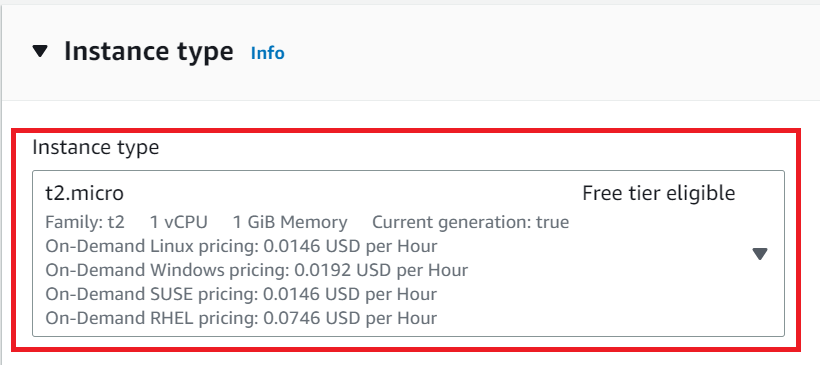
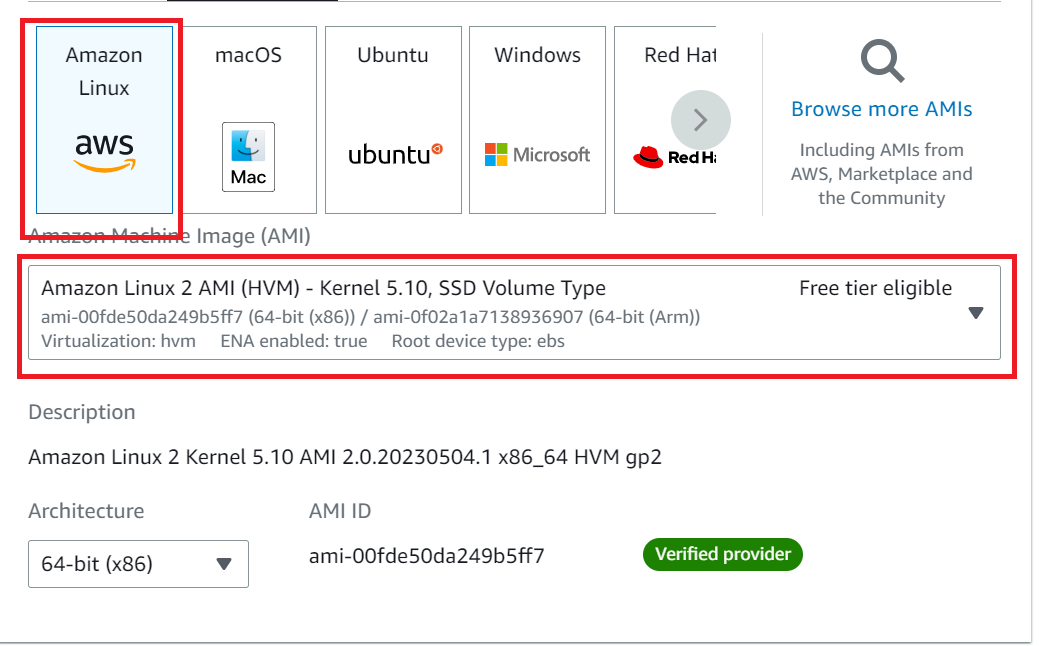
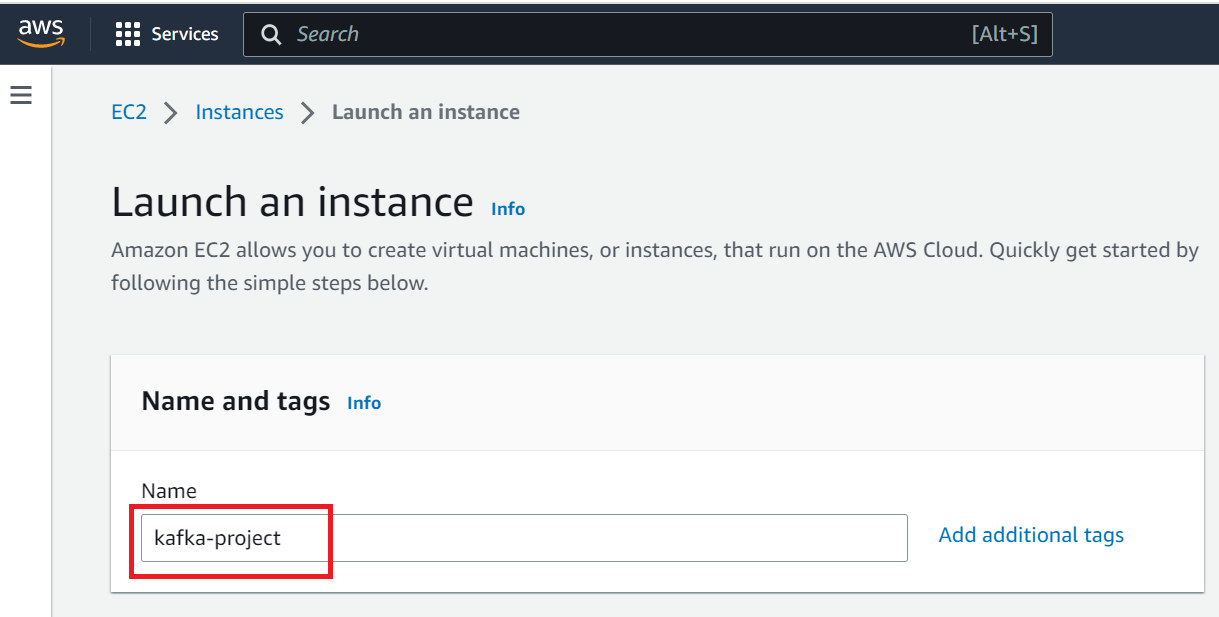
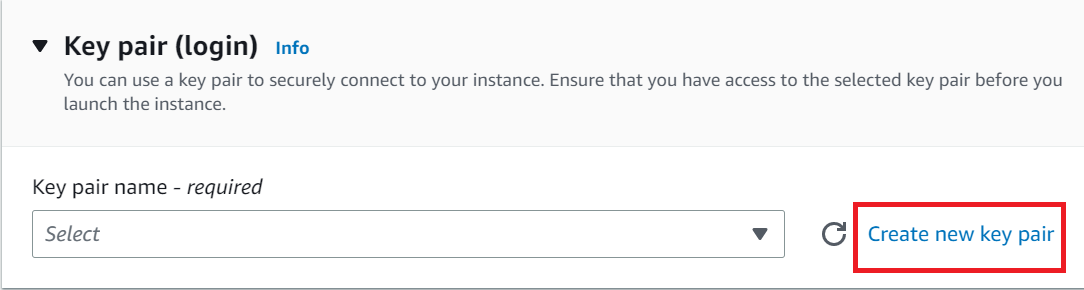
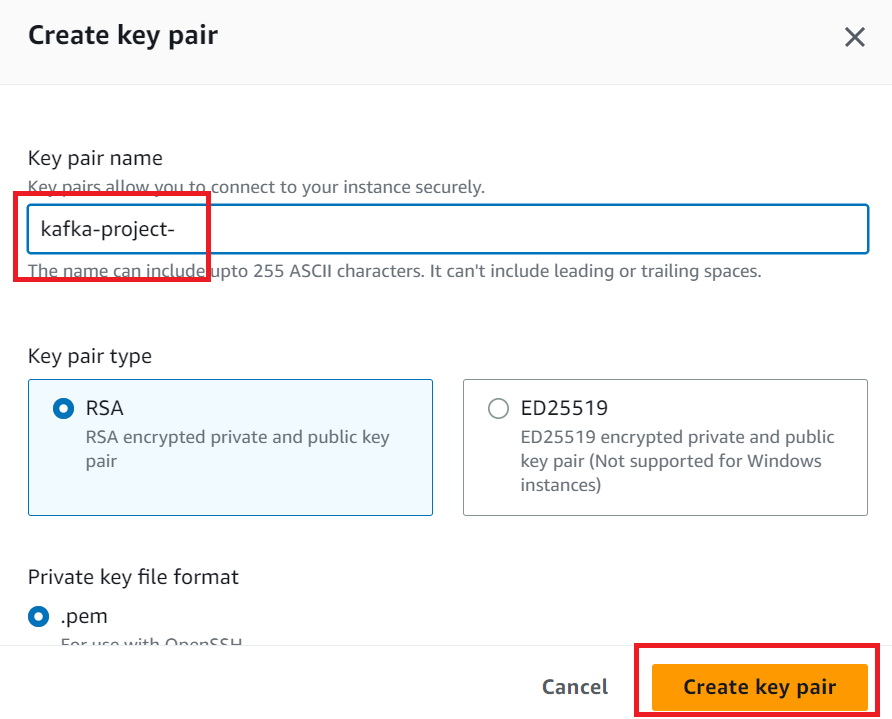
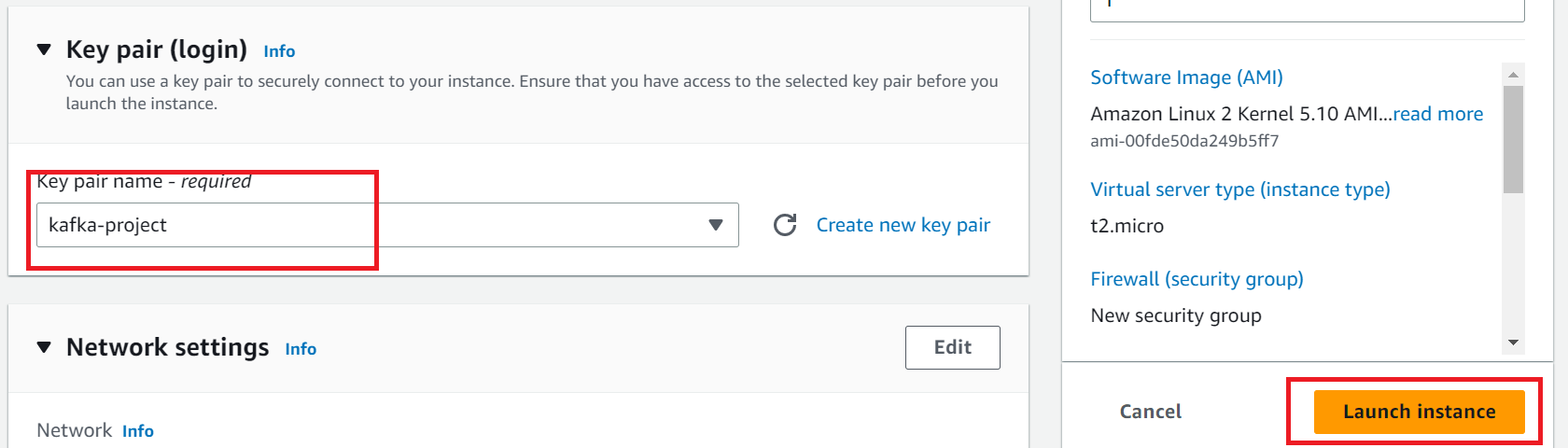
Launch instance EC2

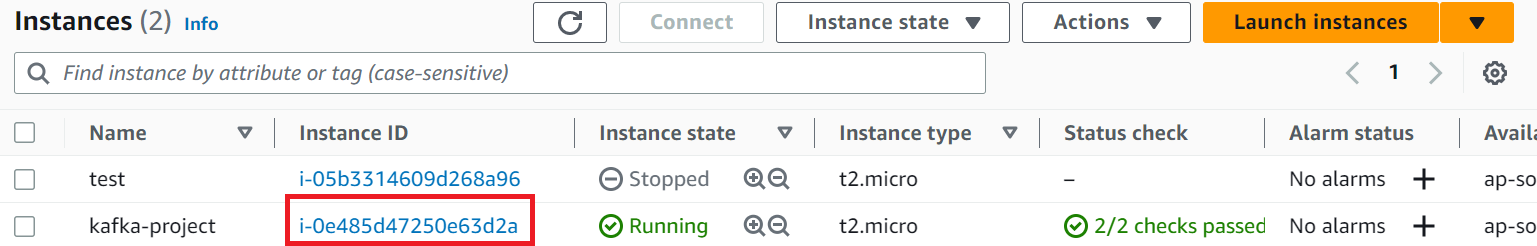
Create new key pair. This key will be using to connect the ec2 machine from our local computer.

