• **Load DATA through pipeline**

nano /usr/local/hadoop/etc/hadoop/core-site.xml

<property>

<name>fs.s3n.awsAccessKeyId</name>

<value>{AWS\_ACCESS\_KEY}</value>

</property>

<property>

<name>fs.s3n.awsSecretAccessKey</name>

<value>{AWS\_SECRET\_KEY}</value>

</property>

hadoop fs -cp s3n://<bucket\_name>/<data\_set> dataset

hadoop fs -cp dataset s3n://<bucket\_name>/<data\_set>

s3n://datasetz/purchases.csv

/usr/local/hadoop/share/hadoop/tools/lib\*

nano .bashrc

export HADOOP\_CLASSPATH=$HADOOP\_CLASSPATH:$HADOOP\_HOME/share/hadoop/tools/lib/\*

ubuntu@nn:~$ hadoop fs -cp s3n://shaan/index.pdf /

ubuntu@nn:~$ hdfs dfs -ls /

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**DISTCP**

hadoop distcp hdfs://ip-172-31-10-128.ap-south-1.compute.internal:9001/user/ubuntu/BigData.tsv hdfs://ip-172-31-17-55.ap-south-1.compute.internal:9001/user/ubuntu/

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

**CREATE SNAPSHOTS**

hadoop fs -mkdir snapshot

hdfs dfsadmin -allowSnapshot /user/ubuntu/snapshot

hdfs dfsadmin -disallowSnapshot <path>

hdfs dfs -createSnapshot /user/ubuntu/snapshot/

hdfs dfs -deleteSnapshot <path> <snapshotName>

hdfs dfs -renameSnapshot <path> <oldName> <newName>

hdfs lsSnapshottableDir

hdfs snapshotDiff /user/ubuntu/snapshot s20160820-000747.522 s20160820-000825.861

**Configure Backup Node**

<property>

<name>dfs.namenode.backup.address</name>

<value>newnode:50100</value>

<description>

The backup node server address and port.

If the port is 0 then the server will start on a free port.

</description>

</property>

<property>

<name>dfs.namenode.backup.http-address</name>

<value>s3n://hdfsbackupnode:0</value>

<description>

The backup node http server address and port.

If the port is 0 then the server will start on a free port.

</description>

</property>

hdfs getconf -backupNodes

hdfs namenode -backup

hdfs namenode -checkpoint

**Decommisioning Node**

nano excludes

nn

hdfs-site.xml

<property>

<name>dfs.hosts.exclude</name>

<value>/usr/local/hadoop/etc/hadoop/excludes</value>

<final>true</final>

</property>

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

yarn-site.xml

<property>

<name>mapred.hosts.exclude</name>

<value>/usr/local/hadoop/conf/excludes</value>

<final>true</final>

</property>

hadoop dfsadmin -refreshNodes

hadoop mradmin -refreshNodes

hadoop dfsadmin -report

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Commissioning Nodes**

vi includes

ip-172-31-30-159.eu-west-1.compute.intern

update the Nodes in slaves file

Remove the Nodes from exclude file

hdfs-site.xml

<property>

<name>dfs.hosts.include</name>

<value>/usr/local/hadoop/etc/hadoop/includes</value>

<final>true</final>

</property>

mapred-site.xml

<property>

<name>mapred.hosts.include</name>

<value>/usr/local/hadoop/etc/hadoop/includes</value>

<final>true</final>

</property>

hadoop mradmin -refreshNodes

hdfs balancer

hadoop dfsadmin -report > report\_aug

hadoop dfsadmin -report

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**• system file checker**

hadoop fsck -location -block -size

hdfs fsck / -locations -blocks -files

hdfs fsck -list-corruptfileblocks

hdfs dfs -rm /user/ubuntu/purchases.csv (example that this is a corrupt)  
hdfs fsck / -delete

hadoop fsck / > /home/ubuntu/

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**• Trash configuration**

nano core-site.xml

<property>

<name>fs.trash.interval</name>

<value>30</value>

<description>Number of minutes between trash checkpoints. If zero, the trash feature is disabled</description>

</property>

<property>

<name>fs.trash.checkpoint.interval</name>

<value>15</value>

</property>

hadoop fs -expunge ---------- remove file from trash

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

nano hdfs-site.xml

**Block Size**

<property>

<name>dfs.block.size</name>

<value>524288</value>

</property>

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_--

**Hadoop Archivals**

hadoop archive -archiveName positive.har -p /user/ubuntu/filename /user/ubuntu/

hadoop fs -ls

hadoop fs -lsr positive.har

hadoop fs -lsr har:///user/ubuntu/positive.har

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

hadoop version

hadoop fs -ls / ---------to check hadoop root

hadoop fs -du hdfs:/ -------------disk usage

hadoop fs -count hdfs:/ --------- file count

**Setting Replication for a dataset**

hdfs dfs -setrep -w 4 purchases.csv

**Setting permissions**

hadoop fs -chmod 600 path

**Adding users in ubuntu**

sudo adduser dfs

sudo adduser dfs sudo

hadoop fs -mkdir /user/dfs

hadoop fs -chown dfs:supergroup /user/dfs

cp /home/ubuntu/.bashrc ~/

bash