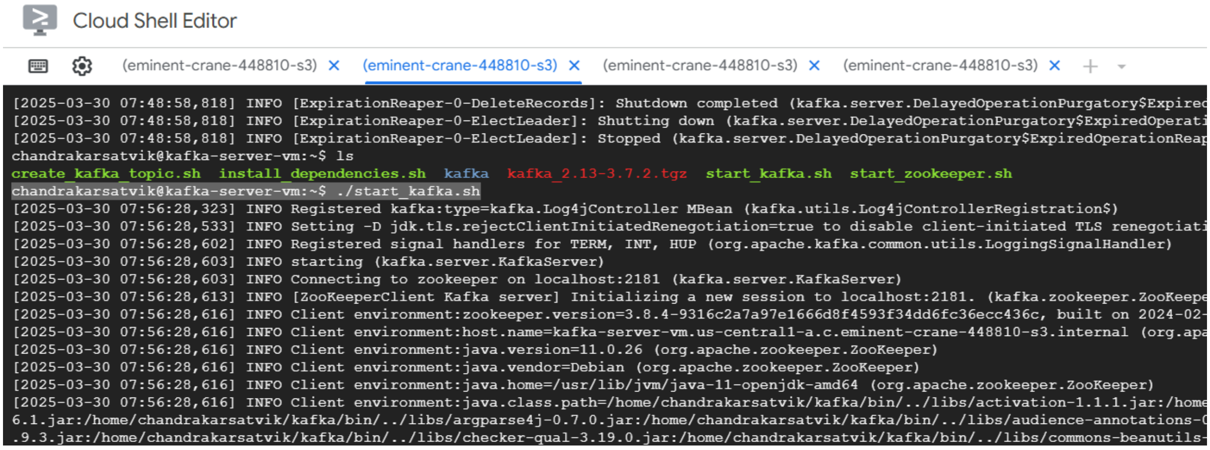


```
# To be executed inside kafka_vm
# Start Kafka

# Navigate to the Kafka directory
cd kafka

KAFKA_EXTERNAL_IP=$(curl -s ifconfig.me)

# Update (or append) the listeners configuration
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
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
```


Cloud Shell Editor
(eminent-crane-448810-s3)
(eminent-crane-448810-s3)
(eminent-crane-448810-s3)
(eminent-crane-448810-s3)

