DATE: 15/Sept
PAGE NO.:

Session-2

- Basics of storage.
- code in hdd (hard disk) is known as program. But when it is in running state then it is called process.
- CLI ≅ console ≅ Terminal
- ctrl + c vs ctrl + z
  - ↳ Terminates ↳ pauses the prog in bg.

fg 2 → forward ground (resume the pro
  - ↳ job number

firefox &
  - ↳ run the cmd in bg.

touch file.txt
  - ↳ creates an empty file

mkdir file

ls ~~-l~~ → longlist
  - field switch

rmdir

~~user~~ useradd gaurav
  - ↳ adds a user

passwd gaurav.
  - ↳ adds the password for the user gaurav

Ctrl + Alt + F1 / F2 / F3
  - ↳ changes the ~~d~~ desktop/screen.

DATE : / /
PAGE NO. :

vi my.txt
↳ esc → cmd / control mode
↳ i → insert mode
↳ esc ;
↳ :w → written or save
↳ :q → exit the file
↳ :wq → saves and exits the file.

df (-h) → converts bits to MBs.
↳ maximum limit of data we can store.
↳ shows h.d.d

Big Data → Major problem
↳ volume ↳ sub-problem
↳ velocity ↳ cost high

Bit → Byte → KB → MB → GB → PB → EB → ZB → YB
← Brontobyte ↓

- If we create hard disk of size YB issue will come of input / output (new problems will emerge)

free -m
↳ shows memory (ram)

watch date
↳ runs the cmd every 2 sec for ∞ time
↳ watch -n 1 free -m

DATE : / /
PAGE NO. :

vi
↳ esc
↳ yy → copies current line
↳ p → paste data
↳ np → pastes n times.

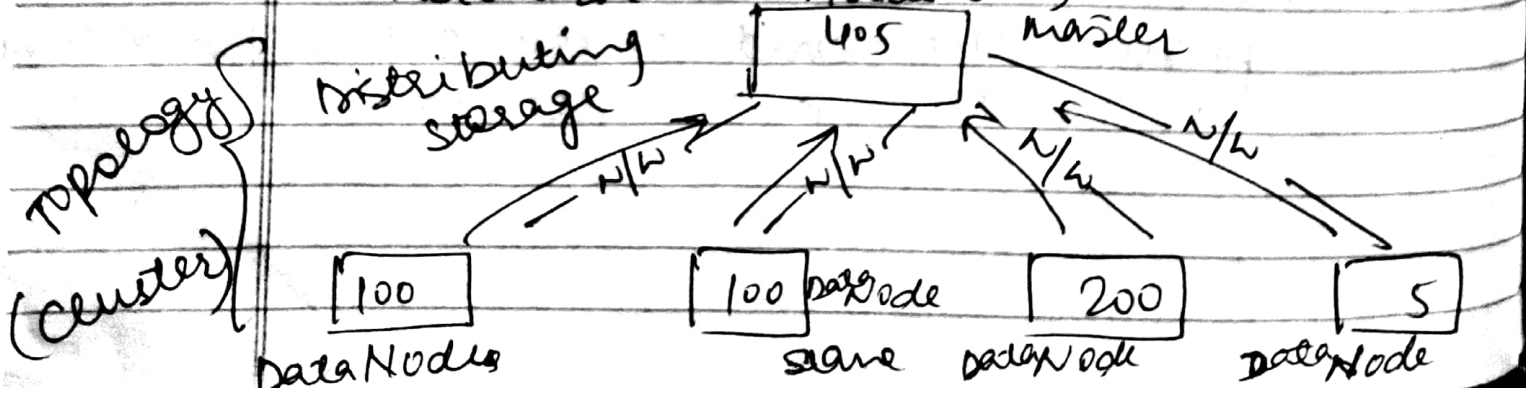
- Concept
  ↳ distributed storage
  ↳ split data

{ Servers
OS
RAM / CPU / HD } • more powerful than that of PCs.
• cost ↑↑.
• Reliable ↑.

Same pc hardware → Commodity

- { Master : Recieving the storage
  Slave : & Contributing the storage.

Master - Slave Relationship.
• All are connected through network.



DATE : / /
PAGE NO. :

- similary we can connect RAM, CPUs, etc
  ↳ Distributed computing

HADOOP → software
↳ ~~def~~ solves the problem of big data through the concept of distributed storage.

* Self - work
  Doc / Article / Post / Blog.
  regarding big data.
  ↳ 99% solved by hadoop.

DATE: 16/Sept
PAGE NO.:

SESSION-3

- Prompt is known as shell prompt.

#. vim my.txt

cat my.txt → used to read a file.

source my.txt
↳ runs all the cmd in my.txt.

- Bash shell scripting

echo hello → prints the text.

espeak-ng hello → speaks the text.

TASK: Find a cmd which will send the data to whatsapp, sms, email. and print the data.

echo `date` or # echo $(date)
↳ special symbol that will differentiate b/t cmd and string
(Ctrl + console / back quote)

rpm -qf /usr/bin/date
↳ package
↳ query → file
↳ displays the name of the software.

man cmd
↳ opens the manual for the cmd.

DATE: / /
PAGE NO.:

TASK: espeak-ng ; Find the cmd to change the language of the cmd.

rpm -q -f 'which firefox'

rpm -e firefox
↳ erase

init 0
↳ shut downs the OS.

rpm -q hadoop

- similar like Hadoop
  ↳ S3, Ceph, Glusterfs

- HDFS protocol is used in Hadoop
  Master —HDFS→ Slave
  ↳ Hadoop Distributed File System

java -version

- winscp - program used to access transfer doc from win to linux

DATE : / /
PAGE NO. :

- Software → Drag & drop in /root.
  - (i) hadoop
  - (ii) jdk.

rpm -i jdk<TAB>.

java -v (HotSpot) ORACLE

rpm -i hadoop<TAB>. ← fails

rpm -i hadoop --force.

rpm -q hadoop
hadoop version //check

DATE : 17 Sept
PAGE NO. :

SESSION - 4

- Basics of Binary
- ASCII – Basic knowledge
- In networking we use bits not bytes
- ASTERICK – software used to create our own conference servers.

- SPOF - single Point of failure.

