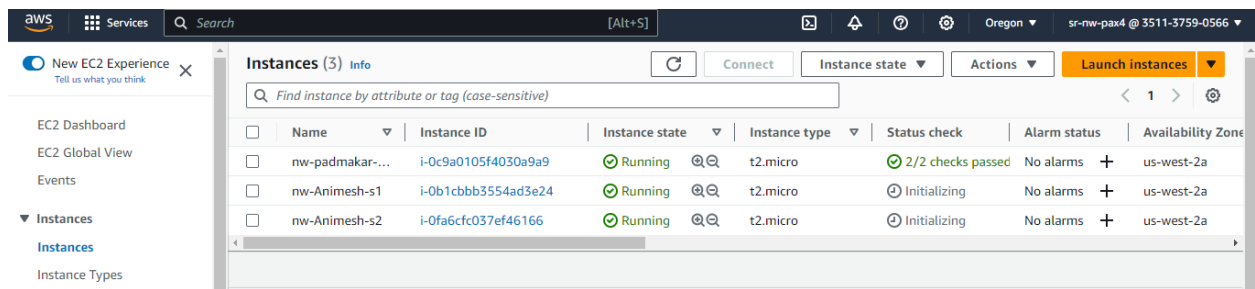


Multi node cluster:

1) Setup an instance on cloud



2) Download a jdk on the instance - jdk-17 (sudo apt update , sudo apt install openjdk-8-jdk)

3) Checking version

```
1 error
ubuntu@ip-172-31-17-18:~$ java -version
openjdk version "17.0.8.1" 2023-08-24
OpenJDK Runtime Environment (build 17.0.8.1)
```

4) Passphraseless connection setup

a) ssh-keygen

b) Add new public key to the authorized_keys file.

```
ubuntu@ip-172-31-23-125:~$ ls
ubuntu@ip-172-31-23-125:~$ ssh-keygen
ssh: Could not resolve hostname keygen: Temporary failure in name resolution
ubuntu@ip-172-31-23-125:~$ ssh-keygen
Generating public/private rsa key pair.
Enter file in which to save the key (/home/ubuntu/.ssh/id_rsa):
Enter passphrase (empty for no passphrase):
```

```
OBwH3uVn+30bcXG230p1OKCEFzjCfA7RWcnYEA8LuQZVfRnfN9zsYo4tbUCcUKBYaC7q
ip-172-31-17-18
ubuntu@ip-172-31-17-18:~/.ssh$ cat id_rsa.pub >> authorized_keys
ubuntu@ip-172-31-17-18:~/.ssh$
```

5) Installing dsh (`sudo apt install dsh -y`),Download hadoop > 1.2.1

```
buntu@ip-172-31-17-18:/etc/dsh$ cd
buntu@ip-172-31-17-18:~$ dsh -a wget https://archive.apache.org/dist/hadoop/core/hadoop-1.2.1/hadoop-1.2.1.tar.gz
-2023-10-03 06:55:23-- https://archive.apache.org/dist/hadoop/core/hadoop-1.2.1/hadoop-1.2.1.tar.gz
```

6) Checking version and moving file

```
ubuntu@ip-172-31-23-125:~$ dsh -a ls
hadoop-2.5.0
hadoop-2.5.0.tar.gz
hadoop-2.5.0
hadoop-2.5.0.tar.gz
ubuntu@ip-172-31-23-125:~$ dsh -a sudo mv hadoop-2.5.0 /usr/local/hadoop/
ubuntu@ip-172-31-23-125:~$ dsh -a ls
hadoop-2.5.0.tar.gz
hadoop-2.5.0.tar.gz
ubuntu@ip-172-31-23-125:~$
```