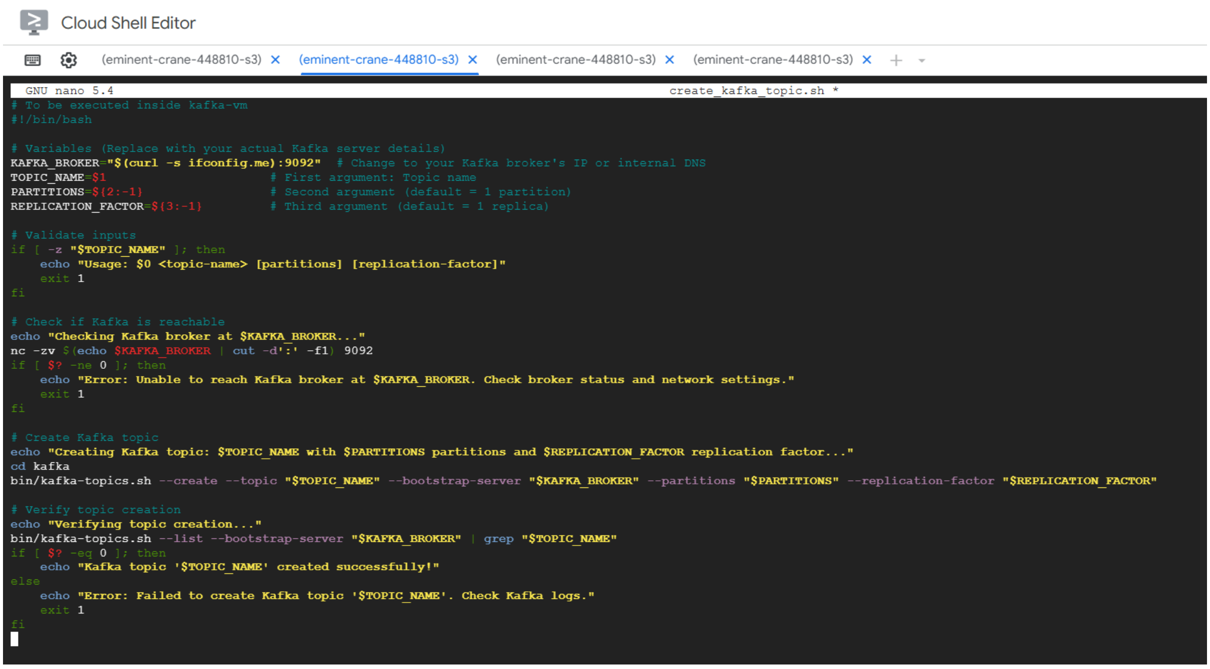
```
create_kafka_topic.sh  install_dependencies.sh  kafka  kafka_2.13-3.7.2.tgz  start_kafka.sh  start_zookeeper.sh
chandrakarsatvik@kafka-server-vm:~$ ./start_zookeeper.sh
[2025-03-30 07:56:14,110] INFO Reading configuration from: config/zookeeper.properties (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2025-03-30 07:56:14,111] WARN config/zookeeper.properties is relative. Prepend ./ to indicate that you're sure! (org.apache.zookeeper.serve
[2025-03-30 07:56:14,113] INFO clientPortAddress is 0.0.0.0:2181 (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2025-03-30 07:56:14,113] INFO secureClientPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2025-03-30 07:56:14,113] INFO observerMasterPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2025-03-30 07:56:14,113] INFO metricsProvider.className is org.apache.zookeeper.metrics.impl.DefaultMetricsProvider (org.apache.zookeeper.se
[2025-03-30 07:56:14,115] INFO autopurge.snapRetainCount set to 3 (org.apache.zookeeper.server.DatadirCleanupManager)
[2025-03-30 07:56:14,115] INFO autopurge.purgeInterval set to 0 (org.apache.zookeeper.server.DatadirCleanupManager)
[2025-03-30 07:56:14,115] INFO Purge task is not scheduled. (org.apache.zookeeper.server.DatadirCleanupManager)
[2025-03-30 07:56:14,115] WARN Either no config or no quorum defined in config, running in standalone mode (org.apache.zookeeper.server.quoru
[2025-03-30 07:56:14,115] INFO Log4j 1.2 jmx support not found; jmx disabled. (org.apache.zookeeper.jmx.ManagedUtil)
[2025-03-30 07:56:14,116] INFO Reading configuration from: config/zookeeper.properties (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2025-03-30 07:56:14,116] WARN config/zookeeper.properties is relative. Prepend ./ to indicate that you're sure! (org.apache.zookeeper.serve
[2025-03-30 07:56:14,116] INFO clientPortAddress is 0.0.0.0:2181 (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2025-03-30 07:56:14,116] INFO secureClientPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2025-03-30 07:56:14,116] INFO observerMasterPort is not set (org.apache.zookeeper.server.quorum.QuorumPeerConfig)
[2025-03-30 07:56:14,116] INFO metricsProvider.className is org.apache.zookeeper.metrics.impl.DefaultMetricsProvider (org.apache.zookeeper.se
[2025-03-30 07:56:14,116] INFO Starting server (org.apache.zookeeper.server.ZooKeeperServerMain)
[2025-03-30 07:56:14,124] INFO ServerMetrics initialized with provider org.apache.zookeeper.metrics.impl.DefaultMetricsProvider@79ca92b9 (org
[2025-03-30 07:56:14,125] INFO ACL digest algorithm is: SHA1 (org.apache.zookeeper.server.auth.DigestAuthenticationProvider)
[2025-03-30 07:56:14,126] INFO zookeeper.DigestAuthenticationProvider.enabled = true (org.apache.zookeeper.server.auth.DigestAuthenticationPr
[2025-03-30 07:56:14,128] INFO zookeeper.snapshot.trust.empty : false (org.apache.zookeeper.server.persistence.FileTxnSnapLog)
[2025-03-30 07:56:14,133] INFO  (org.apache.zookeeper.server.ZooKeeperServer)
[2025-03-30 07:56:14,133] INFO   ______                  _                                          (org.apache.zookeeper.server.ZooKeeperSe
[2025-03-30 07:56:14,133] INFO  |___  /                 | |                                         (org.apache.zookeeper.server.ZooKeeperSe
[2025-03-30 07:56:14,133] INFO     / /    ___     ___   | | __   ___    ___   _ __     ___   _ __   (org.apache.zookeeper.server.ZooKeeperSe
[2025-03-30 07:56:14,133] INFO    / /    / _ \   / _ \  | |/ /  / _ \  / _ \ | '_ \   / _ \ | '__| (org.apache.zookeeper.server.ZooKeeperServ
[2025-03-30 07:56:14,133] INFO   / /__  | (_) | | (_) | |   <  |  __/ |  __/ | |_) | |  __/ | |     (org.apache.zookeeper.server.ZooKeeperSe
[2025-03-30 07:56:14,133] INFO  /_____|  \___/   \___/  |_|\_\  \___|  \___| | .__/   \___| |_| (org.apache.zookeeper.server.ZooKeeperServer)
[2025-03-30 07:56:14,134] INFO                                               | |                     (org.apache.zookeeper.server.ZooKeeperS
[2025-03-30 07:56:14,134] INFO                                               |_|                     (org.apache.zookeeper.server.ZooKeeperS
[2025-03-30 07:56:14,134] INFO  (org.apache.zookeeper.server.ZooKeeperServer)
[2025-03-30 07:56:14,134] INFO Server environment:zookeeper.version=3.8.4-9316c2a7a97e1666d8f4593f34dd6fc36ecc436c, built on 2024-02-12 22:1
[2025-03-30 07:56:14,135] INFO Server environment:host.name=kafka-server-vm.us-central1-a.c.eminent-crane-448810-s3.internal (org.apache.zoo
[2025-03-30 07:56:14,135] INFO Server environment:java.version=11.0.26 (org.apache.zookeeper.server.ZooKeeperServer)
[2025-03-30 07:56:14,135] INFO Server environment:java.vendor=Debian (org.apache.zookeeper.server.ZooKeeperServer)
[2025-03-30 07:56:14,135] INFO Server environment:java.home=/usr/lib/jvm/java-11-openjdk-amd64 (org.apache.zookeeper.server.ZooKeeperServer)
[2025-03-30 07:56:14,135] INFO Server environment:java.class.path=/home/chandrakarsatvik/kafka/bin/../libs/activation-1.1.1.jar:/home/chandr
```

Zookeeper started in the first terminal of kafka-vm


Cloud Shell Editor
(eminent-crane-448810-s3)
(eminent-crane-448810-s3)
(eminent-crane-448810-s3)
(eminent-crane-448810-s3)

