**Working with AWS**

Different Type of Storages in AWS

S3 : A cloud based storage which you can access from any of your client machines using SSH (using public key)

EC2(Elastic Cloud Compute) instance storage : This is the storage which is available free of cost with EC2. But dis advantage is , It will be auto cleared once we stop the instance

EBS (Elastic Block Storage) : This is the network attached storage. Once you will create an EC2 instance with EBS storage, you will be provided with an option whether to delete the volume when

instance is terminated.If we do not check that check box , volume and data will be remain as it is in the EBS storage and we can re use it once we restart the same instance or

even with another instance.But untill the allocated volume is there, amazon will charge you for the volume.

VPC : Virtual Private Cloud - this is nothing but the grouping of EC2 instances to groups.

We use different ips for different purposes in AWS

private ip which does not change until terminated is used to communicate with other EC2 instances with in VPC

Public ip for external world to communicate with EC2 instance.Default public ip is free. But it change every time if we stop and start,or terminate and create EC2 instance

Elastic IP can be used to overcome this , but it comes at nominal cost.We can create and then attach or detach elastic ip to any EC2 instance

Apche tomhat webserver in centos 7

Note: To install any type of tools in centos ,use sudo yum -y install <package name>

>>Installing apche webserver in AWS EC2 instance with centos 7

sudo yum -y install httpd

yum : this is centos tomhat package manager

>>Starting apache webserver

sudo systemctl start httpd

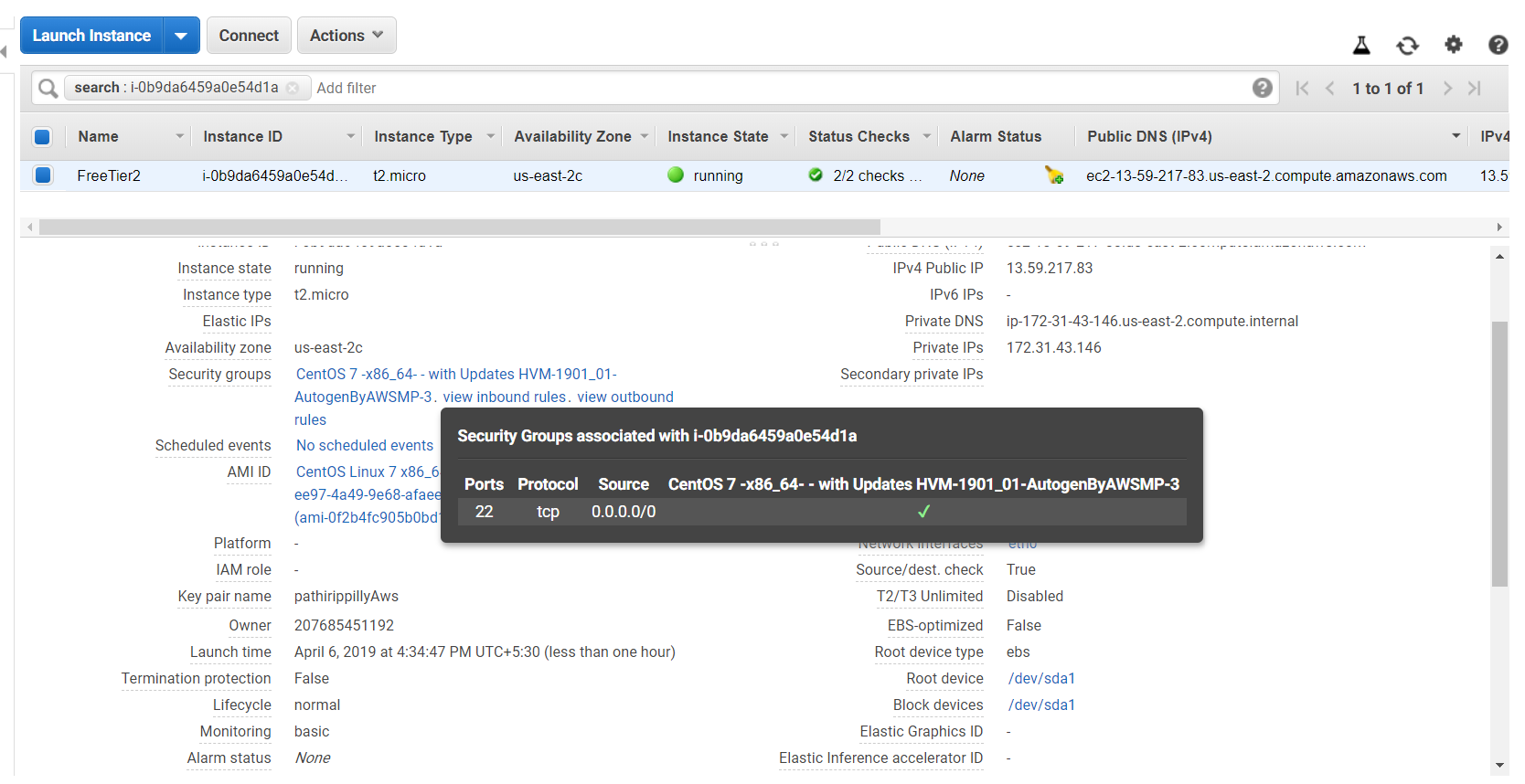
>>Check the status of the running webserver