7) configure the files

Mapred.xml

```
<property>
<name>mapred.job.tracker</name>
<value>hdfs://localhost:9001</value>
</property>
```

```
... The site specific property override

<configuration>
<property>
<name>mapred.job.tracker</name>
<value>hdfs://dn:9001</value>
</property>
```

hdfs-site.xml

```
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
```

```
<!-- Put site-specific property o

<configuration>
<property>
<name>dfs.replication</name>
<value>3</value>
</property>
</configuration>
```

nano .bashrc

```
GNU nano 4.8                               .bashrc
fi

# enable programmable completion features (you don't need to enable
# this, if it's already enabled in /etc/bash.bashrc and /etc/profile
# sources /etc/bash.bashrc).
if ! shopt -oq posix; then
  if [ -f /usr/share/bash-completion/bash_completion ]; then
    . /usr/share/bash-completion/bash_completion
  elif [ -f /etc/bash_completion ]; then
    . /etc/bash_completion
  fi
fi
export HADOOP_PREFIX=/usr/local/hadoop/
export PATH=$PATH:$HADOOP_PREFIX/bin
export HADOOP_HOME=/usr/local/hadoop/
export PATH=$PATH:$HADOOP_HOME/sbin
export JAVA_HOME=/usr/lib/jvm/java-17-openjdk-amd64
export PATH=$PATH:$JAVA_HOME
```

core-site.xml

```
<property>
<name>fs.default.name</name>
<value>hdfs://localhost:9000</value>
</property>
<property>
<name>hadoop.tmp.dir</name>
<value>/usr/local/hadoop/tmp</value>
</property>
```

Hadoop.env

```
export JAVA_HOME=/usr/lib/jvm/java-17-openjdk-amd64
export HADOOP_OPTS=-Djava.net.preferIPv4Stack=true
```

8) Format namenode

ubuntu@ip-172-31-25-39: /usr/local/hadoop

```
23/10/03 13:27:10 INFO util.GSet: VM type          = 64-bit
23/10/03 13:27:10 INFO util.GSet: 0.029999999329447746% max memory 966.7 MB =
7.0 KB
23/10/03 13:27:10 INFO util.GSet: capacity        = 2^15 = 32768 entries
23/10/03 13:27:10 INFO namenode.NNConf: ACLs enabled? false
23/10/03 13:27:10 INFO namenode.NNConf: XAttrs enabled? true
23/10/03 13:27:10 INFO namenode.NNConf: Maximum size of an xattr: 16384
Re-format filesystem in Storage Directory /usr/local/hadoop/tmp/dfs/name ? (Y
N) y
23/10/03 13:27:16 INFO namenode.FSImage: Allocated new BlockPoolId: BP-2130924
6-172.31.25.39-1696339636536
23/10/03 13:27:16 INFO common.Storage: Storage directory /usr/local/hadoop/tmp
fs/name has been successfully formatted.
23/10/03 13:27:16 INFO namenode.NNStorageRetentionManager: Going to retain 1 i
ges with txid >= 0
23/10/03 13:27:16 INFO util.ExitUtil: Exiting with status 0
23/10/03 13:27:16 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at ip-172-31-25-39/172.31.25.39
```

9) Checking output on both the server

```
ubuntu@ip-172-31-23-93: /usr/local/hadoop
starting namenodes on [nn]
a: starting namenode, logging to /usr/local/hadoop/logs/hadoop-ubuntu-namenode-
ip-172-31-23-93.out
a: starting datanode, logging to /usr/local/hadoop/logs/hadoop-ubuntu-datanode-
ip-172-31-18-186.out
starting secondary namenodes [0.0.0.0]
Warning: Permanently added '0.0.0.0 (0.0.0.0)' can't be established.
ECDSA key fingerprint is SHA256:LPDqgLqEan3O+h9diKCLFfjvsWwQlZuUdpTIvn60lSo.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
0.0.0.0: Warning: Permanently added '0.0.0.0' (ECDSA) to the list of known hosts

0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-ub
untu-secondarynamenode-ip-172-31-23-93.out
ubuntu@ip-172-31-23-93:/usr/local/hadoop$ jps
0467 SecondaryNameNode
0250 NameNode
0570 Jps
ubuntu@ip-172-31-23-93:/usr/local/hadoop$ dsh -a jps
0658 Jps
0467 SecondaryNameNode
0250 NameNode
0757 Jps
0621 DataNode
ubuntu@ip-172-31-23-93:/usr/local/hadoop$
```

just 2 commands. Guys.