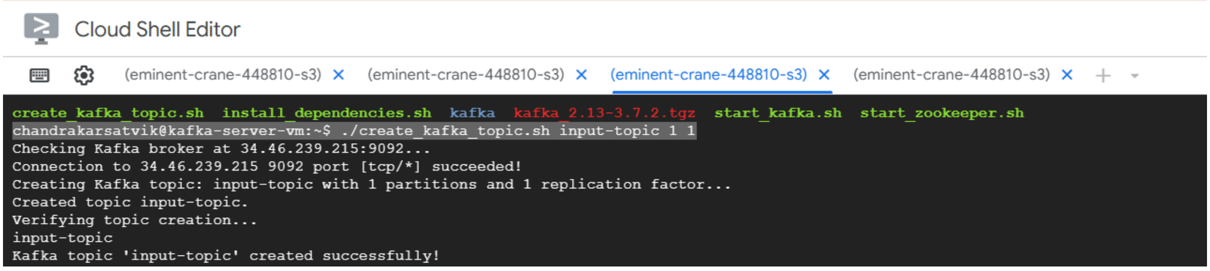
```
[2025-03-30 07:48:58,818] INFO [ExpirationReaper-0-DeleteRecords]: Shutdown completed (kafka.server.DelayedOperationPurgatory$Expired
[2025-03-30 07:48:58,818] INFO [ExpirationReaper-0-ElectLeader]: Shutting down (kafka.server.DelayedOperationPurgatory$ExpiredOperati
[2025-03-30 07:48:58,818] INFO [ExpirationReaper-0-ElectLeader]: Stopped (kafka.server.DelayedOperationPurgatory$ExpiredOperationReap
chandrakarsatvik@kafka-server-vm:~$ 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
[2025-03-30 07:56:28,323] INFO Registered kafka:type=kafka.Log4jController MBean (kafka.utils.Log4jControllerRegistration$)
[2025-03-30 07:56:28,533] INFO Setting -D jdk.tls.rejectClientInitiatedRenegotiation=true to disable client-initiated TLS renegotiati
[2025-03-30 07:56:28,602] INFO Registered signal handlers for TERM, INT, HUP (org.apache.kafka.common.utils.LoggingSignalHandler)
[2025-03-30 07:56:28,603] INFO starting (kafka.server.KafkaServer)
[2025-03-30 07:56:28,603] INFO Connecting to zookeeper on localhost:2181 (kafka.server.KafkaServer)
[2025-03-30 07:56:28,613] INFO [ZooKeeperClient Kafka server] Initializing a new session to localhost:2181. (kafka.zookeeper.ZooKeepe
[2025-03-30 07:56:28,616] INFO Client environment:zookeeper.version=3.8.4-9316c2a7a97e1666d8f4593f34dd6fc36ecc436c, built on 2024-02-
[2025-03-30 07:56:28,616] INFO Client environment:host.name=kafka-server-vm.us-central1-a.c.eminent-crane-448810-s3.internal (org.apa
[2025-03-30 07:56:28,616] INFO Client environment:java.version=11.0.26 (org.apache.zookeeper.ZooKeeper)
[2025-03-30 07:56:28,616] INFO Client environment:java.vendor=Debian (org.apache.zookeeper.ZooKeeper)
[2025-03-30 07:56:28,616] INFO Client environment:java.home=/usr/lib/jvm/java-11-openjdk-amd64 (org.apache.zookeeper.ZooKeeper)
[2025-03-30 07:56:28,616] INFO Client environment:java.class.path=/home/chandrakarsatvik/kafka/bin/../libs/activation-1.1.1.jar:/home
6.1.jar:/home/chandrakarsatvik/kafka/bin/../libs/argparse4j-0.7.0.jar:/home/chandrakarsatvik/kafka/bin/../libs/audience-annotations-0
.9.3.jar:/home/chandrakarsatvik/kafka/bin/../libs/checker-qual-3.19.0.jar:/home/chandrakarsatvik/kafka/bin/../libs/commons-beanutils-
```

Kafka server started in the second terminal of kafka-vm


Cloud Shell Editor
(eminent-crane-448810-s3)
(eminent-crane-448810-s3)
(eminent-crane-448810-s3)
(eminent-crane-448810-s3)

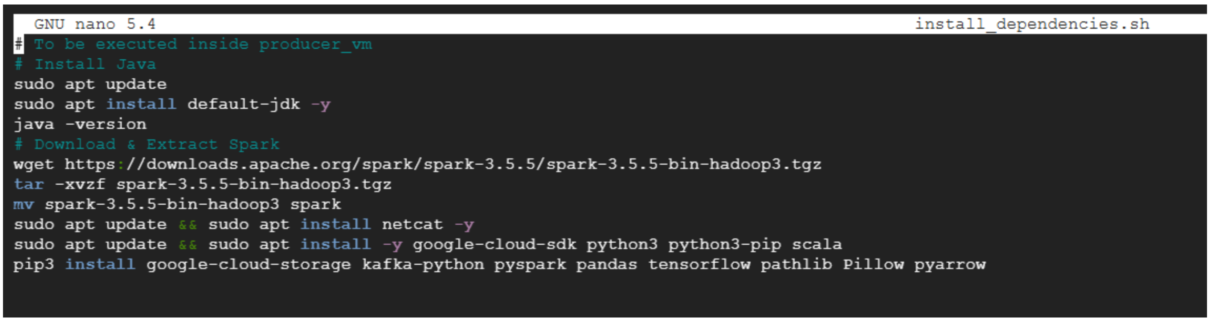
```
  GNU nano 5.4                                                                    create_kafka_topic.sh *
# To be executed inside kafka-vm
#!/bin/bash

# Variables (Replace with your actual Kafka server details)
KAFKA_BROKER="$(curl -s ifconfig.me):9092"  # Change to your Kafka broker's IP or internal DNS
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 input-topic. create_kafka_topic.sh was executed on the third terminal of the kafka-vm.

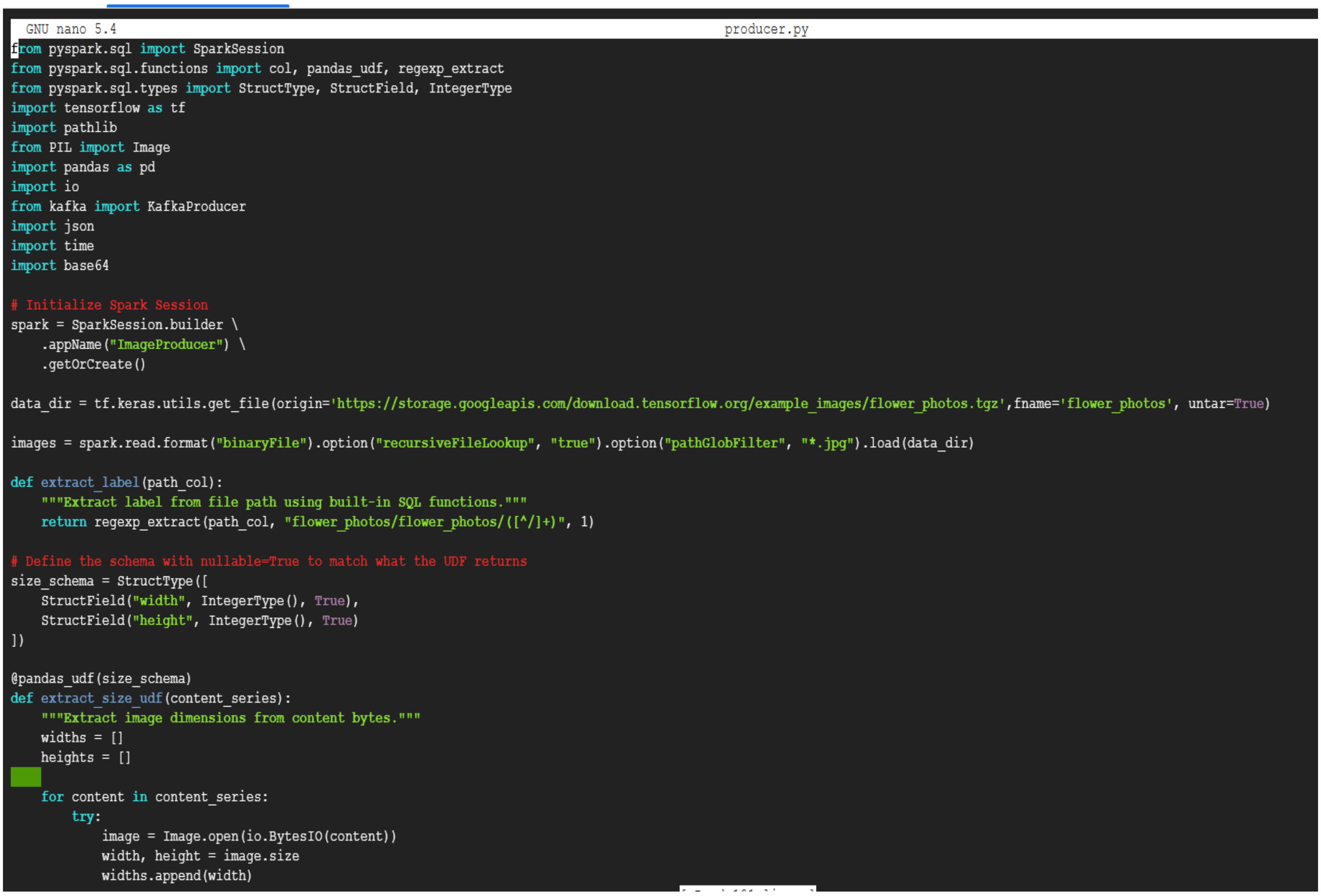

Cloud Shell Editor
(eminent-crane-448810-s3)
(eminent-crane-448810-s3)
(eminent-crane-448810-s3)
(eminent-crane-448810-s3)