- Horizontal scaling & Vertical scaling
  (SPOF) ↔ (SPOF)
  ↳ Replica

- valid ip address.

jps → does the system is master or slave

cd /etc/hadoop/
  ↳ configuration file of hadoop.

vim hdfs-site.xml
Master
```
<property>
<name> dfs.name.dir </name>
<value> /nn </value>
</property>
```

mkdir /nn ~~(in / dir)~~ // created in / dir.

rpm -i -v jdk ___.rpm
  ↳ verbose (shows all the bg process).

DATE : / /
PAGE NO. :

<Same >>

cd /etc/hadoop/

vim hdfs-site.xml

// same as master

<name> dfs.data.dir </name>
</value>
</property>
</configuration>

DATE: 18/Sept
PAGE NO.:

SESSION - 5

- Ip address consists of 4 octets, each of 8 bits. 32 bits / 4 bytes in total.
- ifconfig in Linux
  ipconfig in Win.
- MAC address — 6 byte.

ifconfig enp0s3 4.3.2.1
  ↳ changes the ip address of a system

ping www.google.com
  ↳ checks the connectivity.

nslookup www.google.com
  ↳ gives the ip of the fav. site.

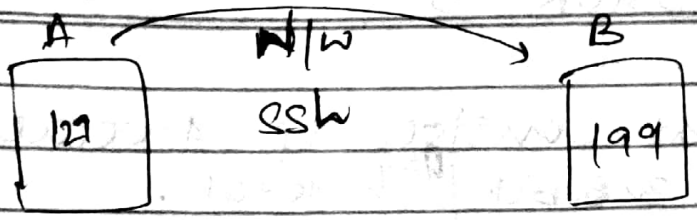
- DNS - domain name system
  ↳ gives the name to ip addresses.

bc
  ↳ binary calculator

  ibase = 10 ← input base
  obase = 2 ← output base

- protocol = = Process
  win ↳ RDP (Remote desktop protocol)
  linux ↳ SSH (Secure Shell)

DATE : / /
PAGE NO. :



$ ssh 192.168.0.199

init 6
↳ reboots the system

- For android - linux connection
  ↳ Juice Ssh (app)

putty
↳ connects win to linux using ssh.

ping 192.168.0.129
↳ from slave to master to check the connectivity.

- we never connect with the ip address. Through we check the connectivity. To connect we use some external software.

~~master # hive core site.xml~~

masters # jps
↳ checks the process of master node.

DATE : / /
PAGE NO. :

vim core-site.xml

- <property>
  <name>fs.default.name</name>
  <value>hdfs://192.168.0.129:9001</value>
  </property>

netstat -tnlp
↳ checks how many ports are running in the linux system

909/cupsd is used by printer

hadoop namenode -format

- process == service == daemon

hadoop-daemon.sh start nodename

jps // check the conn.

hadoop dfsadmin -report
↳ shows how many slaves are connected.

DATE : / /
PAGE NO. :

slave.

jps

vim core-site.xml

<property>
<name> fs. default. name </name>
<value> hdfs:// 192.168.0.129:9001 </value>
</property>

↑ MASTER (ip)

→ SAME IN ALL MASTER & SLAVE.

jps

hadoop-daemon.sh start datanode

The service will not start.

master

systemctl stop firewalld.

slave

hadoop-daemon.sh start datanode
jps.

It WORKS!!!

master

hadoop dfsadmin -report.

DATE: 22 Sept
PAGE NO.:

SESSION - 8

AWS cloud ⟷ Hadoop

- Launched an instance – RedHat.
- Use putty to interact with the OS.
- Change .pem to .ppk using puttygen
- Run putty
- Upload JDK & Hadoop via WinSCP
→ (WinSCP) scp protocol used to move files
→ ssh - remote login (putty)

- Connecting private key in WinSCP.
  Advanced... → Advanced... → SSH (Auth)

- OOM – out of memory
  Hadoop is heavy if not provided with sufficient memory.

cat /proc/sys/vm/drop_caches
0 ⟶ each no. has its own meaning.
echo 3
echo (3) > /proc/sys/vm/drop_caches
→ cleans the cache mem.

- configure hdfs-site.xml
  some change in core-site.xml

DATE : / /
PAGE NO. :

slave

<property>
<name> fs.default.name </name>
~~hdfs://~~
<value> hdfs://15.207.72.215:9001 </value>
</property>

↓ public ip of master

only Master

<value> hdfs://0.0.0.0:9001

open ip

EC2 → pub ip
→ pr ip

master only # hadoop namenode -format
↳ format the disk

jps

master only # hadoop-daemon.sh start namenode
↳ to start the master node

hadoop ~~dfsadmin~~ -report
b dfsadmin

DATE: 23 Sept
PAGE NO.:

SESSION - 9

- when we stop & start the instance, the ip address changes.
- Logical storage != physical storage
- Logical storage ≠ virtual storage
- Hadoop 1.2.1 ; by default size of strip is 64 Mb

ip of master: 50070 ← url
↳ shows the report in web.

- configured client.
- In client we do not have to start the service though have to be configured in order to connect to the master node.

Master
vi core-site.xml } ll do not have to do with hdfs-site.xml

```
<property>
<name>fs.default.name</name>
<value>hdfs://65.0.31.143:9001</value>
</property>
```
master pub ip

client # hadoop fs -ls /

hadoop fs -put lw.txt /
copy filename directory

hadoop fs -rm /zzz.txt

hadoop fs -cat /zzz.txt

hadoop fs -D no space dfs.block.size=33554432 -put zzz.txt
changes the block size

- It does not support Mb, only Bytes is supported. we can change block size per size

yum install bc
rpm -q bc
bc
32*1024
32*1024*1024 ≈ 32M into bytes.

DATE: 24 Sept.
PAGE NO.:

SESSION -10

#. hadoop-daemon.sh start datanode/namenode
↳ starts the service temporarily.

