**Kafka-Based Instagram Engagement Data Pipeline**

1. Launching EC2 Instance

In AWS, EC2 🡪 Launch instance

Specify a name and download the key value pair

1. Connecting key value pair with EC2

For Windows user, if it fails change the security setting of the file

Key value pair file -> Properties -> Security -> Advanced -> Advanced Security Settings -> Disable inheritance

In the prompt, select Remove all inherited permissions.

Click Add, then Select a principal.

Type your Windows username and click Check Names (it should underline your name).

Click OK.

In the "Permissions" window, check Read (this should be enough for SSH).

Click OK to save.

1. Open cmd from the place where the key value pair file is saved.

ssh -i "kafka-stock-market-project.pem" [ec2-user@ec2-54-205-28-228.compute-1.amazonaws.com](mailto:ec2-user@ec2-54-205-28-228.compute-1.amazonaws.com)

if you click on connect on the EC2 instance -> SSH Client   
you can find the command to connect to EC2

1. Download Kafka

wget <https://dlcdn.apache.org/kafka/3.9.0/kafka_2.12-3.9.0.tgz>  
check for the latest verions (2.12-3.9.0- latest version as of March 13th 2025)

1. Uncompress the downloaded file

tar -xvf kafka\_2.12-3.9.0.tgz

1. Install Java (check the compatible version with Kafka)

sudo dnf install java-17-amazon-corretto -y

1. cd kafka\_2.12-3.9.0

It is pointing to private server , change server.properties so that it can run in public IP

Stop both server, sudo nano config/server.properties  
In the line “advertised.listener=PLAINTEXT://your.host.name:9092”  
change your.host.name to the public IPv4 address 3.90.104.168  
You can find the IPv4 address in the Instance.

1. Starting the zookeeper

Navigate to the kafka folder

cd kafka\_2.12-3.9.0

bin/zookeeper-server-start.sh config/zookeeper.properties

1. Setup additional space in Kafka.

Open new cmd and run the EC2 instance

set KAFKA\_HEAP\_OPTS=-Xmx256M -Xms128M

1. Start the Kafka in the same cmd where additional heap was given

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

1. Now start both the server
2. In AWS, click on instance -> Security -> Security Group

Click on the security group number -> Edit Inbound Rules

Add rule -> Type: All Traffic -> Source: My IP

1. Create Topics

bin/kafka-topics.sh --create --topic demo\_test --bootstrap-server 3.86.184.36:9092 --replication-factor 1 --partitions 1

demo\_test -> topic name

3.86.184.36-> Public IP address of the EC2 instance

In the same window create producer

bin/kafka-console-producer.sh --topic demo\_test --bootstrap-server 3.86.184.36:9092

1. Create Consumer

bin/kafka-console-consumer.sh --topic demo\_test --bootstrap-server 3.86.184.36:9092

what ever you type in the producer, you can view them in real time in consumer.

1. In the jupyter notebook, create files kafkaProducer and kafkaConsumer
2. If Port is already in use

sudo netstat -tulnp | grep 2181 -> check it the port 2181 is in use

sudo kill -9 <PID>

1. Create a bucket in S3 with Programmatic access
2. Create User

IAM 🡪 User 🡪 Create User 🡪 Give User name 🡪 Attach policies directly 🡪 Give AdministratorAccess 🡪 Next 🡪 Create User

Then create access key with local root and download the access keys.

1. Download AWS CLI

In cmd prompt, type “aws configure” and enter the downloaded access key and password.

1. from s3fs import S3FileSystem  
   this import will help to send file to S3

Producer:  
while True:

dict\_stock = df.sample(1).to\_dict(orient="records")[0]

producer.send('demo\_test', value=dict\_stock)

sleep(1)

Consumer:  
for count, i in enumerate(consumer):

with s3.open("s3://stock-market-analysis-with-kafka-aishwarya/stock\_market\_{}.json".format(count), 'w') as file:

json.dump(i.value, file)

This is send the json objects to the S3 buckets

1. Create IAM Role

IAM 🡪 Create Role 🡪 Specify Use case as “Glue” 🡪 Next 🡪 Give “AdministratorAccess” 🡪 Give role name 🡪 Create Role

1. Create Crawler

AWS Glue 🡪 Crawlers 🡪 Create Crawler 🡪 Specify Name 🡪 Add Data Source 🡪 Add IAM Role 🡪 Next 🡪 Create Database 🡪 Create Crawler

Run the Crawler.

1. Go to “Athena” 🡪 Select stock\_market\_kafka 🡪 Click the three dots on table and Preview table to see the table there.
2. In the python code, if you run the producer and consumer, the files will be added automatically added to the S3 continuously. And you refresh the athena, the rows will keep increasing continuously.