```
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:~$ ./create_kafka_topic.sh input-topic 1 1
Checking Kafka broker at 34.46.239.215:9092...
Connection to 34.46.239.215 9092 port [tcp/*] succeeded!
Creating Kafka topic: input-topic with 1 partitions and 1 replication factor...
Created topic input-topic.
Verifying topic creation...
input-topic
Kafka topic 'input-topic' created successfully!
```

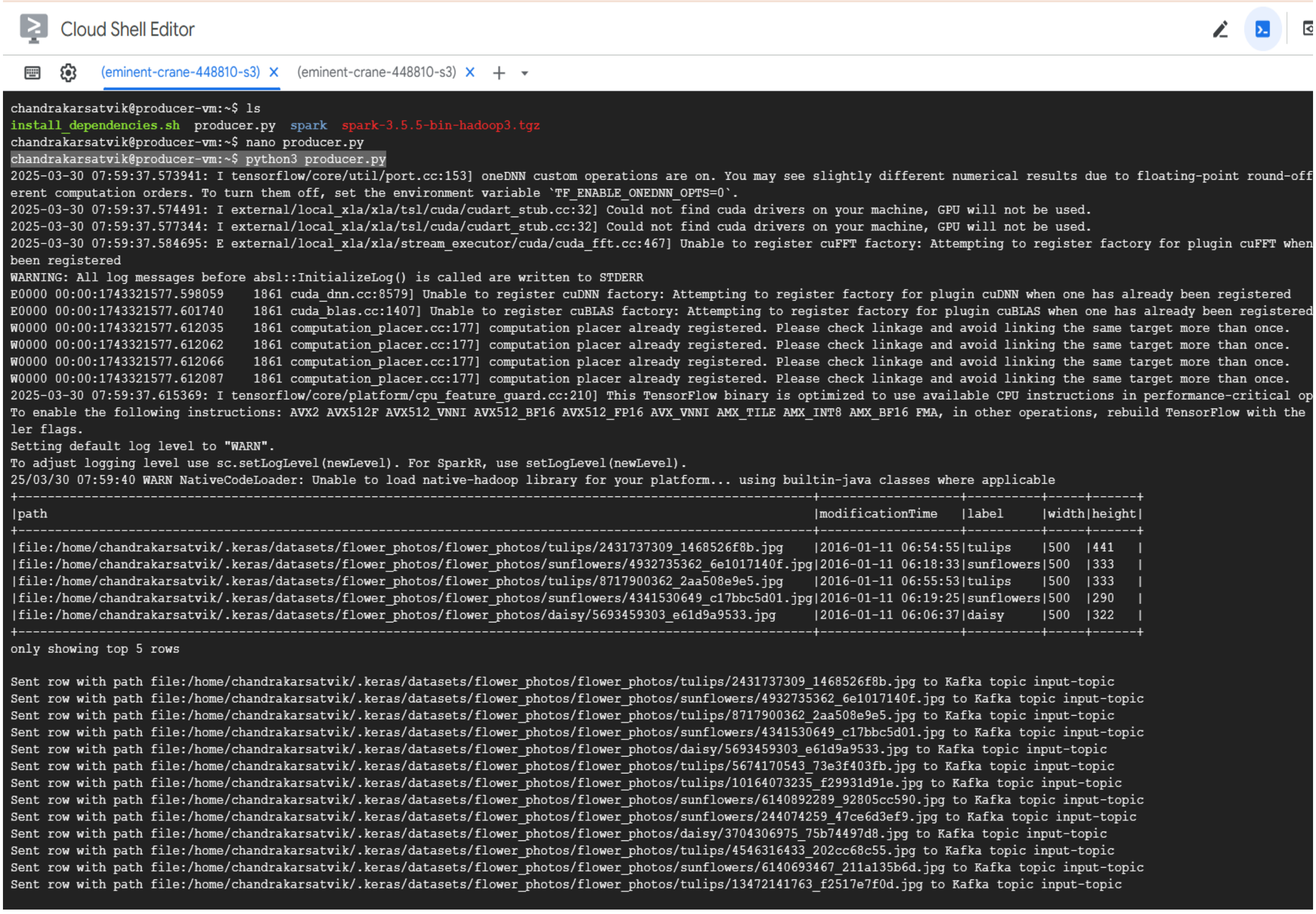
- Step 6 :- SSH into the producer-vm, installed the dependencies, created the producer.py file to download the dataset and write one image data at a time to the kafka topic.