- Port no: 50070 → web ui
- Port no: 9001 → service port of hadoop master.

cd /etc/rc.d/
cat rc.local

13.233.54.
13.233.254.248

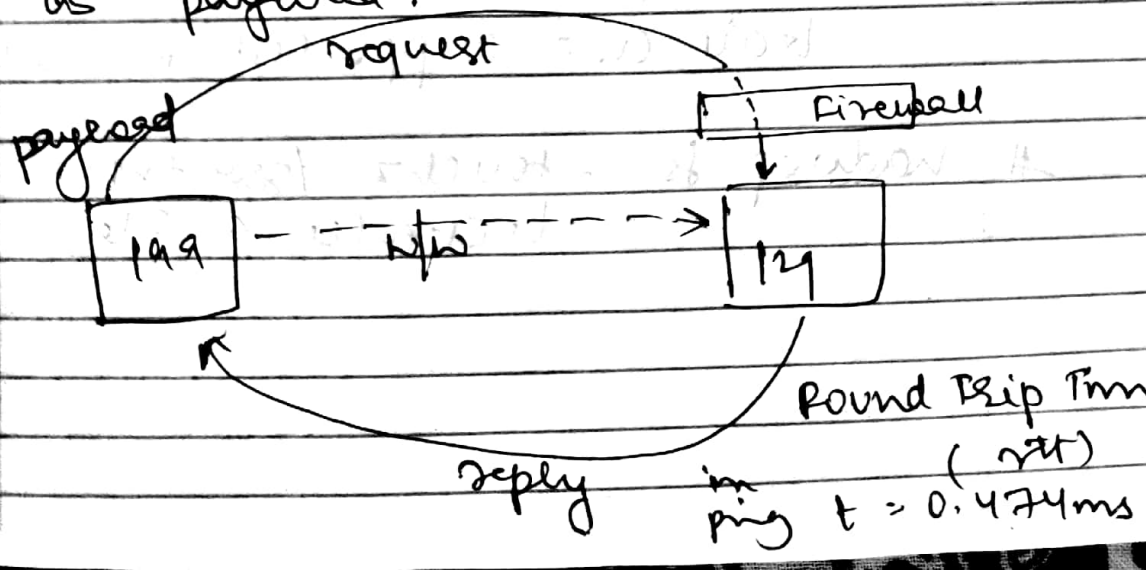
yum whatprovides vim

vim rc.local
hadoop-daemon.sh start namenode

chmod +x /etc/rc.d/rc.local

- For client we only need to update the core-site.xml file.

- Information sent on network is known as payload.



DATE : / /
PAGE NO. :

- 64 bytes is the size of payload attached with the packet.
- ICMP : Internet control messaging protocol
- * ICMP Tunnel
- PING : Packet Internet Grouping

ping -c 2 192.168.0.129
↳ counter (it will stop after 2 pings)

alias vimal="date"
↳ gives a nickname to cmd.

useradd -s /usr/bin/cal tom
passwd tom
↳ tom is restricted only to cal cmd.

hadoop fs
↳ all cmd supported in hadoop

hadoop fs -touchz 188g.txt
↳ creates a file

DATE: 29 Sept.
PAGE NO.:

SESSION - 12

- System (CS) crashes after year 2038.
  - proof ; date cmd.
  - memory overflow
- when more space is ~~used~~ ~~to~~ required than given memory then the extra data will be automatically discarded by the computer.
- Dec 13, 1901 15:46:00 — 0 sec
- Time is stored in seconds.
- Jan, 18, 2038 22:14:07
  - All 32 bits are full
  - Last date
- Dec, 13, 1901 15:46:00
  - It resets.
- Difference bet 32bit & 64bit OS.

DATE: 30 Sept;
PAGE NO.:

SESSION-13

- Master (Hadoop) ; SG → Rule in inbound traffic,
  Type - All
  ↳ Client can do all
  ↳ ssh
  ↳ ping
  ↳ scp , all.

vim core-site.xml
<value> hdfs:// ip(pub) master:9001 </value>

- We only have one public ip but they can have diff. private ips.

rm -rf /dn
↳ deletes the file.
mkdir /dn
- # Delete & recreate the folder (Quick troubleshooting)

- If the master node is in safe mode we can't make any changes in it.

master # hadoop dfsadmin -safemode get
↳ displays the mode.
" # hadoop dfsadmin -safemode leave.
↳ changes the state
(Safe_mode is off)

DATE : / /
PAGE NO. :

- Hadoop has a replication factor of 3.
- Block size : slices the file (64MB def.).
  Replication : Makes copy of file (3 by def.).

TASK : Create a cluster
- 4 slaves
- 1 client
- when you upload a file

- Installation of Py environment :

DATE: 1/Oct.
PAGE NO.:

SESSION - 14

- Basics of networking

tcpdump -i enp0s3 -n
↳ lists all the ICMP packets
↳ monitor/ packet capturing

pgrep firefox
↳ shows PID (process ID) of the program.

- clear text : Text anyone can read
↳ It has no encryption.

tcpdump -i enp0s3 icmp -n
↳ filters the icmp packages.

- But ssh will be failed.
Alt. cmd
tcpdump -i enp0s3 tcp port 22 -n
inside tcp; port : 22 (ssh).

- Basics of Python (Introduction)
↳ crossplatform language.

SSDID → data transfer port.

DATE : 2 / 004.
PAGE NO. :

SESSION - 15

- Only & only client decides the block size & repication of a file. as well as client.
- Hardware address ☠ == Mac Address

master# vi hdfs-site.xml

```
<property>
<name> dfs.replication </name>
<value> 2 </value>
</property>
```

↳ If ☠ fails. Replication factor is still 3.

client# vi hdfs-site.xml <name> dfs.replication </name>

```
<value> 4 </value>
</property>
```

client# vi hdfs-site.xml

```
<name> dfs.block.size </name>
<value> 1024 </value>
```

↳ changes the block size

- Replication ∝ Size ∝ Money
- REPL does not save the file

DATE : / /
PAGE NO. :

gedit lw.py

python3 lw.py

- IDE - Integrated Development Environment
- Jupyter Notebook
- Mostly in DS & ML we use jupyter notebook
- alt + enter
  shift + enter : Runs the prog.
  esc + DD : deletes the selected cell
- Basics of data type in py
  - int
  - string
  - bool
  - list
  - tuple.
- : - slice operator.
- List can only perform row-wise operation.
- Array can perform column-wise operation.

DATE : / /
PAGE NO. :

- programming file in which the code is written is known as module.

```
#import numpy
 numpy.array(name)
```

↳ columns are accessed by ',' comma

print setaf 3 1 → bl 7 → white
↳ changes the color of terminal

- If we want to print the content exactly we use triple quotes (''' ''').