sudo systemctl status httpd

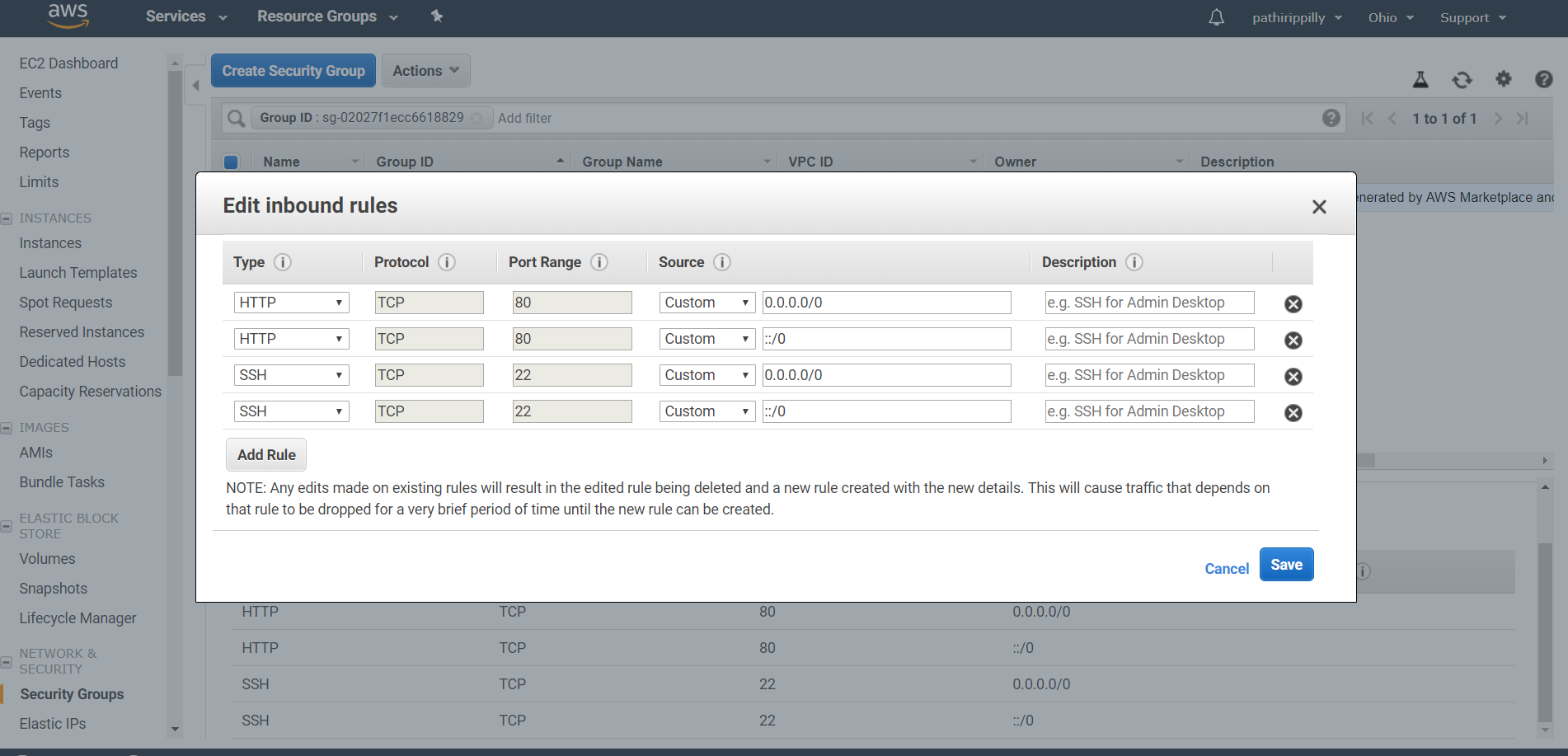
Now this webserver is running on port number 80. But Security group has enabled only port number 22 through which we are doing the SSH.