```
  GNU nano 5.4                                                         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 netcat -y
sudo apt update && sudo apt install -y google-cloud-sdk python3 python3-pip scala
pip3 install google-cloud-storage kafka-python pyspark pandas tensorflow pathlib Pillow pyarrow
```

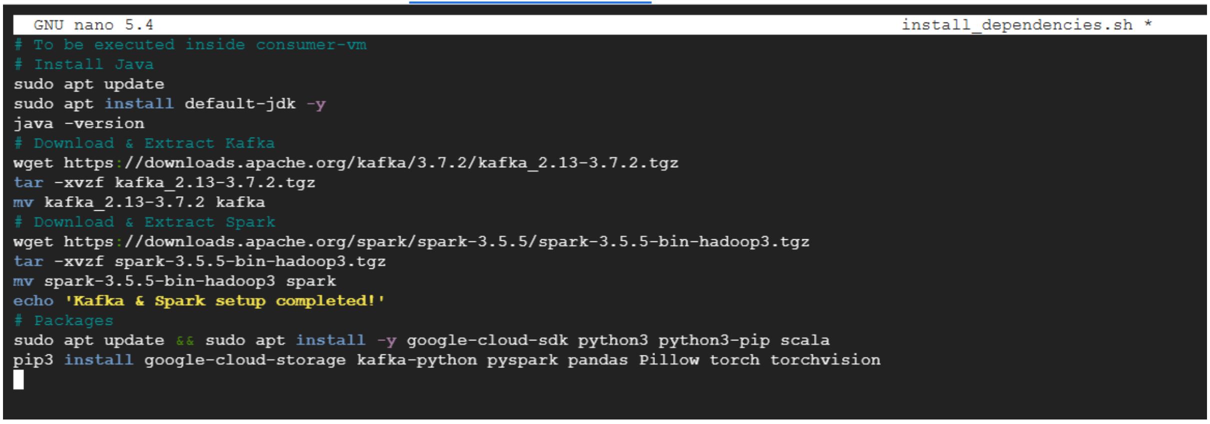


```
  GNU nano 5.4                                                                                          producer.py
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, pandas_udf, regexp_extract
from pyspark.sql.types import StructType, StructField, IntegerType
import tensorflow as tf
import pathlib
from PIL import Image
import pandas as pd
import io
from kafka import KafkaProducer
import json
import time
import base64

# Initialize Spark Session
spark = SparkSession.builder \
    .appName("ImageProducer") \
    .getOrCreate()

data_dir = tf.keras.utils.get_file(origin='https://storage.googleapis.com/download.tensorflow.org/example_images/flower_photos.tgz',fname='flower_photos', untar=True)

images = spark.read.format("binaryFile").option("recursiveFileLookup", "true").option("pathGlobFilter", "*.jpg").load(data_dir)

def extract_label(path_col):
    """Extract label from file path using built-in SQL functions."""
    return regexp_extract(path_col, "flower_photos/flower_photos/([^/]+)", 1)

# Define the schema with nullable=True to match what the UDF returns
size_schema = StructType([
    StructField("width", IntegerType(), True),
    StructField("height", IntegerType(), True)
])

@pandas_udf(size_schema)
def extract_size_udf(content_series):
    """Extract image dimensions from content bytes."""
    widths = []
    heights = []

    for content in content_series:
        try:
            image = Image.open(io.BytesIO(content))
            width, height = image.size
            widths.append(width)
```

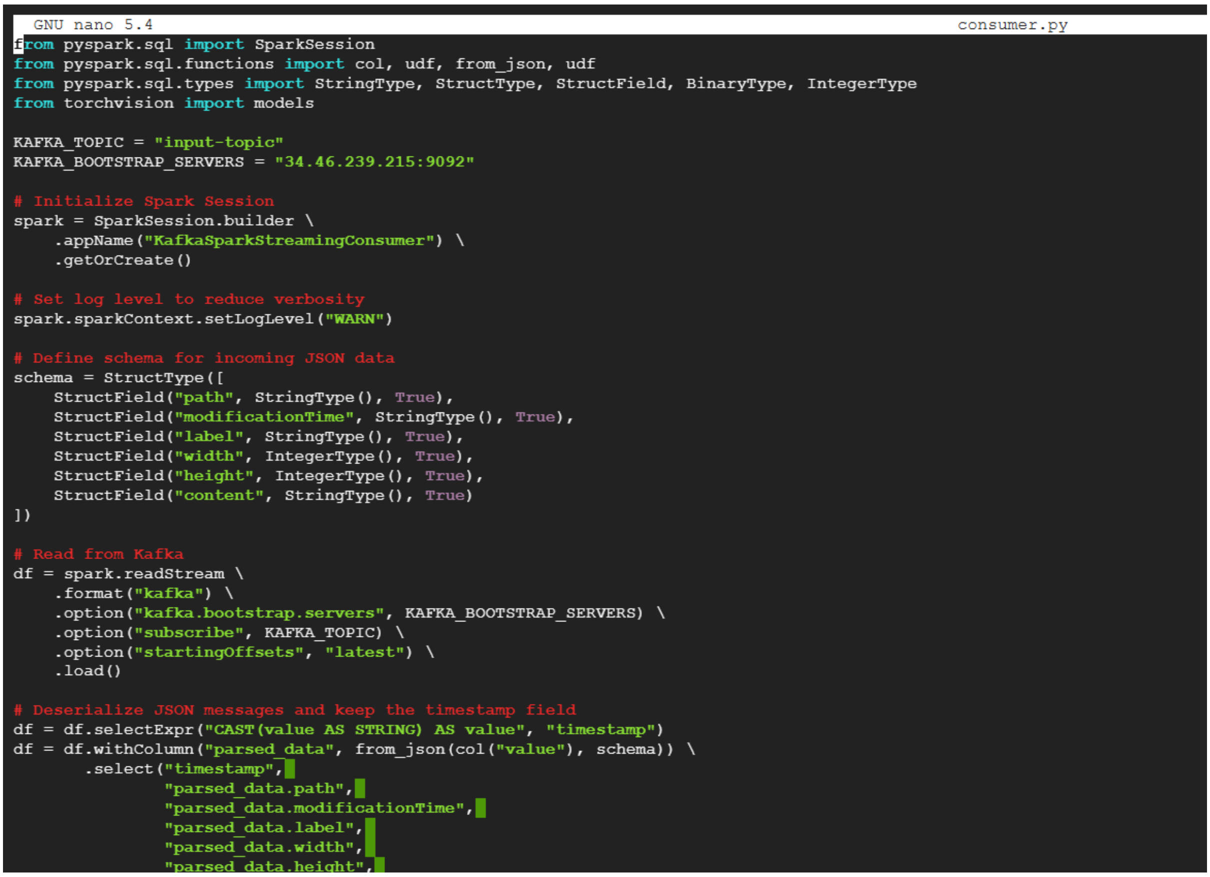