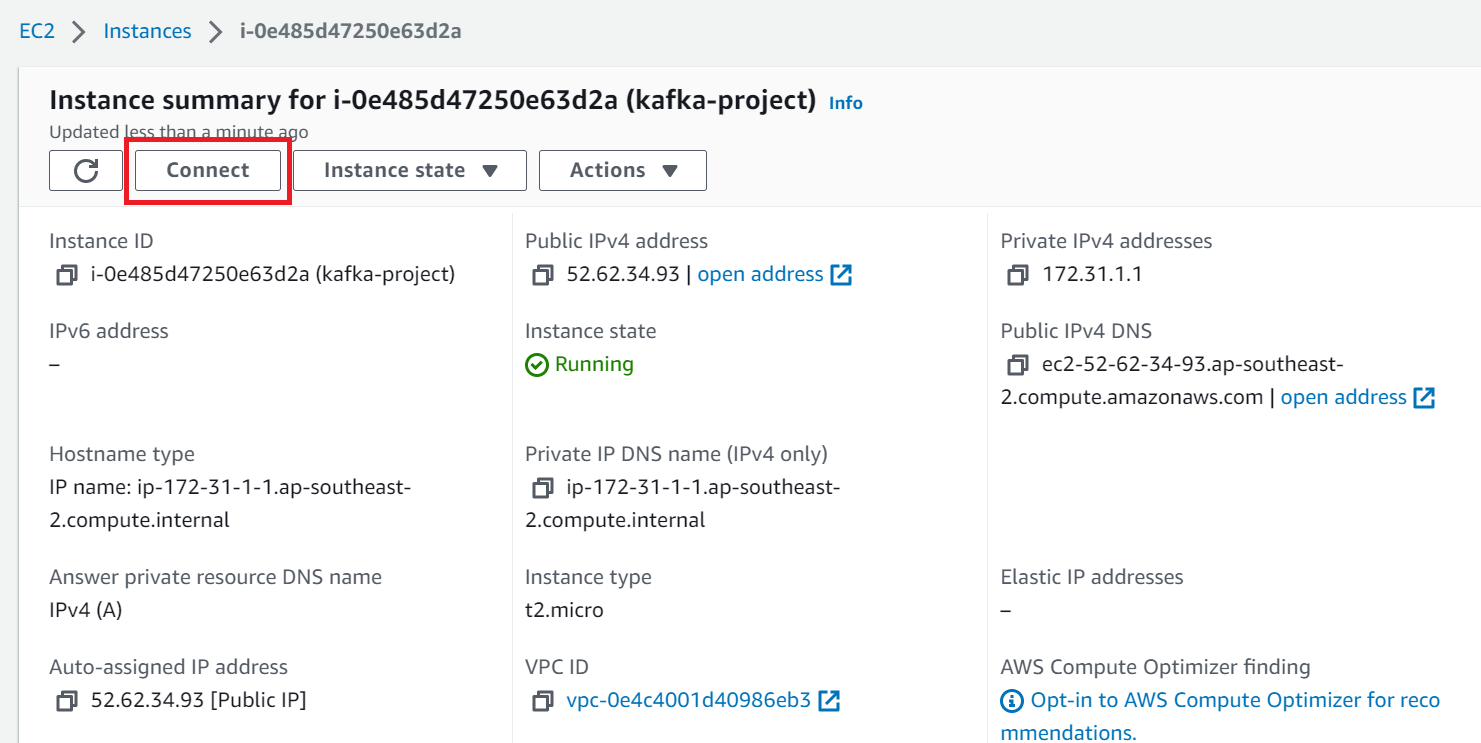


Click on create key pair and it will download automatically

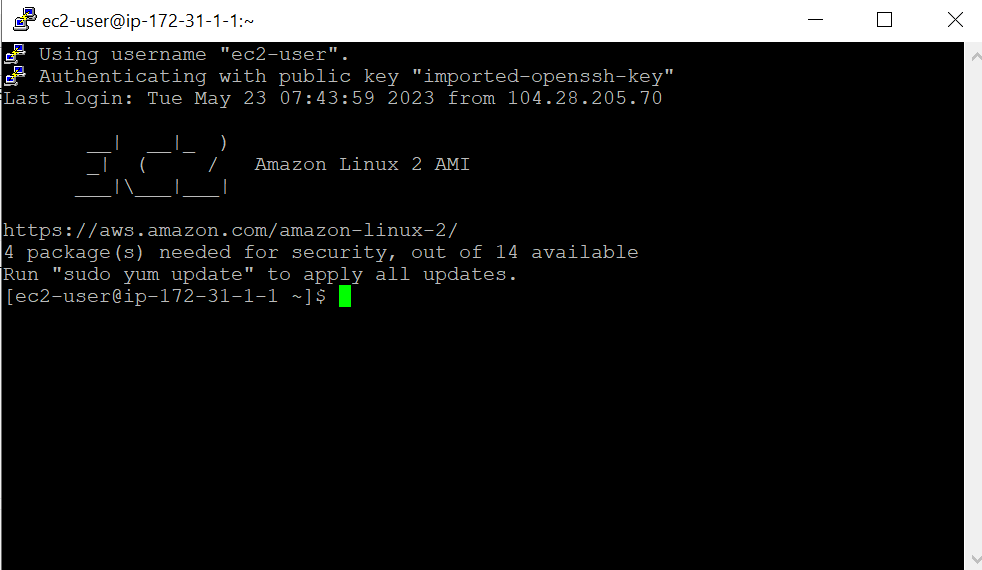
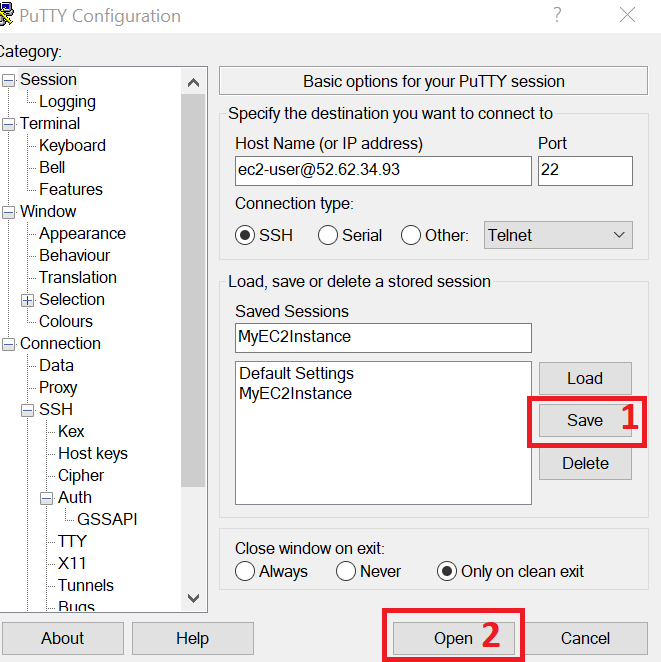
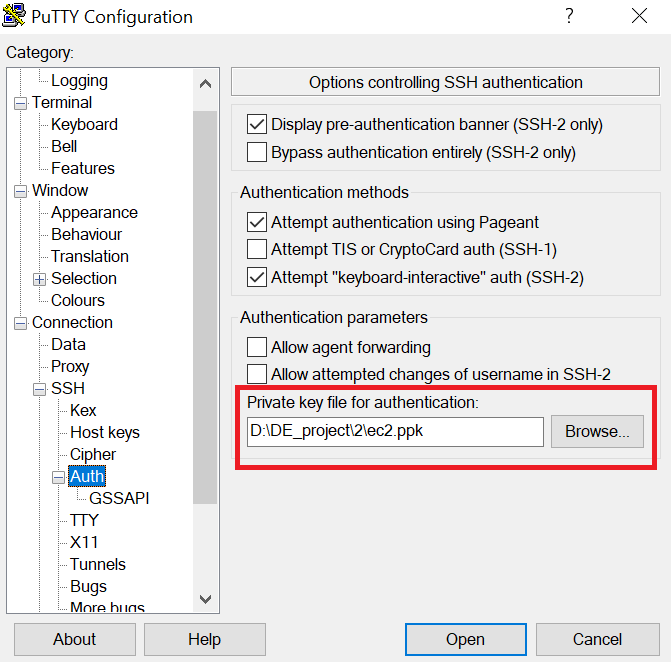
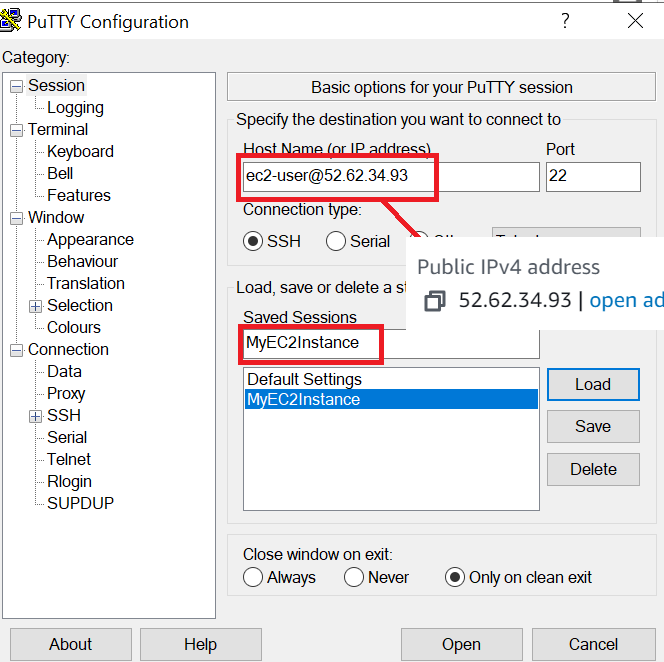
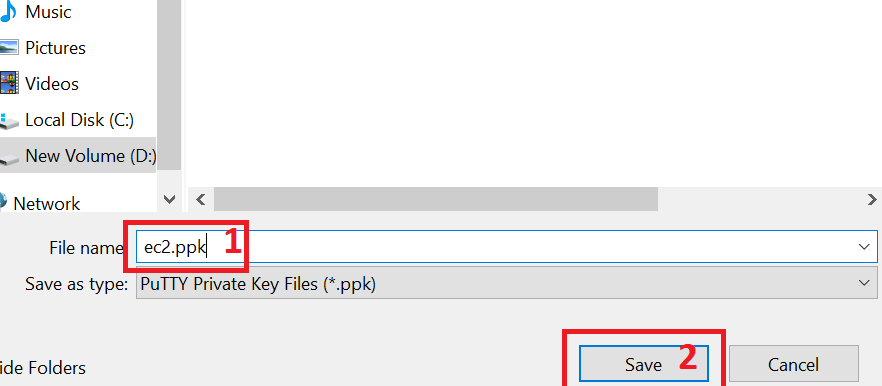
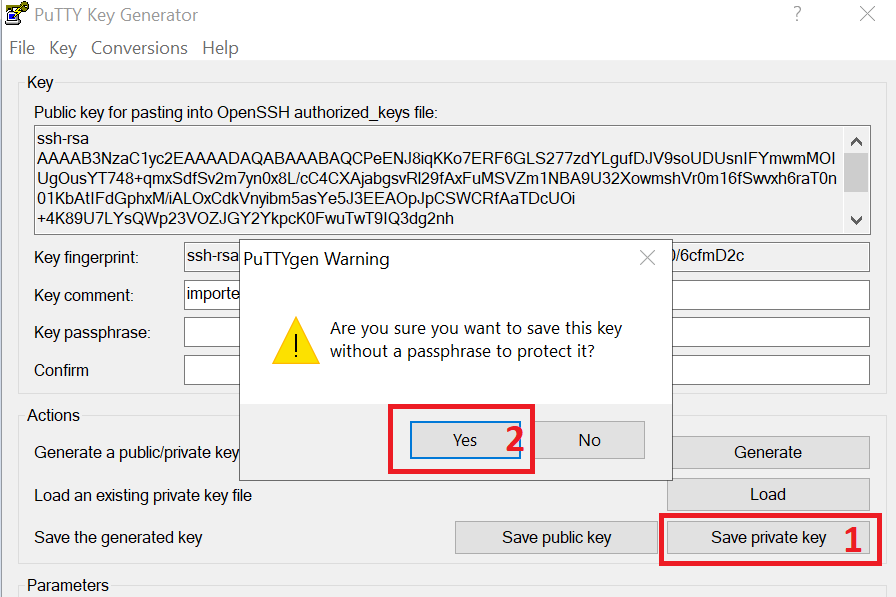
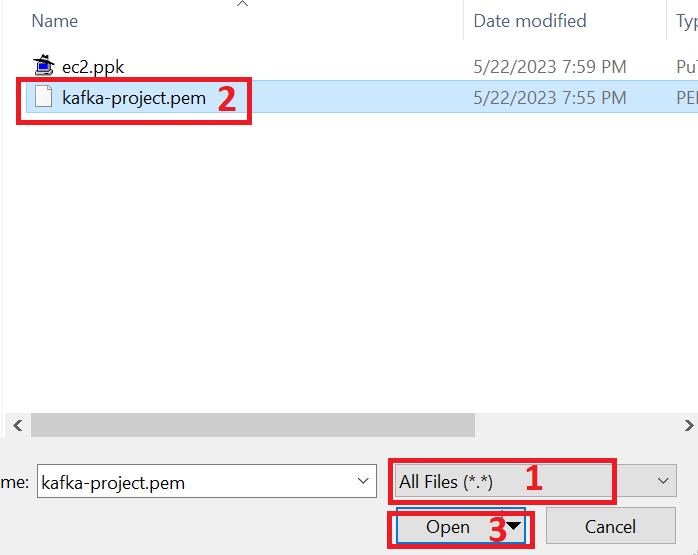
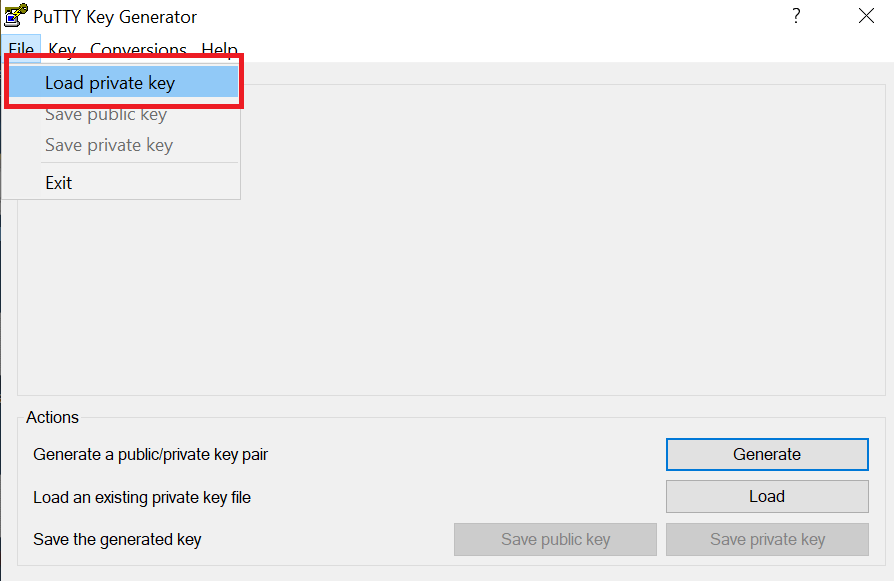