- If - else loop

```
if int(exp) = 1:
    print("no job")
elif int(exp) >=1
    print("job")
else:
```

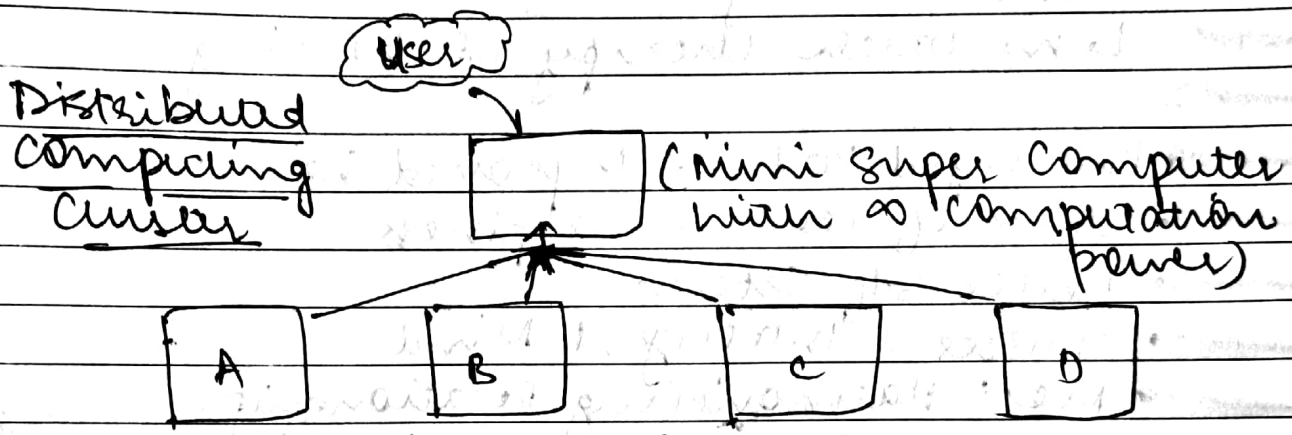
- Python does not support switch-case.

PAGE NO.:
DATE 14 Apr 21

SESSION: 143 Hadoop

*. Problem
- Large RAM (some PB)
- No SPOF

this all can be solved by BigData Concept (Super Computer)

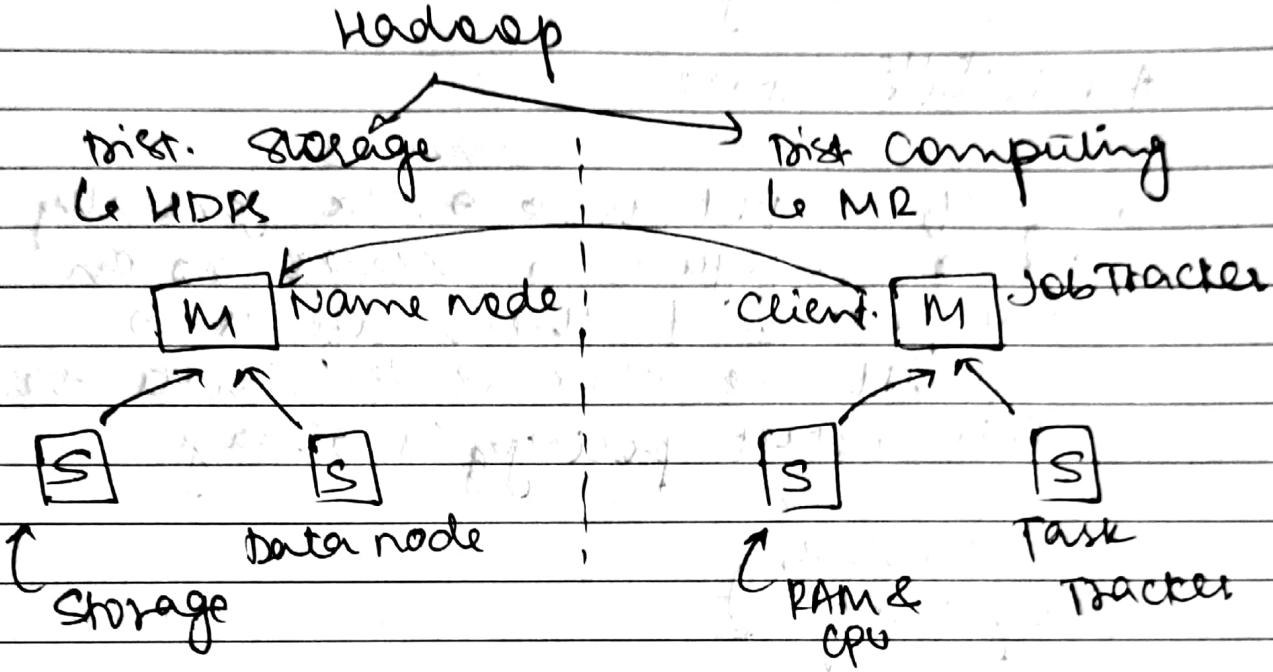


- Sharing of Computing (CPU, RAM)
- Storage → HDFS (Hadoop)
  → S3 (Cloud)
- Computing → MR Cluster (Map Reduced cluster)

- In this setup, we create the chunks/blocks of data and send it to the memory of all the slaves/nodes.

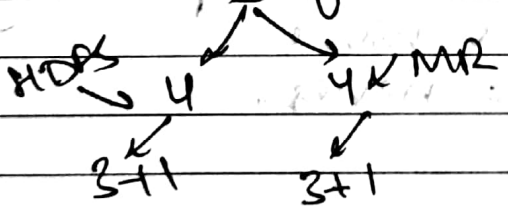
*. Product for distributed computing:
Hadoop (Map Reduced Cluster)
- only for distributed storage but we can take its internal concepts & with some changes we can reuse it.

PAGE NO. :
DATE : / /



- We need HDFS cluster for MR because, for computing we need data. ~~And~~ The large data is stored in the distributed storage.