```
chandrakarsatvik@producer-vm:~$ ls
install_dependencies.sh  producer.py  spark  spark-3.5.5-bin-hadoop3.tgz
chandrakarsatvik@producer-vm:~$ nano producer.py
chandrakarsatvik@producer-vm:~$ python3 producer.py
2025-03-30 07:59:37.573941: I tensorflow/core/util/port.cc:153] oneDNN custom operations are on. You may see slightly different numerical results due to floating-point round-off
erent computation orders. To turn them off, set the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-03-30 07:59:37.574491: I external/local_xla/xla/tsl/cuda/cudart_stub.cc:32] Could not find cuda drivers on your machine, GPU will not be used.
2025-03-30 07:59:37.577344: I external/local_xla/xla/tsl/cuda/cudart_stub.cc:32] Could not find cuda drivers on your machine, GPU will not be used.
2025-03-30 07:59:37.584695: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:467] Unable to register cuFFT factory: Attempting to register factory for plugin cuFFT when
been registered
WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
E0000 00:00:1743321577.598059    1861 cuda_dnn.cc:8579] Unable to register cuDNN factory: Attempting to register factory for plugin cuDNN when one has already been registered
E0000 00:00:1743321577.601740    1861 cuda_blas.cc:1407] Unable to register cuBLAS factory: Attempting to register factory for plugin cuBLAS when one has already been registered
W0000 00:00:1743321577.612035    1861 computation_placer.cc:177] computation placer already registered. Please check linkage and avoid linking the same target more than once.
W0000 00:00:1743321577.612062    1861 computation_placer.cc:177] computation placer already registered. Please check linkage and avoid linking the same target more than once.
W0000 00:00:1743321577.612066    1861 computation_placer.cc:177] computation placer already registered. Please check linkage and avoid linking the same target more than once.
W0000 00:00:1743321577.612087    1861 computation_placer.cc:177] computation placer already registered. Please check linkage and avoid linking the same target more than once.
2025-03-30 07:59:37.615369: I tensorflow/core/platform/cpu_feature_guard.cc:210] This TensorFlow binary is optimized to use available CPU instructions in performance-critical op
To enable the following instructions: AVX2 AVX512F AVX512_VNNI AVX512_BF16 AVX512_FP16 AVX_VNNI AMX_TILE AMX_INT8 AMX_BF16 FMA, in other operations, rebuild TensorFlow with the
ler flags.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
25/03/30 07:59:40 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
+------------------------------------------------------------------------------------------------------------+-------------------+----------+-----+------+
|path                                                                                                        |modificationTime   |label     |width|height|
+------------------------------------------------------------------------------------------------------------+-------------------+----------+-----+------+
|file:/home/chandrakarsatvik/.keras/datasets/flower_photos/flower_photos/tulips/2431737309_1468526f8b.jpg    |2016-01-11 06:54:55|tulips    |500  |441   |
|file:/home/chandrakarsatvik/.keras/datasets/flower_photos/flower_photos/sunflowers/4932735362_6e1017140f.jpg|2016-01-11 06:18:33|sunflowers|500  |333   |
|file:/home/chandrakarsatvik/.keras/datasets/flower_photos/flower_photos/tulips/8717900362_2aa508e9e5.jpg    |2016-01-11 06:55:53|tulips    |500  |333   |
|file:/home/chandrakarsatvik/.keras/datasets/flower_photos/flower_photos/sunflowers/4341530649_c17bbc5d01.jpg|2016-01-11 06:19:25|sunflowers|500  |290   |
|file:/home/chandrakarsatvik/.keras/datasets/flower_photos/flower_photos/daisy/5693459303_e61d9a9533.jpg     |2016-01-11 06:06:37|daisy     |500  |322   |
+------------------------------------------------------------------------------------------------------------+-------------------+----------+-----+------+
only showing top 5 rows

Sent row with path file:/home/chandrakarsatvik/.keras/datasets/flower_photos/flower_photos/tulips/2431737309_1468526f8b.jpg to Kafka topic input-topic
Sent row with path file:/home/chandrakarsatvik/.keras/datasets/flower_photos/flower_photos/sunflowers/4932735362_6e1017140f.jpg to Kafka topic input-topic
Sent row with path file:/home/chandrakarsatvik/.keras/datasets/flower_photos/flower_photos/tulips/8717900362_2aa508e9e5.jpg to Kafka topic input-topic
Sent row with path file:/home/chandrakarsatvik/.keras/datasets/flower_photos/flower_photos/sunflowers/4341530649_c17bbc5d01.jpg to Kafka topic input-topic
Sent row with path file:/home/chandrakarsatvik/.keras/datasets/flower_photos/flower_photos/daisy/5693459303_e61d9a9533.jpg to Kafka topic input-topic
Sent row with path file:/home/chandrakarsatvik/.keras/datasets/flower_photos/flower_photos/tulips/5674170543_73e3f403fb.jpg to Kafka topic input-topic
Sent row with path file:/home/chandrakarsatvik/.keras/datasets/flower_photos/flower_photos/tulips/10164073235_f29931d91e.jpg to Kafka topic input-topic
Sent row with path file:/home/chandrakarsatvik/.keras/datasets/flower_photos/flower_photos/sunflowers/6140892289_92805cc590.jpg to Kafka topic input-topic
Sent row with path file:/home/chandrakarsatvik/.keras/datasets/flower_photos/flower_photos/sunflowers/244074259_47ce6d3ef9.jpg to Kafka topic input-topic
Sent row with path file:/home/chandrakarsatvik/.keras/datasets/flower_photos/flower_photos/daisy/3704306975_75b74497d8.jpg to Kafka topic input-topic
Sent row with path file:/home/chandrakarsatvik/.keras/datasets/flower_photos/flower_photos/tulips/4546316433_202cc68c55.jpg to Kafka topic input-topic
Sent row with path file:/home/chandrakarsatvik/.keras/datasets/flower_photos/flower_photos/sunflowers/6140693467_211a135b6d.jpg to Kafka topic input-topic
Sent row with path file:/home/chandrakarsatvik/.keras/datasets/flower_photos/flower_photos/tulips/13472141763_f2517e7f0d.jpg to Kafka topic input-topic
```

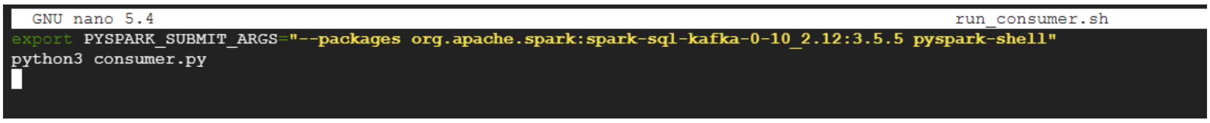
- Step 7 :- SSH into the consumer-vm, installed the dependencies, created the consumer.py file to the image data from the kafka topic using spark streaming, process the image and predict its class using the using pretrained mobilenet_v2 from torchvision package.