Connect to your EC2 instance

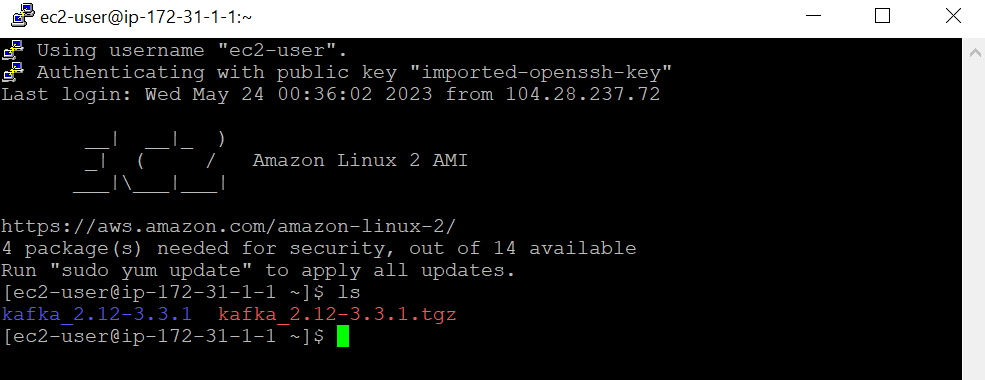


Install Kafka:

-----------------------

wget https://downloads.apache.org/kafka/3.3.1/kafka\_2.12-3.3.1.tgz

tar -xvf kafka\_2.12-3.3.1.tgz



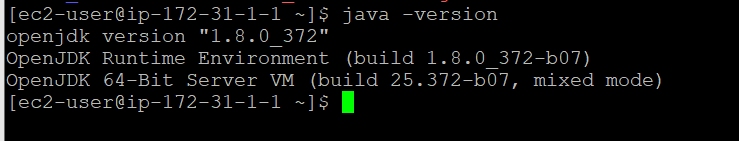
Install Java:

-----------------------

java -version

sudo yum install java-1.8.0-openjdk

java -version

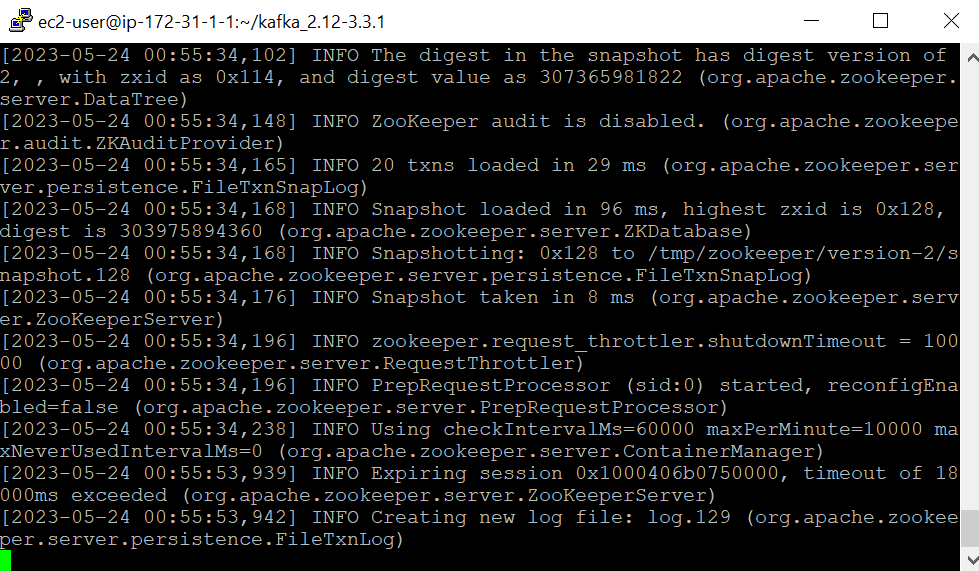


Start Zoo-keeper:

-------------------------------

cd kafka\_2.12-3.3.1

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



Open another window to start kafka

But first ssh to to your ec2 machine as done above

Start Kafka-server:

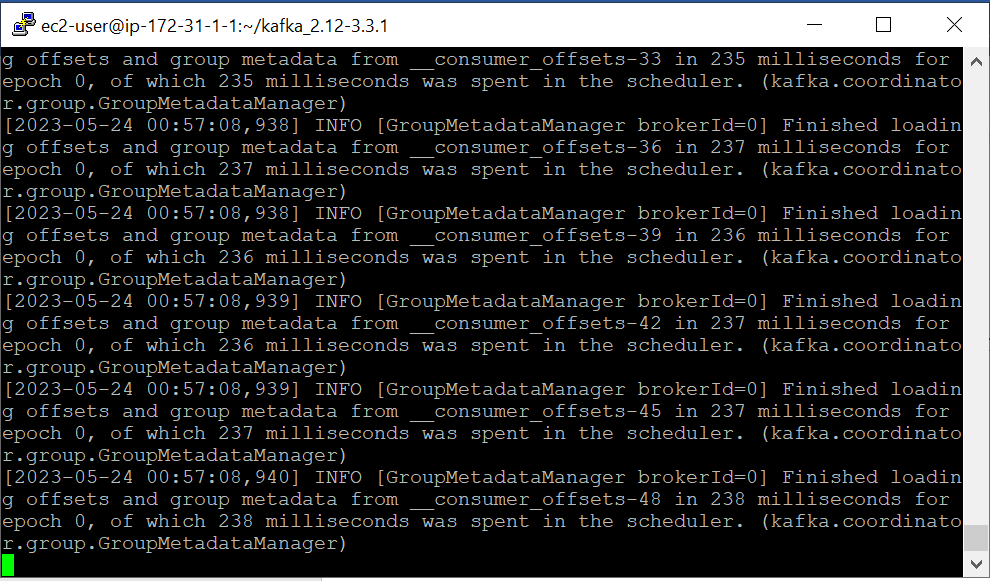
----------------------------------------