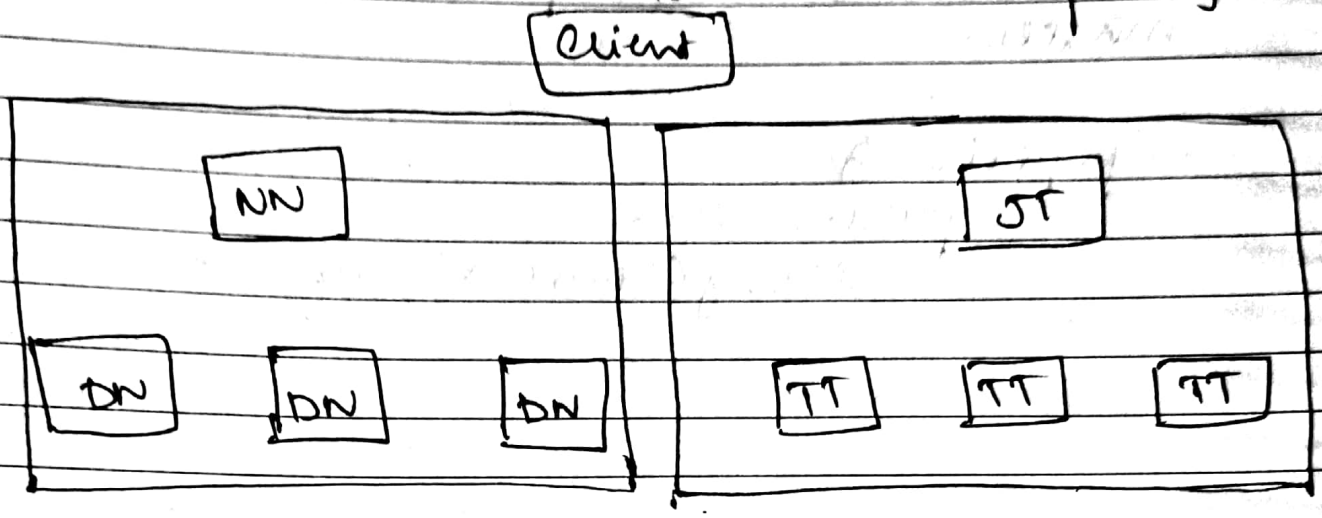
∴ Min 9 system we require
8 for HDFS & MR & 1 client



PAGE NO.:
DATE 16 Apr 21

Session: 145 Hadoop

Distributed Computing Cluster



Requirements: JDK & Hadoop software in every node.
(ec2 instance)

- Create one hadoop_template, and uninstall JDK & Hadoop.

```
# cd /home/ec2-user/
# ls
  JDK & hadoop
# rpm -ivh <jdk & hadoop>,
          <hadoop> --force
# init 0
```

- Now we will clone the hadoop_template
  # Actions > Image & Templates > Create a Image
  Image name: hadoop
  >> Create Image

PAGE NO. :
DATE : / /

- Create 9 clones of hadoop_template AMI.
  - Allow all traffic
  - Public IP enabled
  - Security Group : allow_all_port_for_hadoop

Learn!!

ec2
- Security gp
  - name : allow_all_port_for_hadoop
  - vpc : default
  - Add rule
    - All Traffic Anywhere

> create SG.

- SG : allow all traffic

> launch Instances.

NN

```
# sudo su -
# cd /etc/hadoop/
# vim hdfs-site.xml
    <property>
      <name> dfs.name.dir </name>
      <value> /nn </value>
    </property>

# mkdir /nn
# hadoop namenode -format
# hadoop-daemon.sh start namenode
```

PAGE NO. :
DATE : / /

vim core-site.xml

```
<property>
<name> fs.default.name </name>
<value> hdfs://0.0.0.0:9001 </value>
</property>
```

start service

netstat -tnlp | grep [redacted] java

↳ 50070

↳ 9001

DN # sudo su -

cd /etc/hadoop/

vim hdfs-site.xml

```
<property>
<name> dfs.data.dir </name>
<value> /dn1 </value>
</property>
```

vim core-site.xml

```
<property>
<name> fs.default.name </name>
<value> hdfs://[redacted] <master_pub>:9001
</property>                                </value>
```

• we can also put priv ip bcz all are in same vpc

mkdir /dn

hadoop-daemon.sh start datanode

NN # ~~hadoop~~ ~~dfs~~ hadoop dfsadmin -report

JOB TRACKER

JT # sudo su -
cd /etc/hadoop/
vim mapred-site.xml
<property>
<name> mapred.job.tracker </name>
<value> 0.0.0.0 : 9002 </value>
</property> ← No protocol used

hadoop-daemon.sh start jobtracker
netstat -tnlp | grep java
Lo :9002
L : 50030 ← Map reduced cluster

url = ip : 50030/jobtracker.jsp

TT # sudo su -
cd /etc/hadoop/
vim mapred-site.xml ← job will also be same
<property>
<name> mapred.~~task~~ job .tracker </name>
<value> ip of JT : 9002 </value>
</property>

PAGE NO. :
DATE : / /

hadoop daemon.sh start tasktracker

Session: 147 Hadoop

PAGE NO.:
DATE: 21/Apr/21

- Data is stored in Data Lake.
  ↑ hdfs / s3
- Hadoop is the product that come under the umbrella of Data engineer.

nn-Hadoop

Security > Inbound rule
  ↳ Delete all inbound traffic rules
  ~~I've don't want~~
  ↳ Add: all traffic sg-087f
  ↑ (Launch wizard: 47)
  Security gp

- Now all the instances with this sg will be able to send the traffic to namenode
- Now we will attach this security group (launch wizard 47) to all the nodes and change their inbound traffic to sg 47.
  (same as nn.)

* we can also change all these sg ~~from~~ by going to the Security Group.

- Configure 1nn, 2dn
  1jt, 2tt

PAGE NO. :
DATE : / /

jt
hadoop job -list-active-trackers
↳ lists all the active task trackers.

vim /etc/rc.local
↳ we can make this svc permanent by putting the start cmd here.
*. and make it executable chmod +x

client
cd /etc/hadoop/
vim core-site.xml
same as dn.
<prop>
<name> fs.default.name </name>
<value> hdfs://<ip>:9001 </value>
</prop>

cd
vim twitter.txt

hadoop fs -put twitter.txt /
hadoop fs -ls /
hadoop fs -cat /twitter.txt

vim etc/hadoop/mapred-site.xml
same as jobtracker.

35

PAGE NO. :
DATE : / /

hadoop job -list
↳ NO jobs running

cd /usr/share/hadoop/
↳ hadoop-examples-1.2.1.jar
cd ~
hadoop jar /usr/share/hadoop/hadoop-ex wordcount
↳ for job /twitter.txt /output.

(data) —input (Extract)→ Program —output (↳ load)→ (data)

- In hadoop map reduce they always read the data from hdfs ~~cluster~~ storage.