```
# To be executed inside consumer-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 Pillow torch torchvision
```

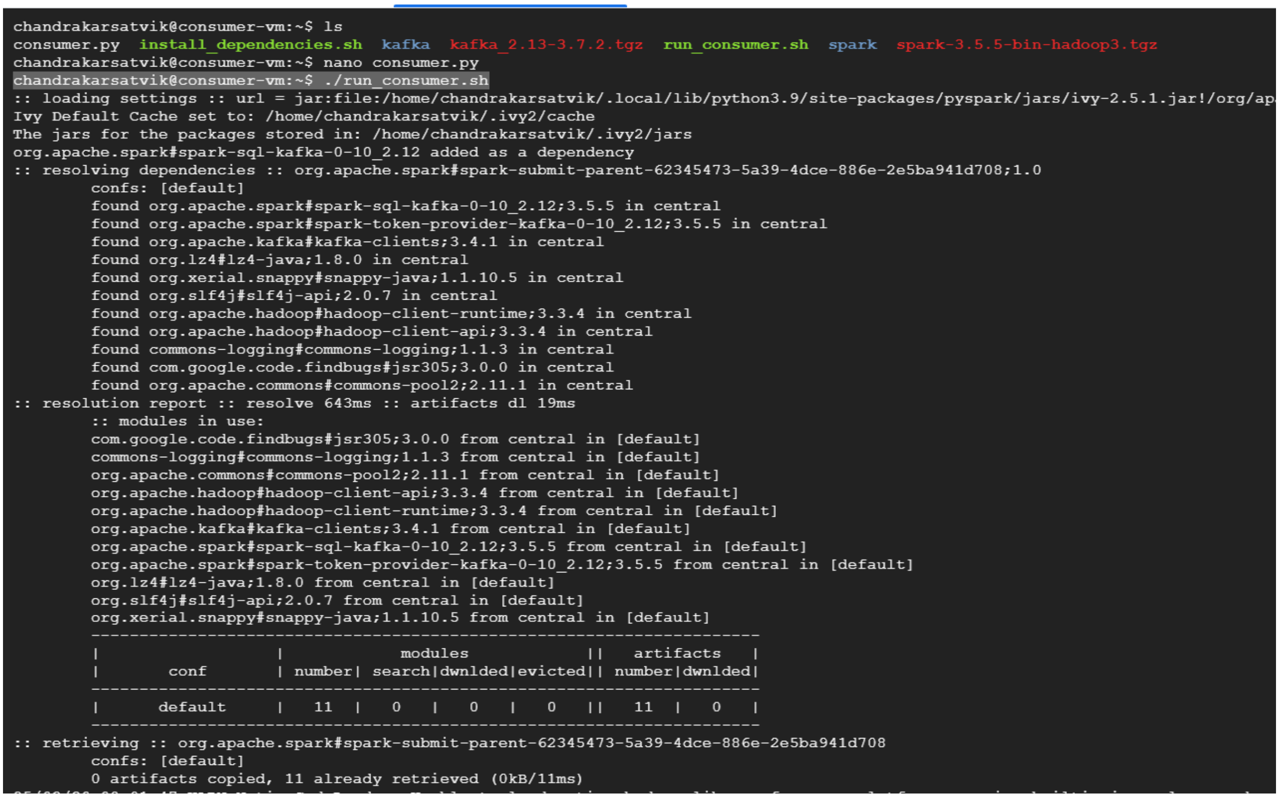


```
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, udf, from_json, udf
from pyspark.sql.types import StringType, StructType, StructField, BinaryType, IntegerType
from torchvision import models

KAFKA_TOPIC = "input-topic"
KAFKA_BOOTSTRAP_SERVERS = "34.46.239.215:9092"

# Initialize Spark Session
spark = SparkSession.builder \
    .appName("KafkaSparkStreamingConsumer") \
    .getOrCreate()

# Set log level to reduce verbosity
spark.sparkContext.setLogLevel("WARN")

# Define schema for incoming JSON data
schema = StructType([
    StructField("path", StringType(), True),
    StructField("modificationTime", StringType(), True),
    StructField("label", StringType(), True),
    StructField("width", IntegerType(), True),
    StructField("height", IntegerType(), True),
    StructField("content", StringType(), True)
])