Duplicate the session & enter in a new console --

export KAFKA\_HEAP\_OPTS="-Xmx256M -Xms128M"

cd kafka\_2.12-3.3.1

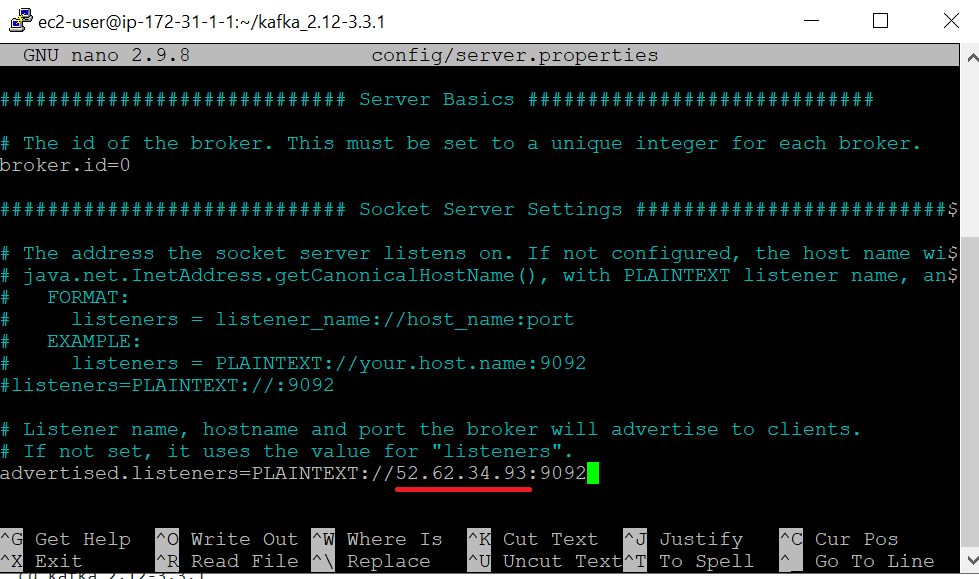
bin/kafka-server-start.sh config/server.properties



It is pointing to private server (you cannot access your private DNS from your local computer unless you are in the same network), so we need to change server properties so that it can run in public IP (so that we can access our ec2 machine outside of the network).

First, stop both of the sever (Ctrl Z -> Ctrl C)

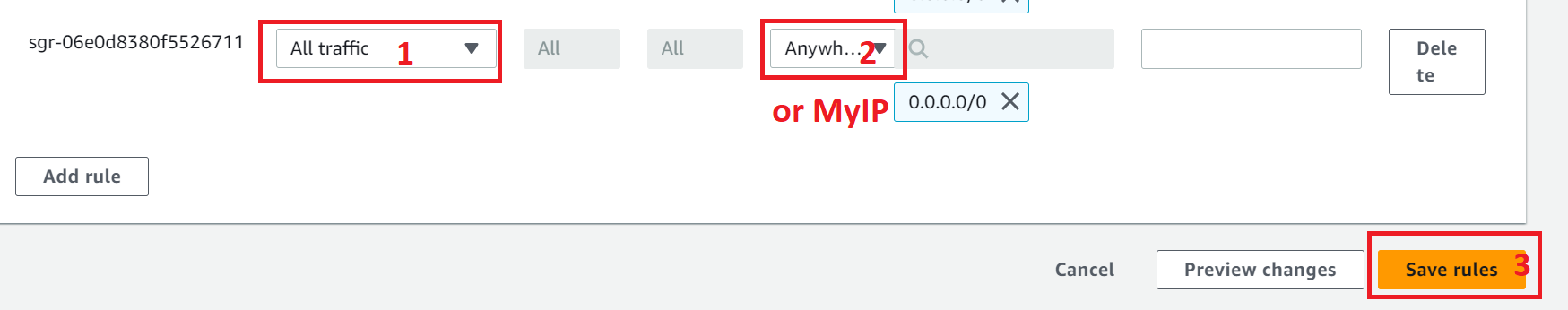
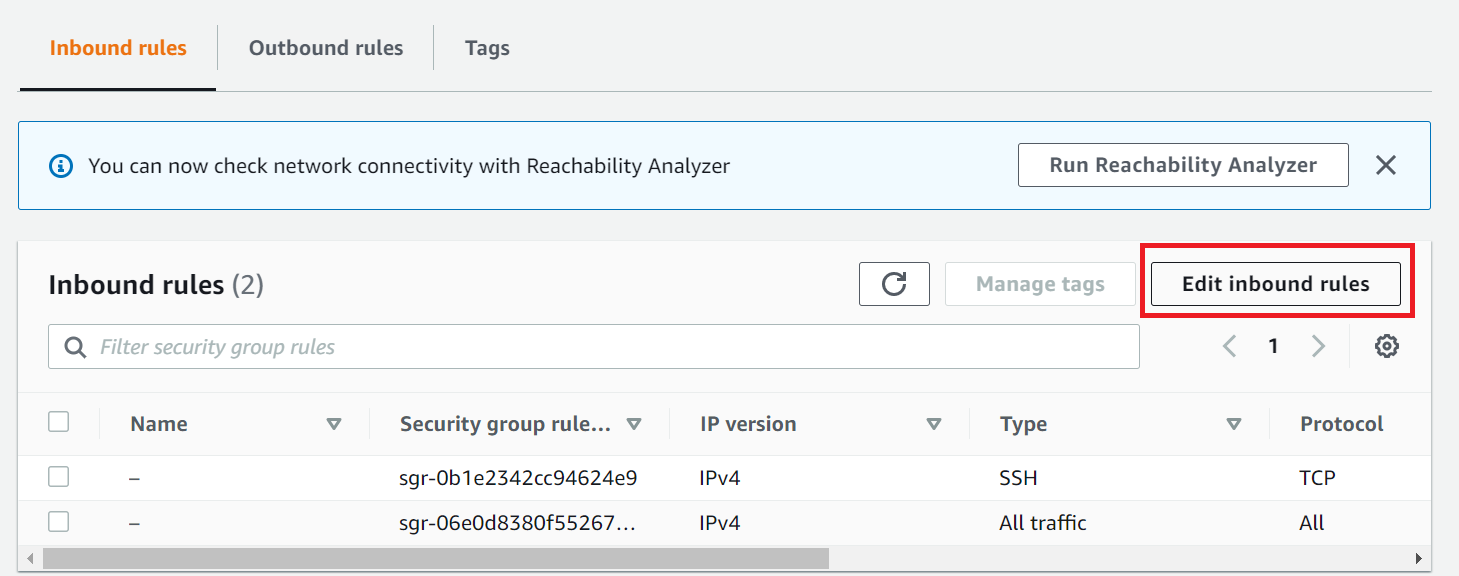
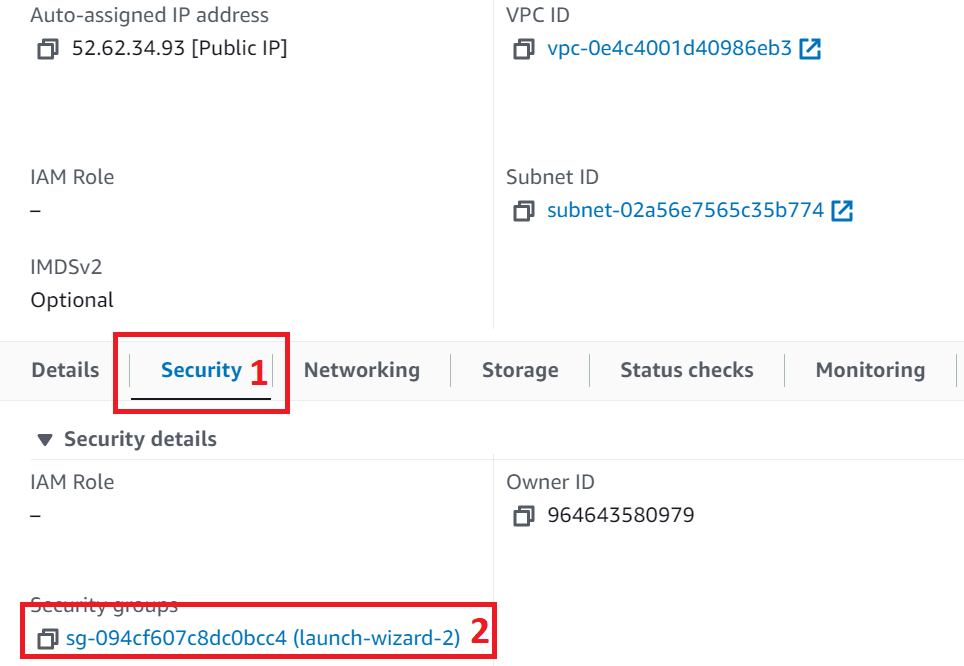
Do a "sudo nano config/server.properties" - change ADVERTISED\_LISTENERS to public ip of the EC2 instance



Ctrl X –> Y –> Enter (Save)

Then start Zoo-keeper and Kafka-server

Provide the security access from our local machine



The thing we just did is like edited the inbound rules of the security group is not the best practice. You should never allow all the traffic from any where in the world.

Create the topic:

-----------------------------

Duplicate the session & enter in a new console --

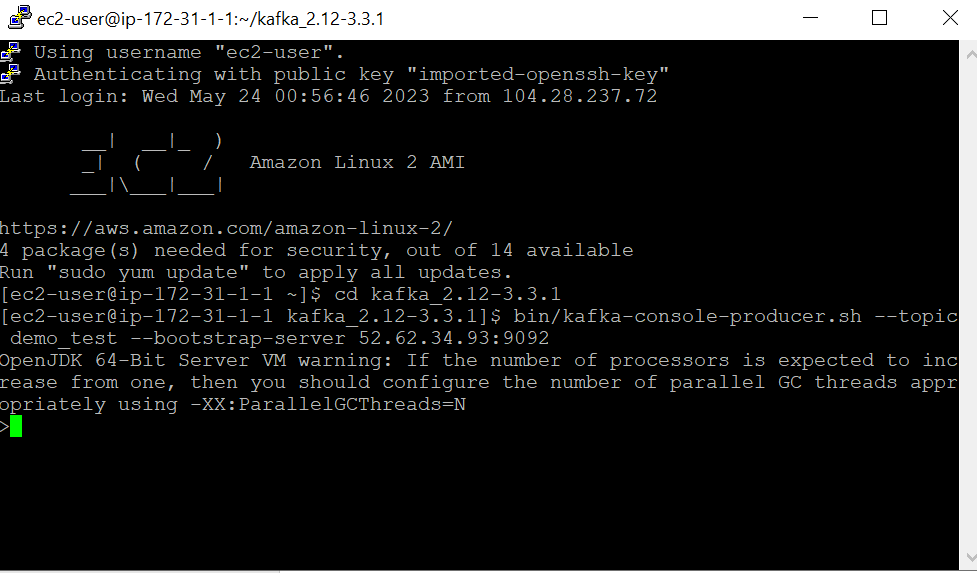
cd kafka\_2.12-3.3.1

bin/kafka-topics.sh --create --topic demo\_test --bootstrap-server 52.62.34.93:9092 --replication-factor 1 --partitions 1

