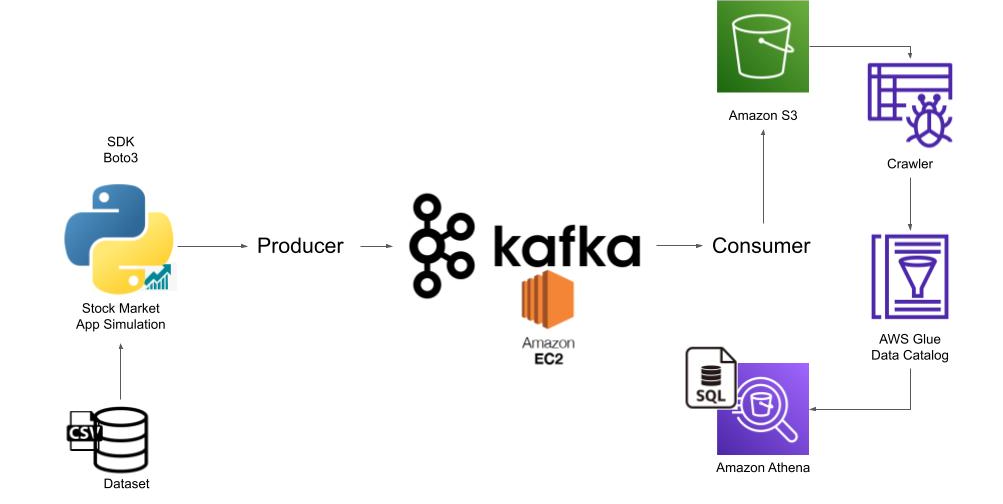
**AWS Data Pipeline Project 2**

**Overview of the Project:**

End-To-End Data Engineering Project on Data using Kafka where we will install necessary Kafka and zookeeper services in AWS EC2 instance and we will read the data using producer code written in python and with the help of consumer we will dump the data into S3 Bucket. And, we will use AWS Glue Crawler to extract the metadata information of the data stored in s3 and we will query the data using AWS Athena.

**Architecture:**



Main Diagram

## AWS and other services used:

## 1.AWS EC2

## 2.S3

## 3.AWS Glue Crawler

## 4.Apache Kafka

## 5.AWS Athena

## 6.IAM Roles.

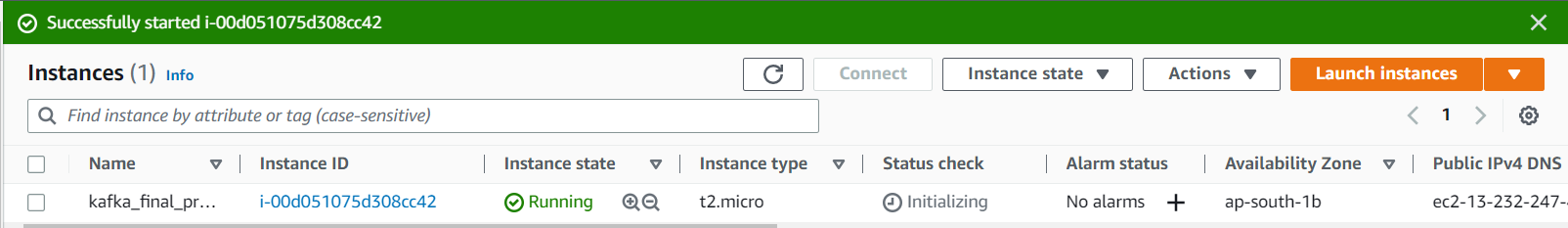
## Prerequisites:

## Requires and AWS account with appropriate AWS Identity and IAM permissions and necessary packages like java, kafka and zookeeper installed on EC2.

**Deploying Solution:**

**Step1 :**

1.Create a EC2 instance in AWS.



2.connect to EC2 using ssh.