- We also need to tell job tracker where the name node is.

jt # cd /etc/hadoop/
vim core-site.xml
<same as dn>

hadoop-daemon.sh stop jobtracker
——— start ———

client

hadoop fs -ls /
↳ new /tmp/ dir is created

hadoop jar /____/hadoop-example wordcount /twitter.txt /output

hadoop fs -ls /
↳ /output

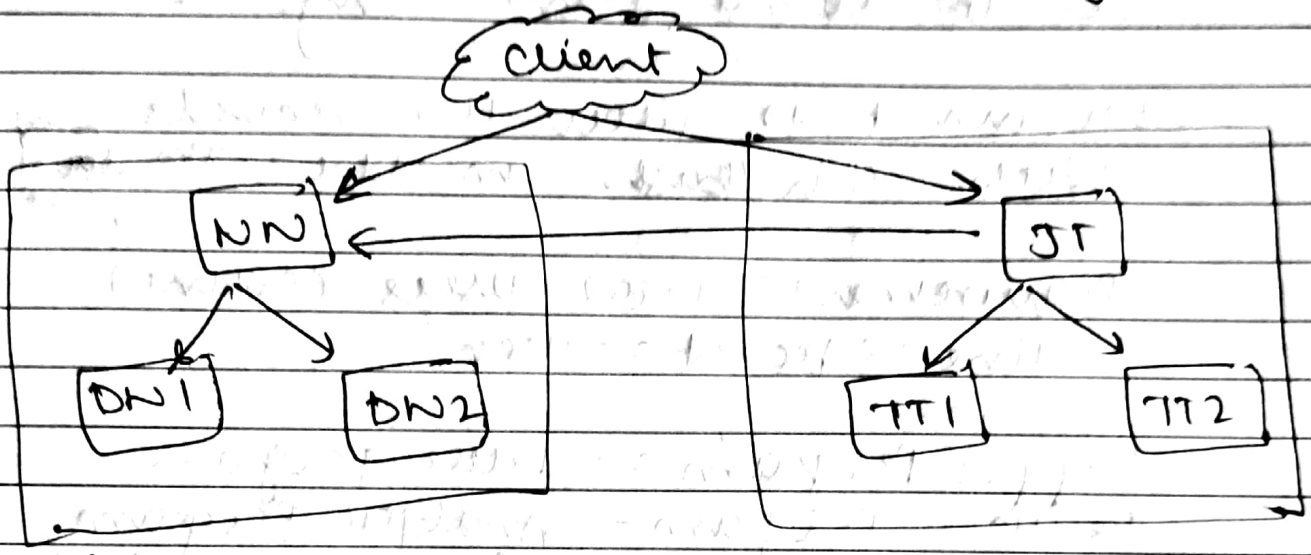
hadoop fs -ls /output
↳ ~~/000~~ /SUCCESS
i.e job is success.
↳ part-r-0000
↳ O/p is stored here.

hadoop fs -cat /output/part-r-00000

PAGE NO.:
DATE: 22/Apr/21

Session 148 Hadoop

Map Reduce programming



hdfs cluster mapred cluster.

- Cluster is created and made permanent.

client

```
# hadoop fs -ls /
# hadoop job -list-active-trackers
    ↳ 2 TASK TRACKERS.

# cat /etc/passwd/
  ↳ lists all the users
  ↳ jack x :1001:1001::/home/jack : /bin/bash
      ↳ : → field seperator/delimiter
      ↳ 7 columns
      ↳ User
  ↳ ——————————————— : /sbin/nologin
```

Encrypted Login

This user can login in bash shell

- If uid ≥ 1000 ; user is general user
  ∴ limited power.
  & root has uid = 0

PAGE NO. :
DATE : / /

- /etc/passwd/ is a raw data. bc. it doesn't have any metadata
- We want to filter the records with /bin/bash in the 7th field

Requirement: Total users (count)
line == record == user

- Mapper Program == filter program
  Reducer Program == analytic program. (gives required o/p)

cat /etc/passwd | grep bash
↳ filter bash (Mapper)

cat /etc/passwd | grep bash | wc -l
↳ 6 (count) (Reducer)

hadoop fs -put /etc/passwd /
——— -ls /
↳ /passwd

- In Hadoop, there is a streaming API where we can use languages other than Java, ex: Python, scala, Shell, etc.

PAGE NO. :
DATE : / /

cd /usr/share/hadoop/contrib/streaming
↳ hadoop-streaming-1.2.1.jar
(Streaming API program)

cd

hadoop jar /usr/share/hadoop/contrib/
streaming/hadoop-streaming-1.2.1.jar
~~passwd grep bash wc~~
-input "/passwd" -mapper "grep bash"
-reducer "wc -l" -output
"/myuser01"

hadoop fs -cat /myuser01/part-*
↳ 6

Map Reducing Programming in Py

Python.

```
s = "vimal:tw:123"
s.split(":")[1]
  ↳ tw
```

```
s.split(":")
['vimal', 'tw', '123']
```

• This will split ":" (seperator)

```
s = "vimal:x:1002:1002::/home/vimal:/bin/bash"
s.split(":")[2]
  ↳ '1002' (string)
```

PAGE NO. :
DATE : / /

if int(user[2]) >= 1000 and user[6]=='/bin/bash':
    print ("user is general")

fh=open("/etc/passwd", "r")
for i in fh:
    print(i,)
    # removes the empty lines.
fh.close()

fh= open("/etc/passwd", "r")
for i in fh:
    print (i.strip())
fh.close()

#mapper.py

import sys
sys.stdin
fh = sys.stdin

#fh = open("
for i in fh:
    i = i.strip
    user = i.split(":")
    uid = user[2]
    shell = user[6]
    if int( ) >= 1000 and
        shell == "/bin/bash":
        print(i)
fh.close()

PAGE NO. :
DATE : / /

- In filtering algo we always need to traverse the whole program.
  ∴ Best case = Worst case = Avg case = O(n)

