AWS EMR Session
==========

1. What is a VM (Virtual Machine)

There will be a big server(which holds lot of resource for example:

128 GB ram
4 Tb hard disk
32 cpu cores

4 small machines

32 gb ram
1 tb hard disk
8 cores of cpu

this machine is called as a virtual machine.

in 1 big physical machine 4 smaller logical machines are running.

that is why it is called a VM or a virtual machine.

2. on-premise

this is a very challenging process.

3 important vendors providing hadoop distribution
==========

1. cloudera

2. Hortonworks

3. MapR

these vendors are not having a good time right now.

world is moving towards cloud where very less pain is there in terms of setting the cluster and doing configuration.

3. big data on cloud

3 main providers of big data on cloud
==========

1. Amazon AWS (EMR) Elastic map reduce
big data hadoop services fully managed by aws on cloud.

2. Microsoft Azure (HDInsight)

3. Google GCP (Google DataProc)

4. HDFS vs S3

S3 is a distributed file system fully managed by AWS.

you are not worried about managing it as it is managed fully by amazon aws.

4 machines - vm's

each of this VM has some:
Memory (Ram)
Storage (hard disk)
CPU (compute)

I store a 2 GB file on HDFS.

1. Start cluster

2. Do the processing

spark programs reads the file from HDFS

do some processing

and writes the result back to HDFS.

3. terminate the cluster

your HDFS data will be gone or lost.

and if we do not the cluster we incurr huge bills for the cluster charges.

The solution is S3

take the data from s3

put to HDFS

do the processing in spark

put the output to S3 and then shut down your cluster.

S3 storage is decoupled from your cluster.

4. There are 3 kinds of instances in amazon aws.

1. on demand instances

the instances which you procure whenever you require on demand and then once work is done you terminate them. we need to pay on hourly basis.

2. spot instances

aws gives this on a heavy discount price upto 90% discount compared to on demand.

but aws can take it back whenever it wants by just giving a 2 minute prior notification.

3. reserved instances

you have to commit upfront that I will use this resource for a long time. and then aws gives some discounts are you are committed for a long time.

5. 3 kind of nodes that we can have in spark cluster
==========

1. Master - this manages the cluster. this will single ec2 instance.

2. core Node - each cluster has one or more core Nodes.

core Node hosts your HDFS data and also it is capable to run tasks

3. Task Node - these are the nodes which can only run tasks. It cannot host the data.

if your application is compute heavy that means it requires lot of processing. then you can opt for few task nodes which can add more computing power to cluster.

spot instances are a good choice for your task nodes.

6. Transient vs Long running cluster
==========

Transient cluster is the one which automatically terminates when all the steps are done.

Long running cluster - where we have to manually terminate.

your reporting job which is lets say a spark job is there which runs 12 pm everyday.
then you can go for a transient cluster.

in amazon S3 we need to create a bucket. A bucket is nothing but its like a folder.

trendytech-sumit

ssh -i sumit_key_pair.pem hadoop@ec2-13-232-140-185.ap-south-1.compute.amazonaws.com

I want to bring the jar in my driver machine.

so first of all I have to put it to s3 bucket

and then I need to download it from S3 bucket into the master machine.

the history server should be configured on port 18080

I should allow inbound traffic to it.

Zeppelin Notebook

pyspark - jupyter notebook

scala - zeppelin notebook

==========

M - general purpose
C - compute optimized
R - memory optimized

Instance storage vs EBS only

AMI - operating system will be installed on ec2 instance. Centos 7

t2.micro free tier

instance states - start , stop , reboot , terminate

when we stop and start the contents on instance store are deleted
when we reboot then its not deleted.

when we reboot public dns wont change.
stop and start we will get new public dns.

storage
========
instance store
instance store is tightly coupled with ec2 instance
its like a local store in ec2 instance

s3 - simple cloud storage service
it is a cloud based storage accessible from anywhere

EBS - elastic block storage
network attached storage accessible from ec2 instances using mount.

Networking
============
public ip vs private ip

concept of grouping multiple ec2 instances is called virtual private cloud.

primarily to group related instances for security reasons.

network switch (private) powerful ethernet cables

network switch (public)

ec2 to ec2 transfer if we misconfigure to use public ip to copy the files then we need to pay a lot.

we should use private ip's to transfer files from one ec2 to other.

private ips wont change util termination

public ips might change during start and stop as well.

we can use elastic ip, it comes at a very nominal cost.

inside network & security we have service called as elastic ip.

3.5 dollars per month for each elastic ip.

to use elastic ip.. after creation of it, we need
actions -> associate address ->

Authentication
==========

by default password login is disabled for ec2 instances

we need to have keypair to connect without password.

public key and private key

.ssh folder on ec2 instance has public key

private key is in pem file.

security group
==========

a logical firewall

one security group can be applied to multiple ec2 instances.

AWS CLI
==========
to automate the things

shell scripting

or programming langauages api for example for python boto is a famous choice.

Pricing
========

ec2, ebs/s3 , elastic ip, data transfer

ebs and elastic ip are fixed cost.

cost of ec2 will will be vm cost + AMI (software cost)

aws pricing calculator

Databases
==========
RDS

DynamoDB

Amazon Redshift

Kinesis

---

for root file system EBS should be used and not instance store

EBS volumes - option delete on termination

create image to generate a template

==========

Athena - to run sql queries on top of s3 data

glue

walmart grocery

AWS CLI Export

/etc/hadoop/conf

classification=hdfs-site.xml,properties=[dfs.blocksize=64000000,p2=v2,p3=v3]

==========

ambari

yarn configs

yarn hosts

spark-shell --conf spark.dynamicAllocation.enabled=false --num-executors 40 --executor-cores 1 --executor-memory 1g

spark.executor.memory 2g
spark.executor.cores 2