Big Data works on distributed system (cluster – Collections of machines)

Hadoop Framework installed on top of linux

Below tools that helps us to connect to cluster - SSH

1. Putty : execute commands
2. WinSCP : help to transfer data/file to the cluster from your laptop
3. MobaXTerm : does both 1 and 2

d/rwx/rwx/rwx

directory or file/owner/group/others

r - read - 4

w – write -2

x – execute – 1

chmod 777 a.txt

chmod 774 a.txt

vi

nano

gedit

Hadoop Framework that works on distributed system

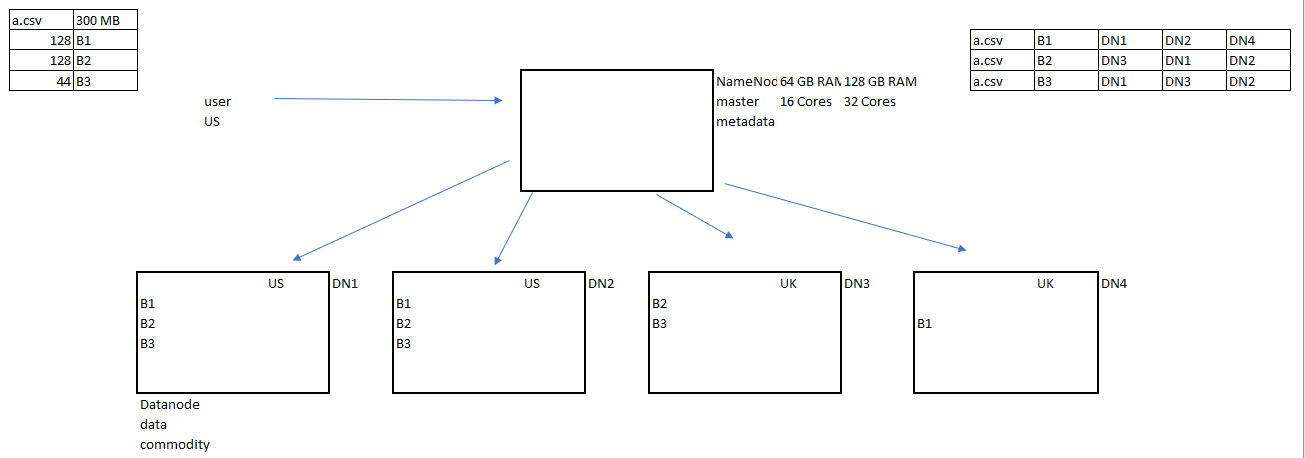
3Main components Hadoop

1. HDFS – Hadoop Distributed File System – Storage
2. MR – MapReduce – Processing/Computation/Transformation
3. YARN – Yet Another Resources Negotiator

Types of cluster

1. On prem : the machines are physically located at some place.
2. Cloud
   1. AWS – Amazon --60
   2. Azure – Microsoft --30
   3. GCP – Google –10

HDFS



1. It works on a master -slave architecture
2. Hadoop supports commodity hardware(cheap/less)
3. Hadoop distributes a file into blocks
4. Block size is 128MB(Hadoop 2 n 3) 64MB(Hadoop 1 )
5. What if one DN goes down?
   1. Replication factor(RF) of 3 by default (Availability, Reliability)
6. How data is served?
   1. Data locality :
7. What if all DN goes down
   1. Rack Awareness :
   2. 2 copies of RF in same rack/region and 1 copy in another rack
8. What happens to the data of the DN when it fails?
   1. When DN goes down, NN will fall short of RF
   2. and hence create a new copy on new DN
9. What happens to the DN which failed comes back again
   1. It will be used for new data storage
   2. Because existing blocks RF of 3 is already met
10. Can we change the RF?
    1. Yes, in configuration files
    2. Is it necessary?
    3. Ideally , No
    4. Why is it configurable then?
    5. To provide flexibility to business need

scalability