Start Producer:

--------------------------

bin/kafka-console-producer.sh --topic demo\_test --bootstrap-server 52.62.34.93:9092



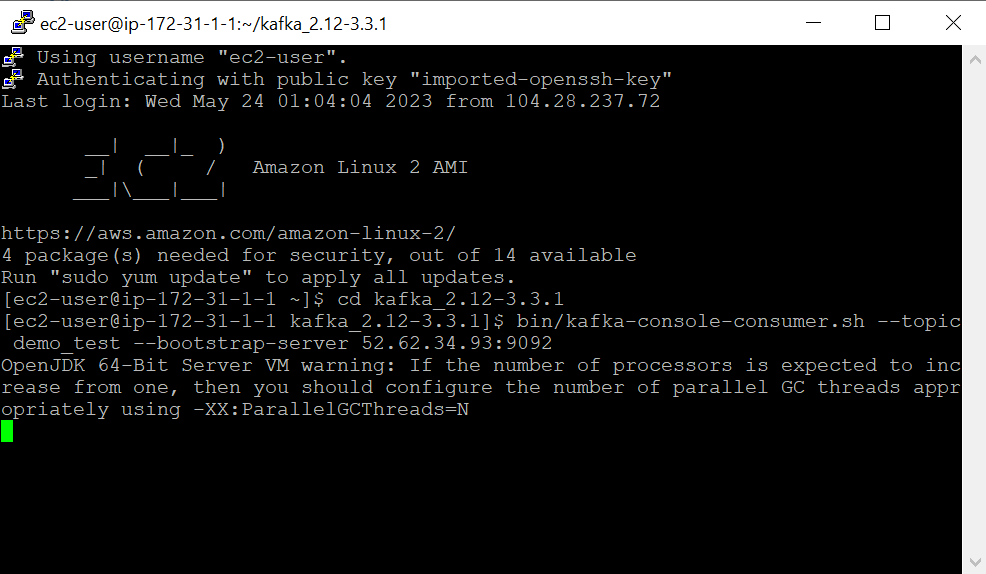
Start Consumer:

-------------------------

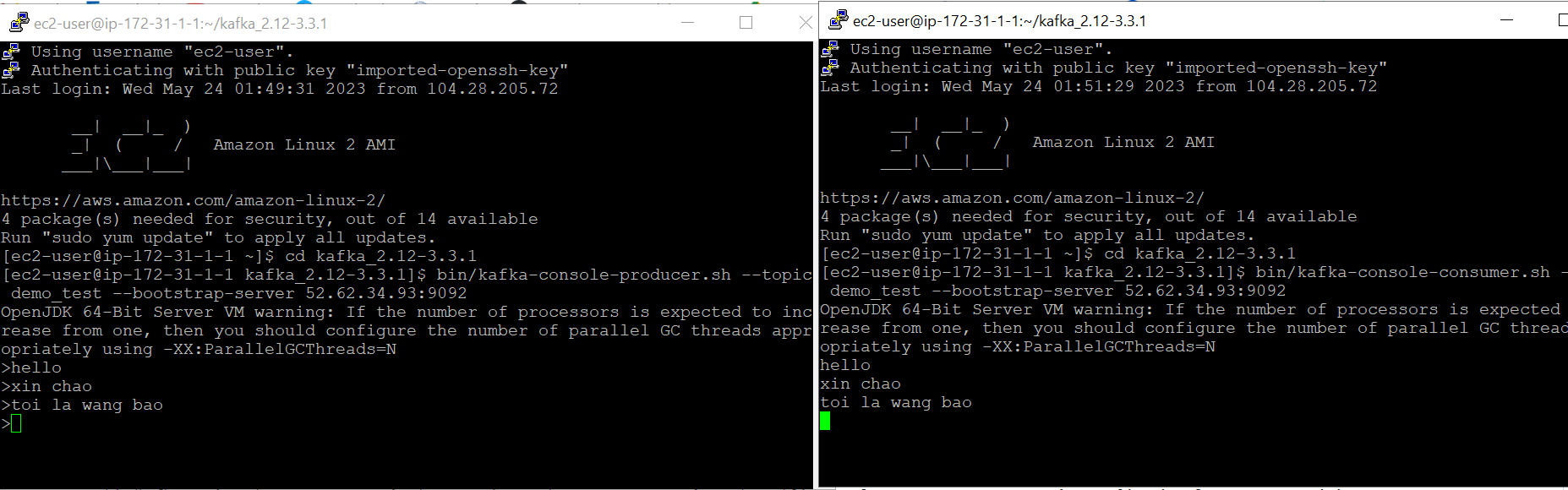
Duplicate the session & enter in a new console --

cd kafka\_2.12-3.3.1

bin/kafka-console-consumer.sh --topic demo\_test --bootstrap-server 52.62.34.93:9092

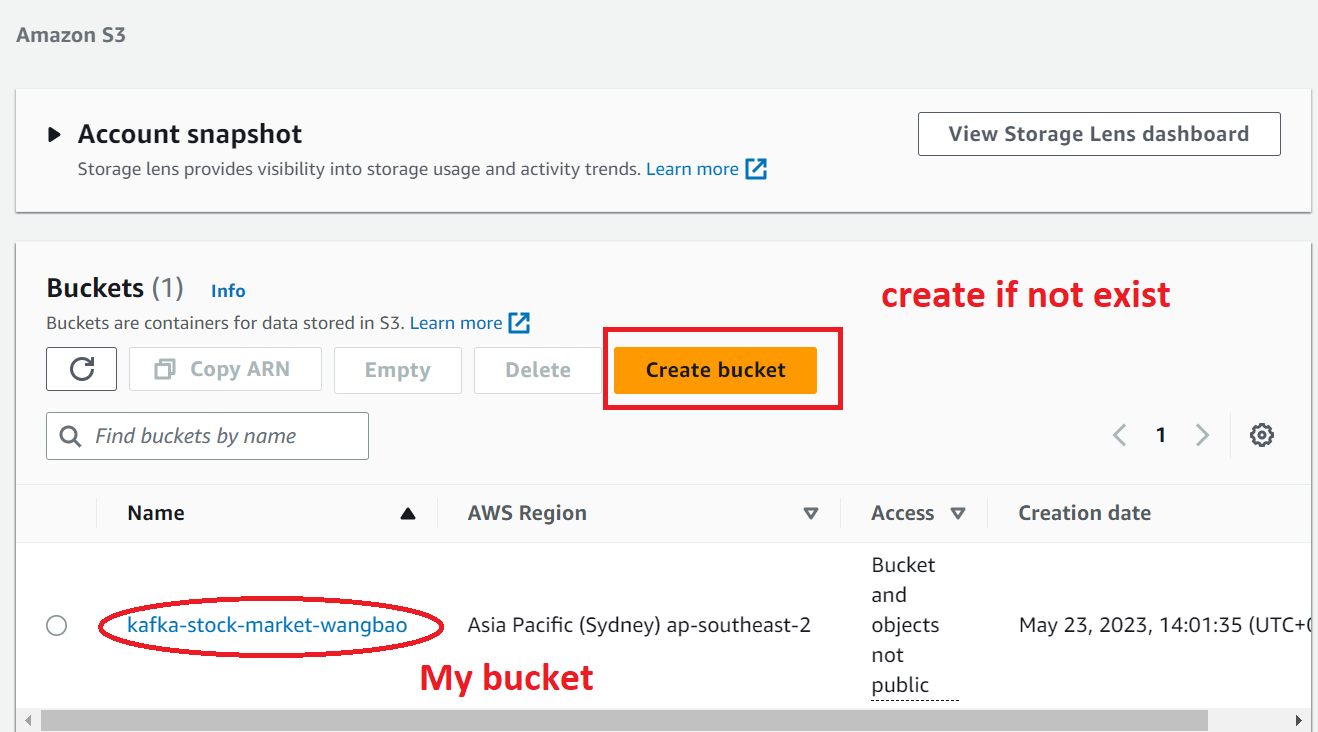


You will see we are producing data from a producer and then we are able to get that data onto the comsumer



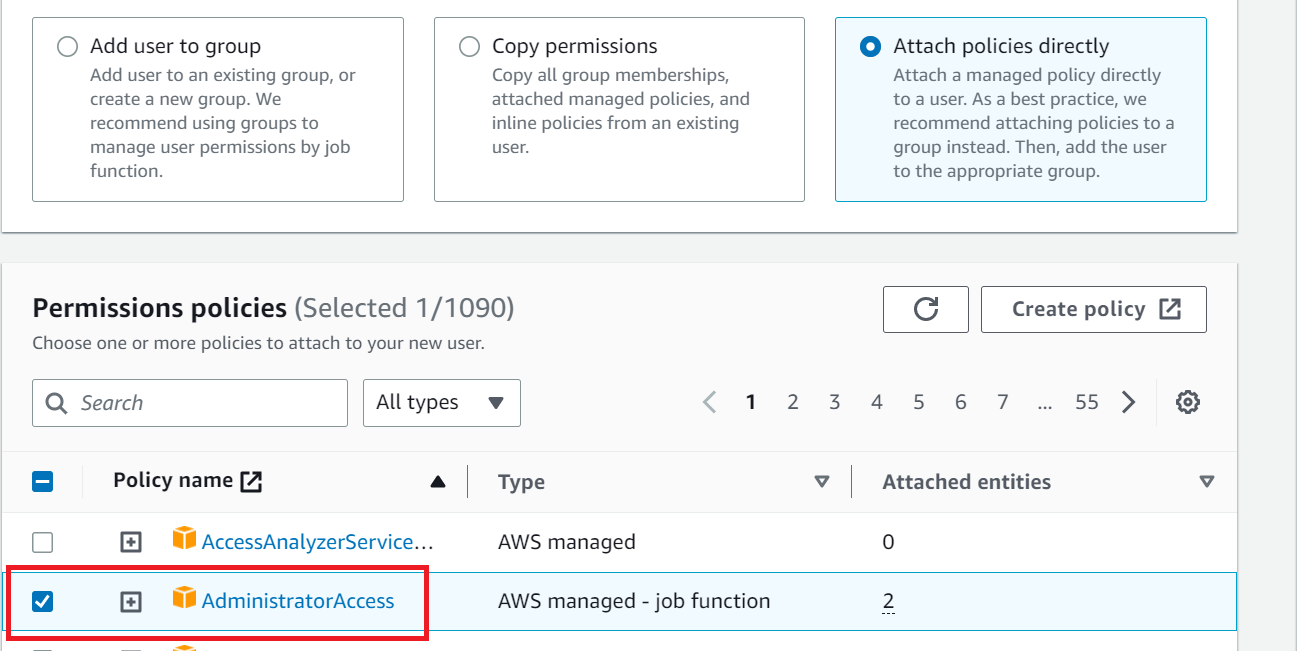
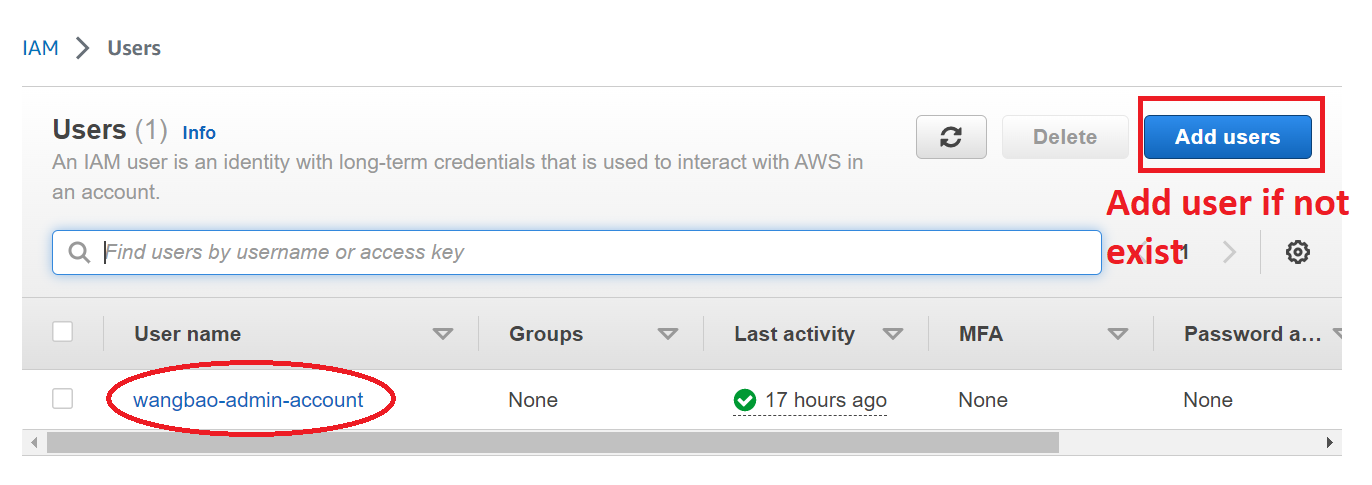
Amazon S3