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

# Deserialize JSON messages and keep the timestamp field
df = df.selectExpr("CAST(value AS STRING) AS value", "timestamp")
df = df.withColumn("parsed_data", from_json(col("value"), schema)) \
       .select("timestamp",
               "parsed_data.path",
               "parsed_data.modificationTime",
               "parsed_data.label",
               "parsed_data.width",
               "parsed_data.height",
```

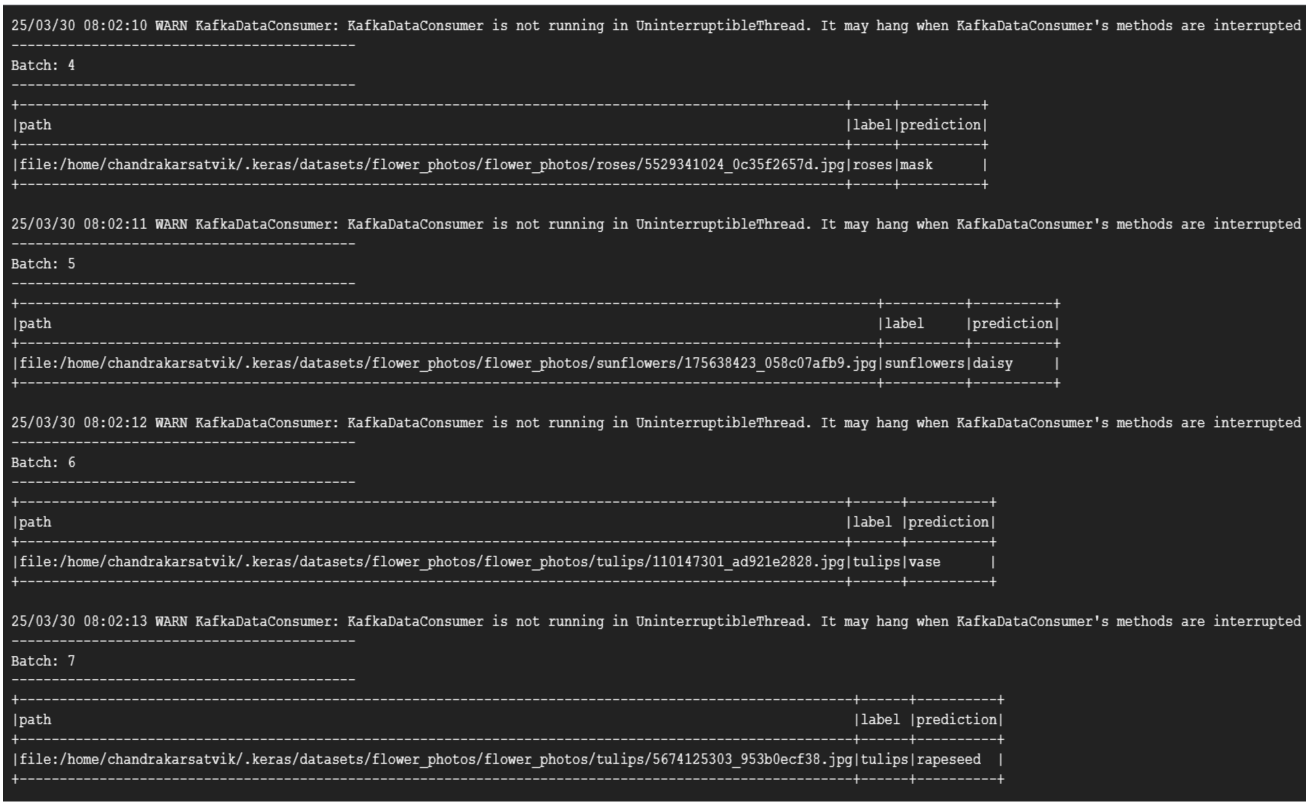


```
export PYSPARK_SUBMIT_ARGS="--packages org.apache.spark:spark-sql-kafka-0-10_2.12:3.5.5 pyspark-shell"
python3 consumer.py
```

run_consumer.sh was use to execute the consumer.py



```
chandrakarsatvik@consumer-vm:~$ ls
consumer.py  install_dependencies.sh  kafka  kafka_2.13-3.7.2.tgz  run_consumer.sh  spark  spark-3.5.5-bin-hadoop3.tgz
chandrakarsatvik@consumer-vm:~$ nano consumer.py
chandrakarsatvik@consumer-vm:~$ ./run_consumer.sh
:: loading settings :: url = jar:file:/home/chandrakarsatvik/.local/lib/python3.9/site-packages/pyspark/jars/ivy-2.5.1.jar!/org/ap
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-62345473-5a39-4dce-886e-2e5ba941d708;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 643ms :: artifacts dl 19ms
        :: 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-62345473-5a39-4dce-886e-2e5ba941d708
        confs: [default]
        0 artifacts copied, 11 already retrieved (0kB/11ms)
```



```
25/03/30 08:02:10 WARN KafkaDataConsumer: KafkaDataConsumer is not running in UninterruptibleThread. It may hang when KafkaDataConsumer's methods are interrupted
-------------------------------------------
Batch: 4
-------------------------------------------
+-------------------------------------------------------------------------------------------------------+-----+----------+
|path                                                                                                   |label|prediction|
+-------------------------------------------------------------------------------------------------------+-----+----------+
|file:/home/chandrakarsatvik/.keras/datasets/flower_photos/flower_photos/roses/5529341024_0c35f2657d.jpg|roses|mask      |
+-------------------------------------------------------------------------------------------------------+-----+----------+

25/03/30 08:02:11 WARN KafkaDataConsumer: KafkaDataConsumer is not running in UninterruptibleThread. It may hang when KafkaDataConsumer's methods are interrupted
-------------------------------------------
Batch: 5
-------------------------------------------
+-----------------------------------------------------------------------------------------------------------+----------+----------+
|path                                                                                                       |label     |prediction|
+-----------------------------------------------------------------------------------------------------------+----------+----------+
|file:/home/chandrakarsatvik/.keras/datasets/flower_photos/flower_photos/sunflowers/175638423_058c07afb9.jpg|sunflowers|daisy     |
+-----------------------------------------------------------------------------------------------------------+----------+----------+

25/03/30 08:02:12 WARN KafkaDataConsumer: KafkaDataConsumer is not running in UninterruptibleThread. It may hang when KafkaDataConsumer's methods are interrupted
-------------------------------------------
Batch: 6
-------------------------------------------
+-------------------------------------------------------------------------------------------------------+------+----------+
|path                                                                                                   |label |prediction|
+-------------------------------------------------------------------------------------------------------+------+----------+
|file:/home/chandrakarsatvik/.keras/datasets/flower_photos/flower_photos/tulips/110147301_ad921e2828.jpg|tulips|vase      |
+-------------------------------------------------------------------------------------------------------+------+----------+

25/03/30 08:02:13 WARN KafkaDataConsumer: KafkaDataConsumer is not running in UninterruptibleThread. It may hang when KafkaDataConsumer's methods are interrupted
-------------------------------------------
Batch: 7
-------------------------------------------
+--------------------------------------------------------------------------------------------------------+------+----------+
|path                                                                                                    |label |prediction|
+--------------------------------------------------------------------------------------------------------+------+----------+
|file:/home/chandrakarsatvik/.keras/datasets/flower_photos/flower_photos/tulips/5674125303_953b0ecf38.jpg|tulips|rapeseed  |
+--------------------------------------------------------------------------------------------------------+------+----------+
```

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