Example:

ssh -i "file\_name.pem" demo-user@ec2-13-232-247-48.ap-south-1.compute.amazonaws.com

**Step2 :**

1. change ADVERTISED\_LISTENERS to public ip of the EC2 instance in config/server.properties so that we can acess the EC2 instance in local.

**Command**: sudo nano config/server.properties

2.Run the Zookeeper on EC2:

**Command**: bin/zookeeper-server-start.sh config/zookeeper.properties



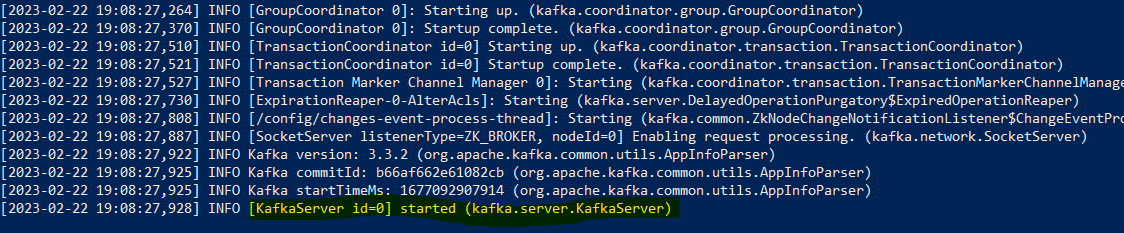
**Step3:**

1.Set heap memory before running Kafka server.

**Command**: export KAFKA\_HEAP\_OPTS="-Xmx256M -Xms128M"

2.Run Kafka:

**Command**: bin/kafka-server-start.sh config/server.properties



3.Create Topic

**Command**:

bin/kafka-topics.sh --create --topic demo\_testing3 --bootstrap-server <public\_ip\_of\_EC2>:9092 --replication-factor 1 --partitions 1

4.Start Producer:

**Command:**

bin/kafka-console-consumer.sh --topic demo\_testing3 --bootstrap-server <public\_ip\_of\_EC2>:9092

5.Start Consumer:

**Command:**

bin/kafka-console-producer.sh --topic demo\_testing3 --bootstrap-server <public\_ip\_of\_EC2>:9092

**Step4:**

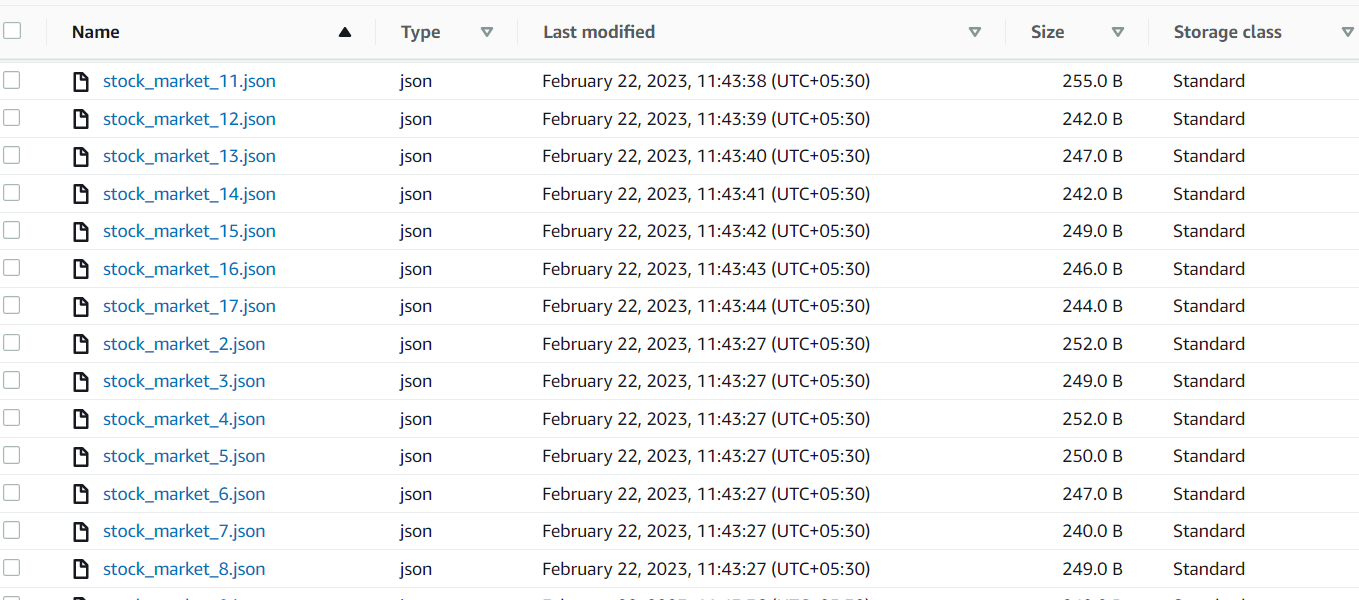
1.Write Producer.py code where the data will be read from Dataset or API and dumped into Kafka Topic.