Create bucket (an object storage you can store whatever the type of file you want)



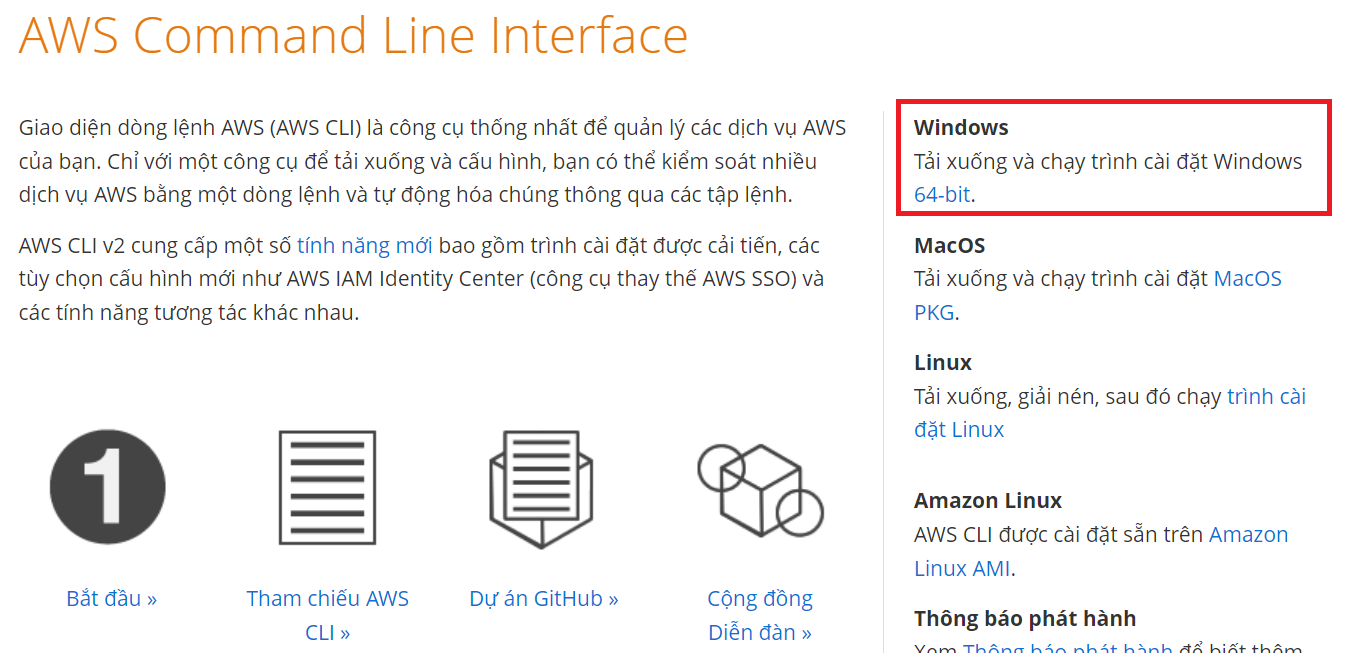
IAM

When you create your AWS account you might have got the AWS secret can access key but if you don’t have it or if you forgot about it. You can go to user click onto add user

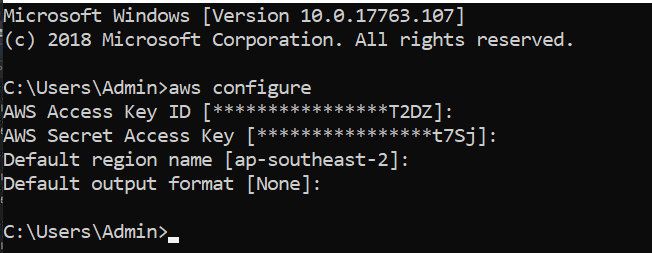


Create Access key -> Dowload csv file (this file will have my access key ID and the secret key ID)

Dowload AWS CLI



AWS account set up on your local machine



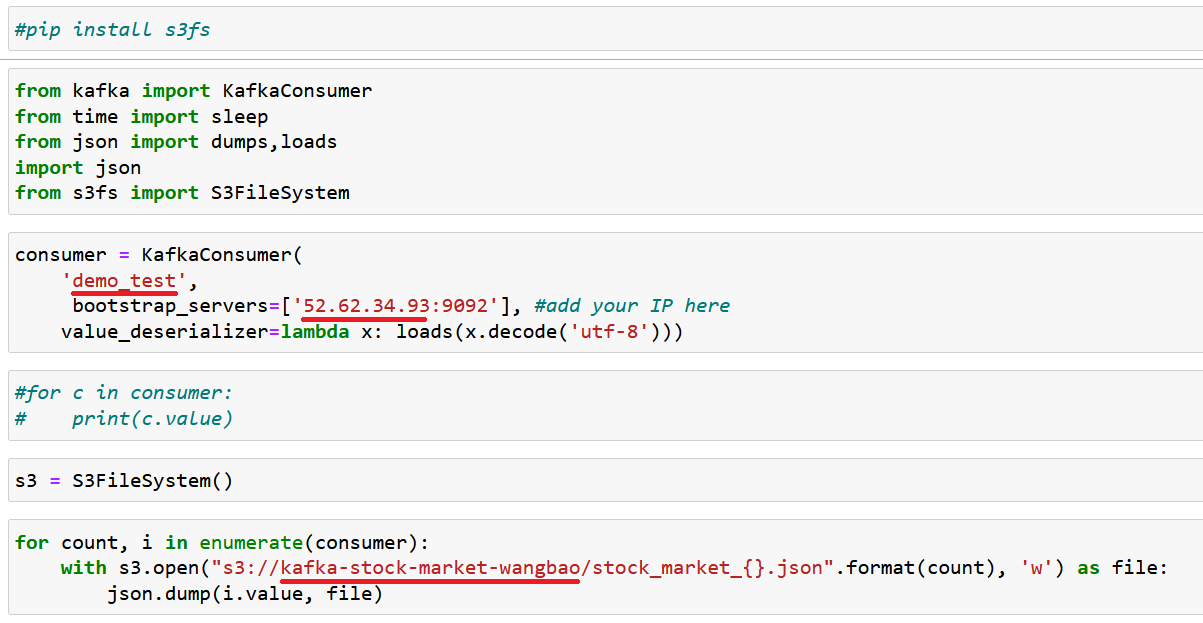
So once you do that then you will be able to send data from your local machine to S3

