HADOOP GROOMING SESSION(21FEB17)

Explain “Big Data” and what are five V’s of Big Data?Please explain?

What is Hadoop and its components.

HDFS - storage unit

MAPREDUCE or YARN - processing framework

What is HDFS? How the data get stored on HDFS?

components of hdfs?

Namenode

Datanode

Secondary Namenode or Standby Namenode

What architecture ?

Master - Namenode and SNN

Slaves - datanodes

2 GB of data on the hadoop cluster

blocks = ?

block size = 128 MB

1024x 2/128 = 16 blocks

block\_file

block\_file.meta

cluster = 50 datanodes

what is replication factor ?

2 replicas of the same block on the same datanode?

in a 2 node cluster - repl factor is 3. how many blocks would be there for each replica?

One as a missing replica fsck command.

under replicated blocks

What are racks ?

What is rack awareness ?

Rack 1 = 3 Nodes , Node 1-2-3

Rack 2 = 2 Nodes Node -4-5

Block 1 - 1-2-3

or

1-2-4

hadoop fs -ls

equal to

hadoop fs -ls /user/hduser

hadoop fs -put file1.txt

/user/hduser

Namenode-master

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

meta data - block id, locations, repl factor, permissions for each file

where is this meta data stored

fs-image and edits

AND

RAM of the namenode-to have the faster accessibility of the data

what happens if the memory gets saturated in namenode?

no more fresh data can be added up in the cluster

Namenode federation was the solution i.e., We can add more primary namenodes

NN-1 and NN-2

50 node - 100 GB extra space unutilised

what is HA in namenode.

High Availibility in Namenode

secondary name node is cold backup

standby namenode is hot backup

zookeeper

Datanodes - slaves

which node is responsible for replication?

Process of replication is done by datanode

on the instruction of the namenode

Reading and Writing in HDFS

1) sequentially

Process is done parallelly

who writes the data on datanodes

1) client

2) namenode

Ans:client

pipeline of datanodes

who reads the data from the datanode

1) client

2) namenode

Ans:client

client will frst approach the NN. Take the metadata for that block

USer or a client interacts with datanode directly.

Difference between 1.x and 2.x

1. Standbynamenode
2. Namenode federation
3. Yarn architecture