2.Write Consumer.py code where the data read from topic will be dumped to S3 Bucket.

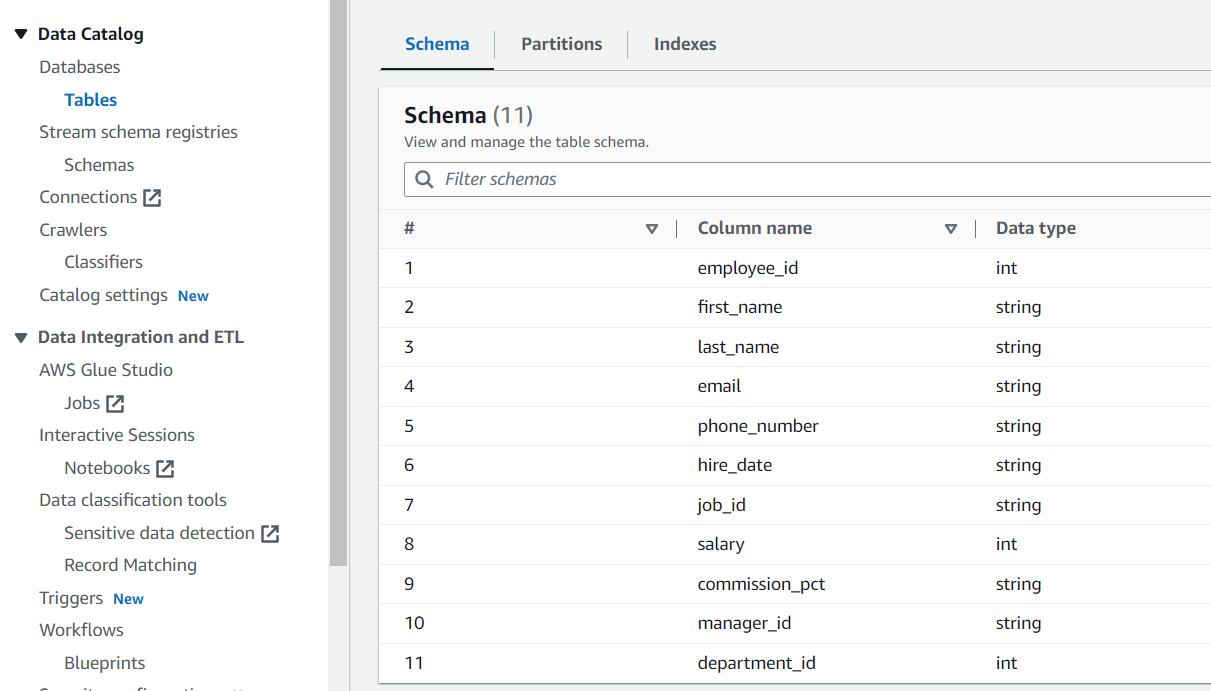


3.Data stored in S3.



**Step5:**

Create Crawler to extract metadata information from data stored in S3.



**Step6:**

Using Crawler read the data in AWS Athena and query on top of it to perform SQL operations.