cat /etc/passwd | python mapper.py

vim reducer.py

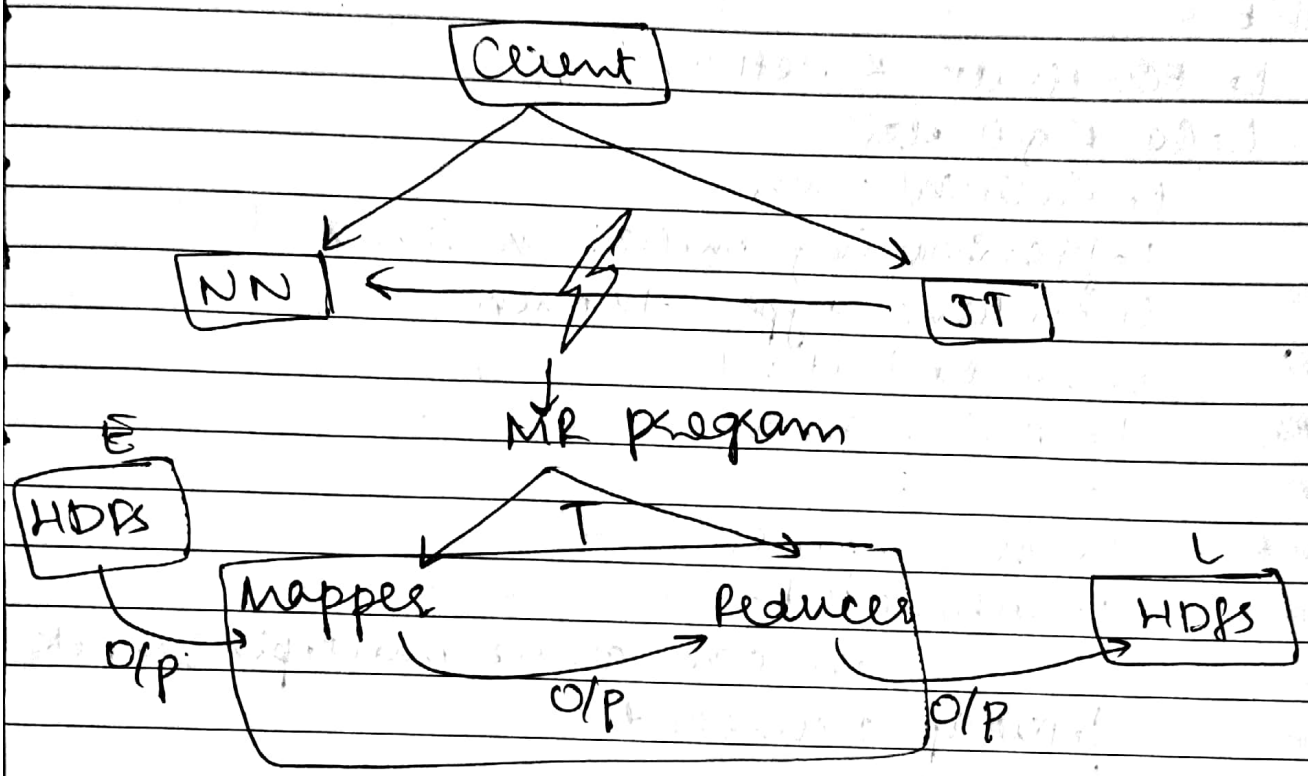
```
import sys
fh = sys.stdin
j = 0
for i in fh:
    j += 1
print(j)
EOF
```

cat /etc/passwd | python mapper.py | python reducer.py

hadoop jar /usr/hadoop/——/——.jar -input "/passwd" -file mapper.py -mapper "python mapper.py" -file reducer.py -reducer "python reducer.py" -output "/op2"

PAGE NO. :
DATE : 27/09/24

SESSION: 152 Hadoop



- ETL process

- Mapper program always run in parallel (3TT → 3mapper program run in parallel).

```
# cat db.txt | tr -s ' ' '\n'
```

↳ all the words come in one line

eg: this is w

→ this
is
w

```
# cat db.txt | tr -s ' ' '\n' |
        sort | uniq -c
```

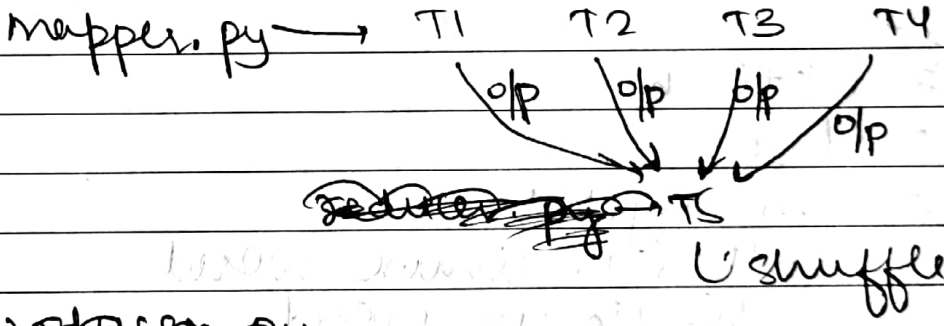
↳ 1 from
3 linux
3 world

PAGE NO.:
DATE: / /

* For word count
↳ mapper.py
- word is the key
- ∴ split each word to dif. line.

eg: linux 1
word 1
hi 2
this 3
} frequency

- It is imp to sort the data first before running the cmd: uniq -c bec it will make our program more optimized.



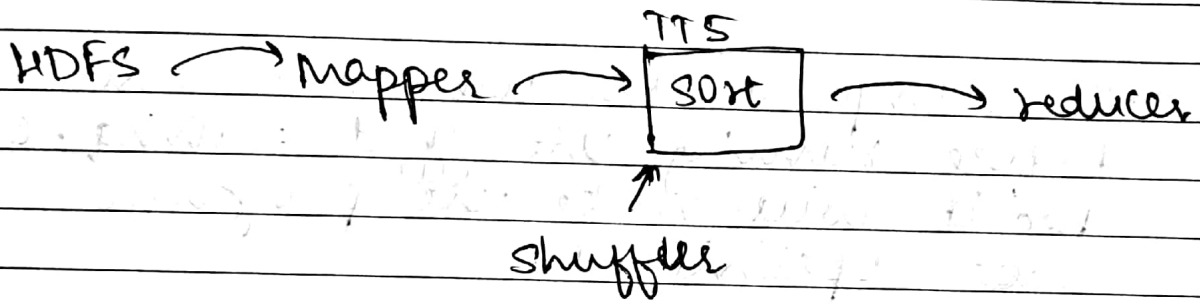
~~reducer.py~~
↳ shuffle

- we already have a sorting algo in hadoop.