Jupyter

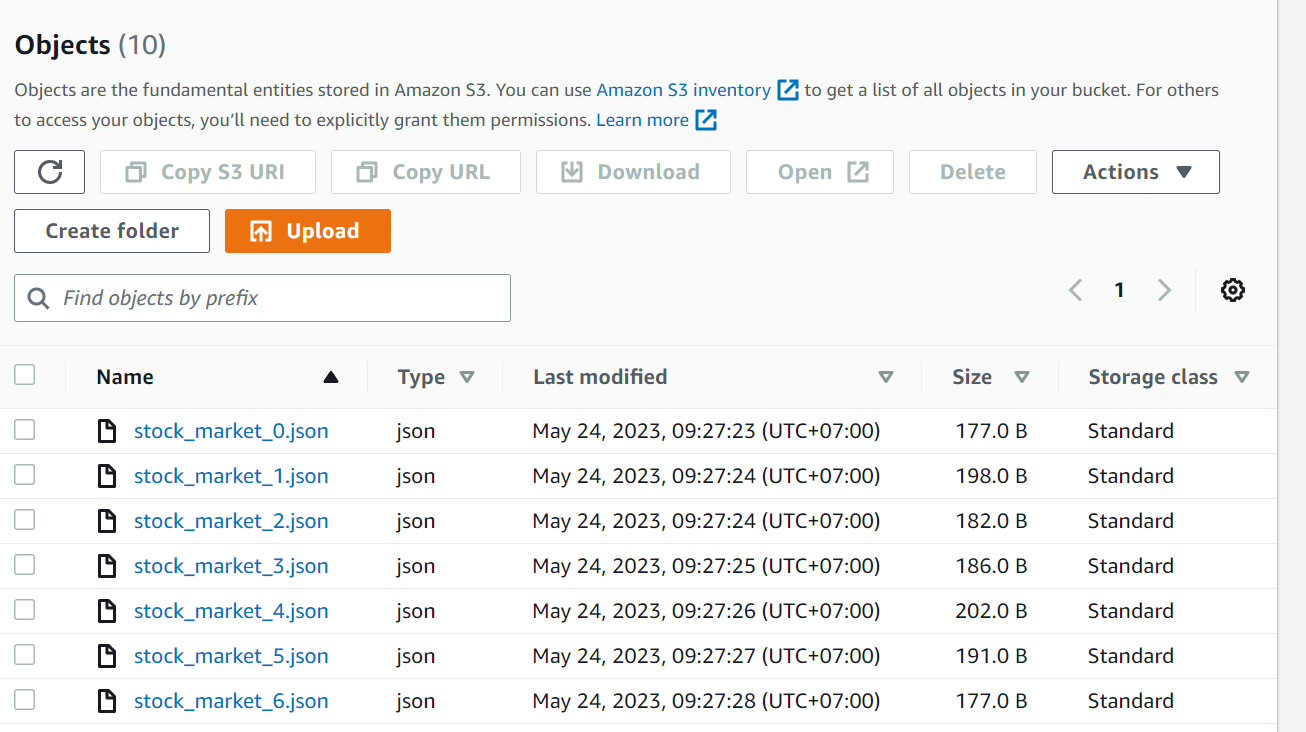
Start producer



Start consumer



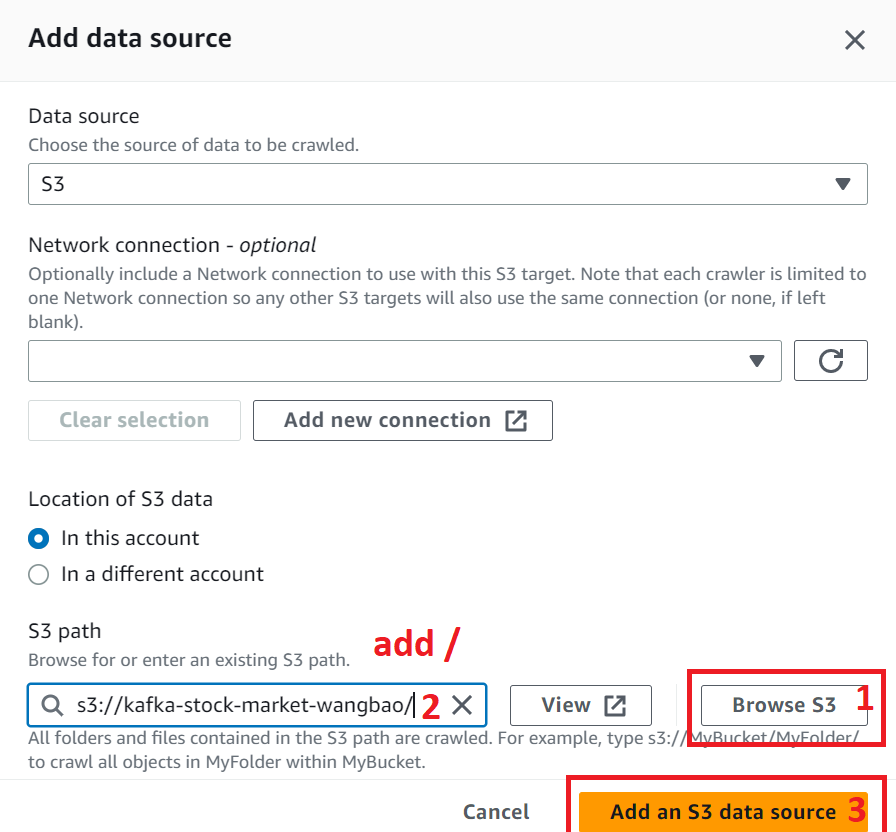
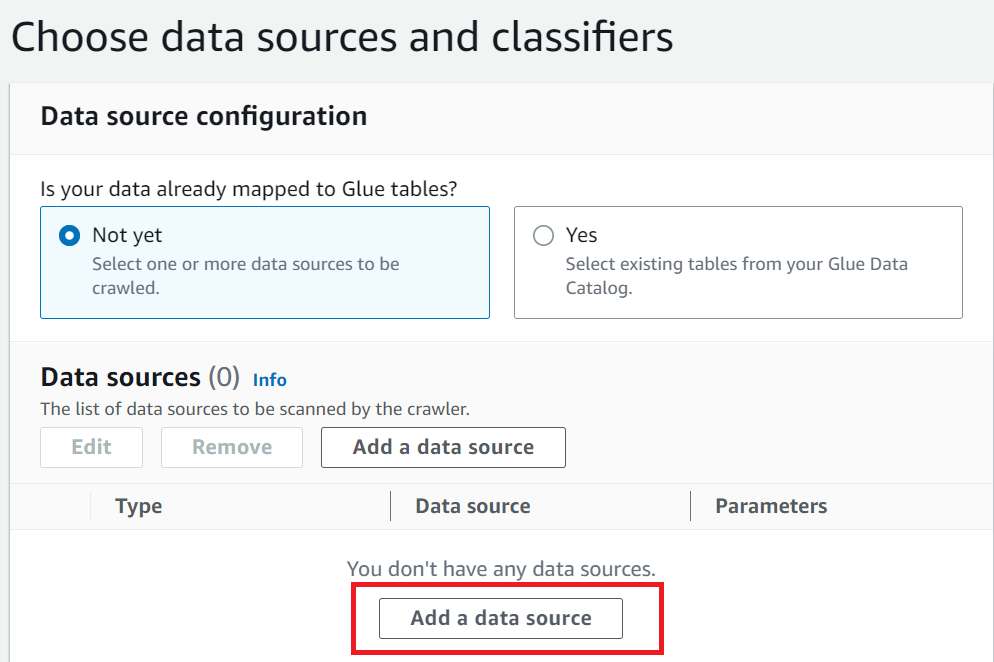
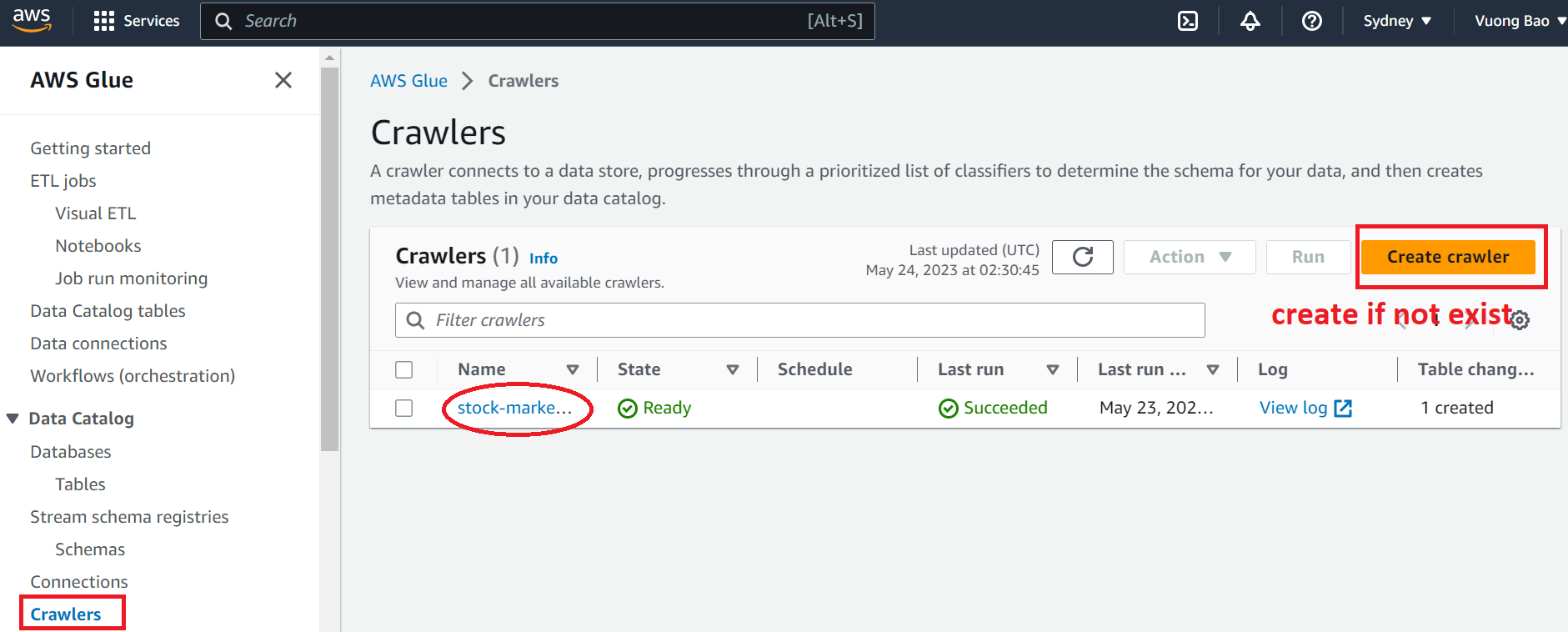
Go back to Amazon S3, refresh this I will be able to see some data over here



Build a crawler (AWS Glue)

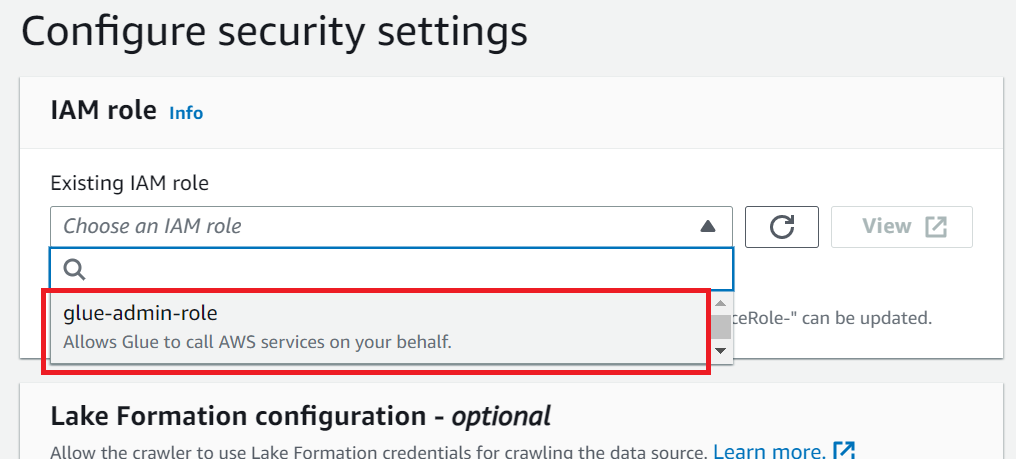
The crawler what we will do is we will craw entire schema from our S3 file so we can directly query on top of it using Athena

Step1:

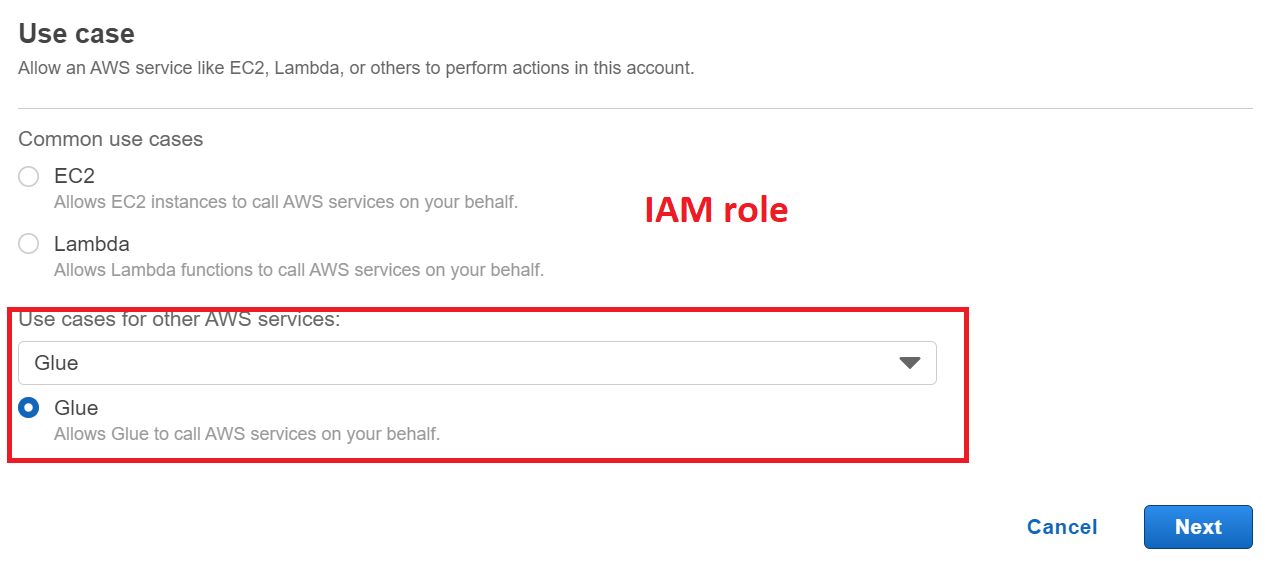


IAM role is basically the gives access to the sevices so in our case, when we want to access AWS S3 and upload data on top of it. We had to configure all of those AWS secret key and access key on our local machine. Now if these sevices such as Glue wants to talk to S3, you need IAM role. IAM role will give access to the Glue to write all of those data onto S3 bucket. If you don’t have IAM role go to Step 2(Option).

