Setting up - MSK cluster (Managed Streaming Kafka Cluster):

MSK enables us to build and run applications that use Apache Kafka to process streaming data.

After creating cluster, 2 Kafka brokers will be created.
Keep the url for each broker, which will be used to setup the during kafka queues.

Setting up - Kafka Cluster and Queues

#EC2 - Kafka Client
Spin a Amazon Linux 2 Instance

#Install Java
sudo yum update -y
sudo yum install java-1.8.0-amazon-corretto -y
java -version

#Install Kafka
Kafka version should match with MSK cluster
Ref: https://kafka.apache.org/downloads

wget https://archive.apache.org/dist/kafka/3.5.1/kafka_2.12-3.5.1.tgz
tar -xzf kafka_2.12-3.5.1.tgz
cd kafka_2.12-3.5.1

#Create Topics
bin/kafka-topics.sh --create --topic <topic-name> --bootstrap-server
<broker1-url>:9092,<broker-2 url>:9092 --replication-factor 1 --partitions 1

#Start Producer
bin/kafka-console-producer.sh --topic <topic-name> --bootstrap-server
<broker1-url>:9092,<broker-2 url>:9092

#Start Consumer (Open duplicate putty/ssh session)
bin/kafka-console-consumer.sh --topic <topic-name>--bootstrap-server
<broker1-url>:9092,<broker-2 url>:9092

#Install Confluent (Use it to publish to Kafka Cluster)

Download in Kafka directory
cd kafka_2.12-3.5.1
wget http://packages.confluent.io/archive/5.1/confluent-5.1.2-2.11.zip
unzip confluent-5.1.2-2.11.zip

export CONFLUENT_HOME=/kafka_2.12-3.5.1/confluent-5.1.2
export PATH=$PATH:$CONFLUENT_HOME/bin

#Make changes to confluent
cd /home/ec2-user/kafka_2.12-3.5.1/confluent-5.1.2/etc/kafka-rest
vi kafka-rest.properties

Edit the below property in kafka-rest.properties:
bootstrap.servers=PLAINTEXT://localhost:9092

Add Bootstrap Server URLs:
bootstrap.servers=PLAINTEXT://<broker-1 url>:9092,PLAINTEXT://<broker-2 url>:9092

#Start Kafka Rest Start
/kafka_2.12-3.5.1/confluent-5.1.2/bin/kafka-rest-start
/kafka_2.12-3.5.1/confluent-5.1.2/etc/kafka-rest/kafka-rest.properties

#Access Kafka REST API
http://<ec2 ip address>:8082/topics/<topic-name>

Setting up - MSK Connector

Create an IAM role which will have access to S3 - this will be used to publish Kafka messages to S3 bucket

Create S3 bucket, in the same regioin in which Kafka cluster has been setup

Create a MSK Connector, linking the Kafka Consumer queue to S3 bucket.
Every message in Consumer queue will be dumped in the S3 bucket

Setting up - DynamoDB for storing the final results

Create a DynamoDB - key,value database which will store the output generated from ML model for each user query.