Hadoop: sorting algo (mapper: quick sort)
(reducer: merge sort)

PAGE NO. :
DATE : / /

eg: is |
linux |
linux |
linux |
world |
world |
world |



client

```
# mkdir ws
# cd ws
# vim my.txt
  this is linux world
  hi linux world
  viral from linux world
  hi
```

```
# hadoop fs -put my.txt /
# hadoop fs -ls /
```

```
# vim m.py
  #!/usr/bin/python
  import sys
  for sys. stdin
```

PAGE NO. :
DATE : / /

```
for line in sys.stdin:
    linelist = line.split()
    for word in linelist:
        print word, 1    # without ().
```

chmod +x m.py

vim r.py (TEMP.)

```
#!/usr/bin/python
import sys

for i in sys.stdin:
    print i.strip()
```

chmod +x r.py

hadoop jar /usr/hadoop/contrib/sh__/
had_jar -input /my.txt -file m.py -mapper
m.py -file r.py -reducer r.py -output /sl

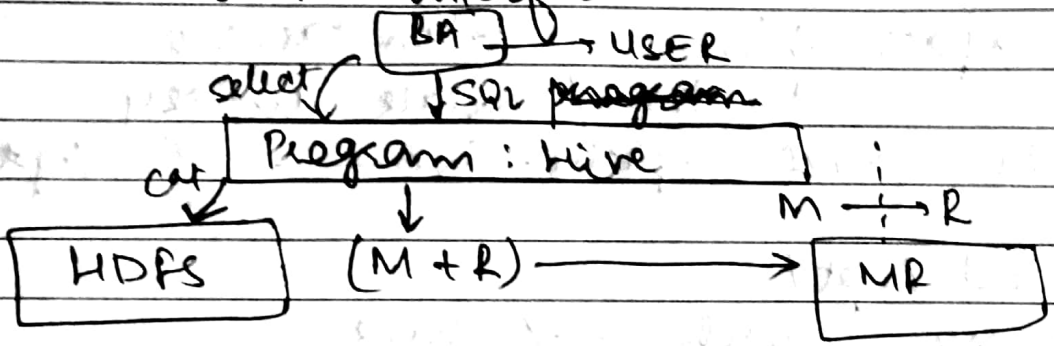
hadoop fs -cat /sl/part*
↳ automatically our data is sorted.

PAGE NO.:
DATE 29 Apr 21

Session: 154 Hadoop

Hive

↳ Just a interface



↳ Generally we use MR program in Analysis.
where we can divide a problem in two parts: Mapper & Reducer.

- Nobody use hadoop directly. They always use intermediate program which automatically converts the program into Mapper & Reducer program.
- This intermediate program is Hive. Or a layer bet. client and hadoop.
- Hive should be installed in the client.

PAGE NO. :
DATE : / /

client

hadoop fs -ls /

apache-hive-0.13.1-bin.tar.gz
↳ drive/install directory

tar -xvzf apache-hive-___ -C /opt/

cd /opt/

mv apach___/ hive

cd hive/

↳ bin

↳ all the cmd are here.

export HIVE_HOME=/opt/hive/

export PATH=/opt/hive/bin/:$PATH

vim /root/.bashrc

↳ export HIVE_HOME=___

export PATH=___

- Generally we store raw data in hadoop but for analysis we need metadata/schema.

- Hive is a meta data server. It only creates/stores the schema of the data. It picks data from hdfs.

PAGE NO.:
DATE: / /

mkdir /hivews
cd /hivews

hive
↳ Initialises hive

- Hive uses derby to store data base.
- Derby db is very light weight.

hive> show databases;
> create database uw;
↳ Creates a user in hdfs.
hdfs: /user/hive/warehouse/uw.db

> use uw;

- Creating a table in hive means creating metadata + raw data.

> create table myuser(username string, password string, uid int, gid int, comment string, homedir string, shell string) row format delimited fields terminated by ":" lines terminated by "\n" stored as textfile;
↳ schema is created.
↳ folder is created in hdfs
↳ /user/hive/warehouse/uw.db/myuser/

PAGE NO. :
DATE : / /

hive > select * from myuser;
↳ nothing will come bec table is empty.

> load data local inpath "/etc/passwd" into table myuser;
↳ loads the dataset into myuser. table.

> select uid from myuser;

> select * from myuser where shell = "/bin/bash";
↳ converts the query into mapper & reducer program & gives the o/p

> selcd count (*) from myuser where shell :"/bin/bash";

> insert overwrite directory '/ho' selcd count (*) -
↳ stores the output in /ho/ dir.

> desc myuser;
↳ describes the table.

> desc formatted myuser;

PAGE NO.:
DATE: 4 May 21

Session: 158 Hadoop

Hive show
hive> create databases;

↑creates the db in hdfs
↳ /user/hive/warehouse

/opt/apache-hive/conf
↳ we can change the default dir
find ~> <workspace>

hive> create database if not exists lw;
> describe database lw;
↳ gives all the details of lw db.

> create database if not exists lwdb1
comment "for team1" with
dbproperties ('creator' = 'vimal', 'date' = '2021-01-24');

> use lwdb1;
> show tables;

2 Types of Tables in hive

Internal
(Managed Type)
• hive have power over data

External
• If the table is deleted the raw data won't be removed.

vim db.csv

1, cricket, india, 1111
2, xyz, us, 2222

- If we have multiple values in a field.
  - 1, cricket:adv:cooking, india, 111
    - we have to change as delimiter eg: space or colon(:).
  - 2, cooking:xyz:abc, us, 2222
  - 3, studing:creating:abc, india, 3333

if not exists
> create table table1 (id int, hobbies array<string> country string, mob int) row format delimited fields terminated by ","
lines terminated by "\n", collection items terminated by ":" stored as textfile; (at the last)

>show tables;
>describe table1;

>load data local inpath '/hivesvs/data/db.csv' into table table1;
- Data is loaded onto the table
>select * from table1;
- prints the table.

- If want to print the column name then go to the conf file and change show columns from false to true

PAGE NO.:
DATE: / /

external table

> create external table table2 (id ———
———) ——— stored as textfile;

> load data ——— ———;

> drop table table2;
↳ The table2 & its schema is removed by the raw data is still intact.
↳ Now if we recreate the table all the data will come up without loading it.

#hadoop fs -mkdir /hr
#hadoop fs -put db.csv /hr

> create database if not exists hrdb;
> create external table hrtable (———) ———
——— stored as textfile location '/hr';

↳ table created
> describe table hrtable;
↳ external table
↳ mapped to /hr from hdfs
↳ automatically picks data from hdfs.

> select country, mob from hrtable;

PAGE NO. :
DATE : / /

> create table newtable (country string, phone int) store as textfile;
> select * from newtable;

> insert into table newtable select country, mob from table3 where country = 'india'

↳ first it runs and runs and the o/p is stored in newtable.

ELT : Extract Load & Transform
↳ Hadoop is used ELT operation.