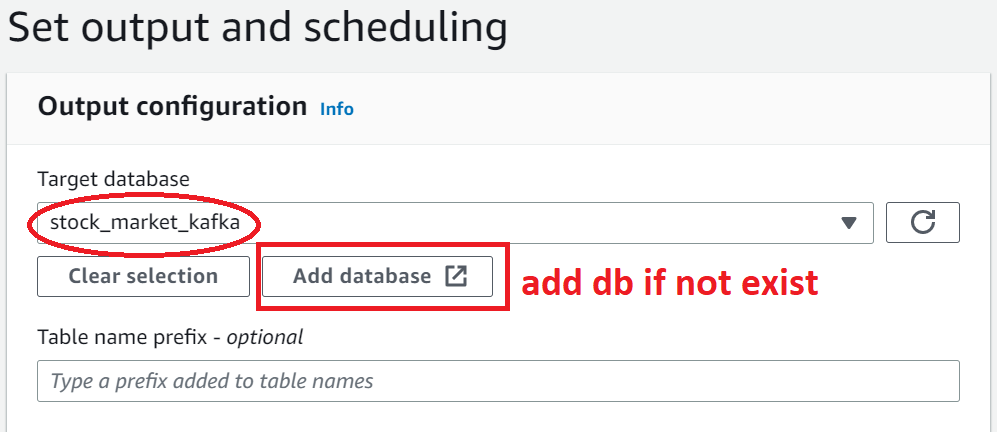
Step 2:



Step 2(Option):

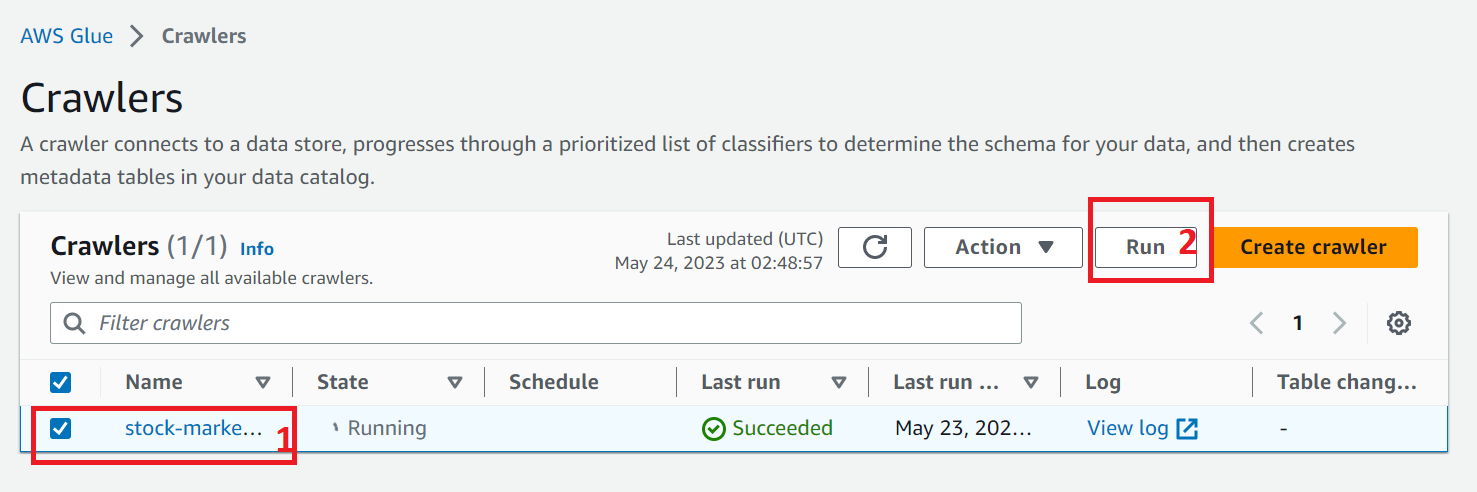


Step 3:

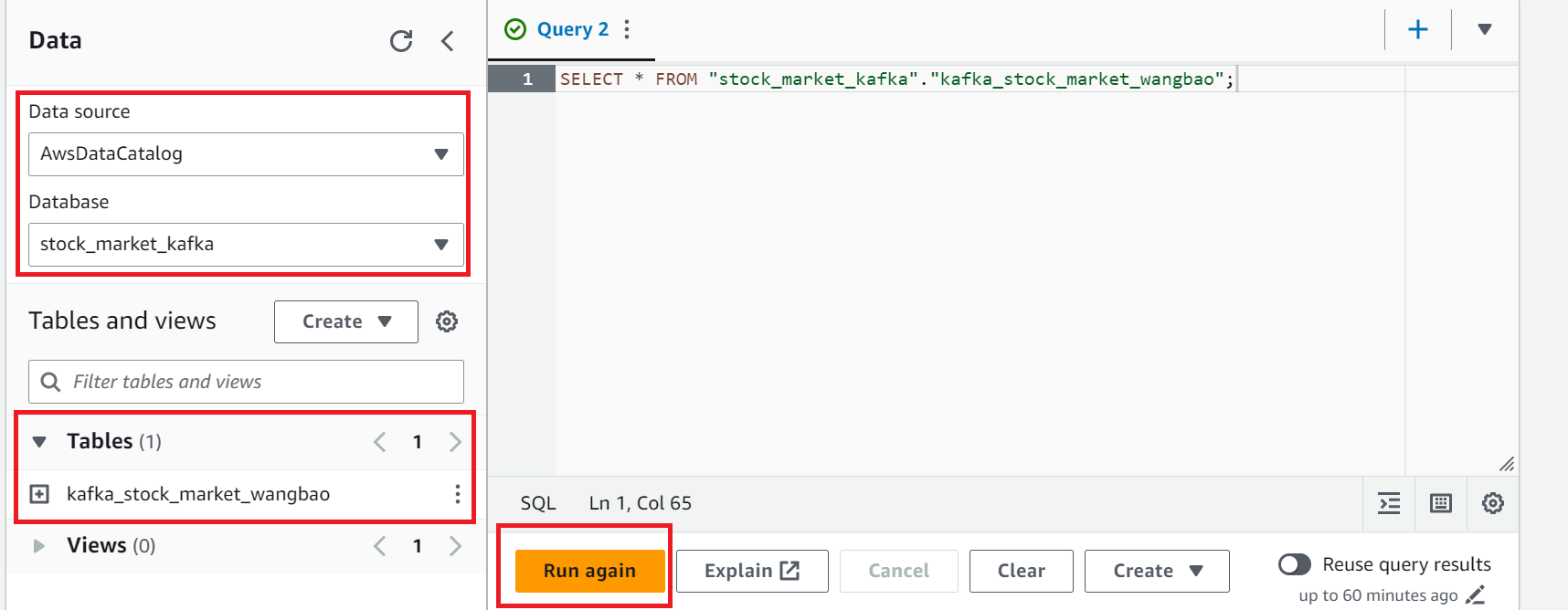


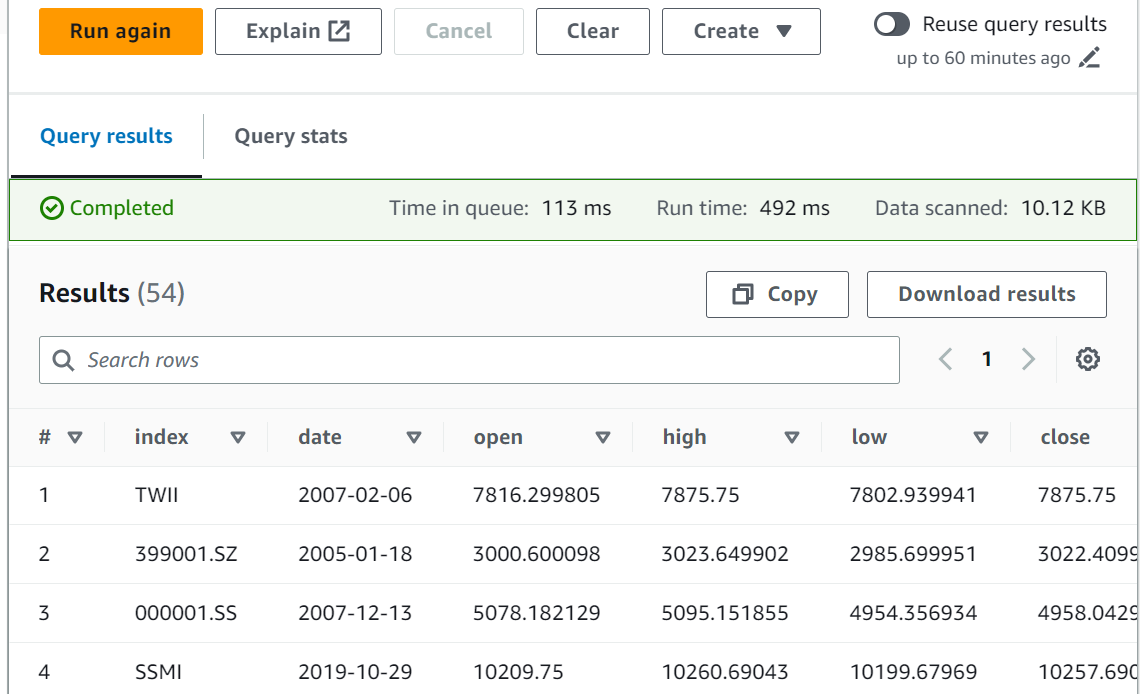
Next -> Create Crawler

Start the crawler running and wait till the crawler finishes



Go to Athena





If you get some error while querying, that is not able to find some target location All you need to do is go to settings.