So to access such a webserver we need to add it in security group

If you see below screen shot , only port 22 is enabled



Now go to the security group and add HTTP type which will come with port number 80. Add that and save. Now you will be able to open any we server running on port 80.



Note: Like this for connecting with any running server, we need to ensure particular port is added in security group.

**AWS CLI**

Install AWS CLI as per the instructions in <https://aws.amazon.com/cli/>

**AWS PRICING**

EC2 instance,S3 storage ,data transfer are variable cost associated with it.It means if you are using it (a running EC2 instance, an occupied S3 storage, volume of data transferred), it will be priced while EBS,elastic ip are having fixed cost which means even if you are using it or not , once it is provisioned , you need to pay.

**Services**

EC2 is a PaaS(Platform as a service) while AWS has many other types of services

RDS – Relational Database Services

* Oracle
* MySQL and Aurora
* Red Shift : Light weight SQL based reporting Database
* Dynamo DB : No SQL reporting database
* And many more

Analytics: Whole suite of big data and tools related to analytics

* EMR : Elastic Map Reduce – Hadoop and Spark as a service
* Kinesis – real time ingestion and streaming
* Athena – SQL on top of S3
* And more

Lamba – For serverless applications

**Difference between Regular on-prem clusters vs** **clusters leveraging cloud’s pay as u go model**

regular on-prem cluster :

* You need to set up the platform with required infrastructure by taking the resources for rent from data centers like ‘colo’. Sometimes based on requirements, you may end up in renting really beast configurations which costs high
* If your company has rules such that the data should reside ion their own data centers and should not take outside, then we will go for on-prem cluster
* For on-prem clusters, we need to set up all infrastructures, and need to configure it by hiring a architect.
* Even if you are using the clusters for only specific period of time in a day, you need to pay the cost for the entire day

clusters leveraging cloud’s pay as u go model:

* You can configure the clusters very easily like AWS EMR
* If you are running the clusters only for a specific period of time in day, you just need to pay for those hours

**Configuring python3 for spark in AWS EMR**

Run the following command to change the default Python environment while the EMR cluster is running:

sudo sed -i -e '$a\export PYSPARK\_PYTHON=/usr/bin/python3' /etc/spark/conf/spark-env.sh

**AWS CLI and Hadoop**

1. Listing what are there in s3 bucket : aws s3 ls
2. Listing what are there in s3 bucket using hadoop fs : hadoop fs -ls s3n://bucketname/