10) Checking result on Ui browser

NameNode '127.0.0.1:9000'

Started: Tue Oct 03 07:57:41 UTC 2023
Version: 1.2.1, r1503152
Compiled: Mon Jul 22 15:23:09 PDT 2013 by mattf
Upgrades: There are no upgrades in progress.

[Browse the filesystem](#)
[Namenode Logs](#)

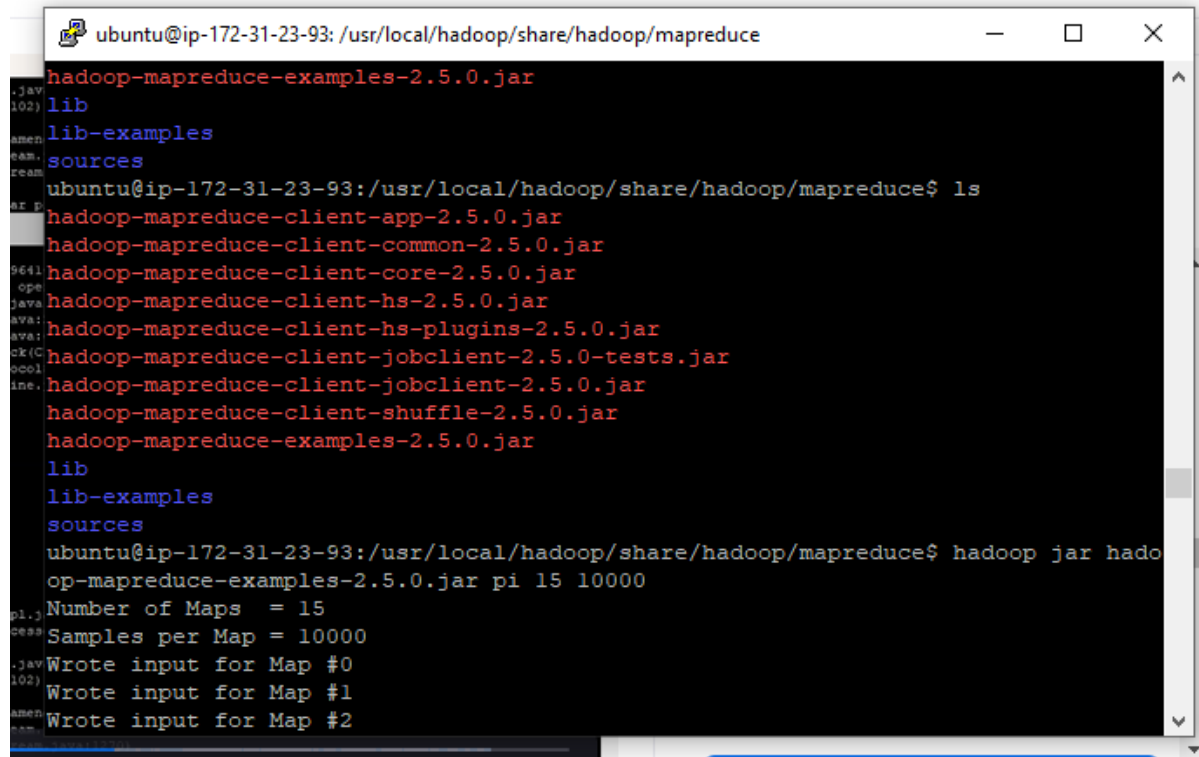
Cluster Summary

8 files and directories, 1 blocks = 9 total. Heap Size is 23.57 MB / 966.69 MB (2%)

Configured Capacity	:	7.57 GB
DFS Used	:	40 KB
Non DFS Used	:	2.9 GB
DFS Remaining	:	4.67 GB
DFS Used%	:	0 %
DFS Remaining%	:	61.68 %
Live Nodes	:	1
Dead Nodes	:	0
Decommissioning Nodes	:	0
Number of Under-Replicated Blocks	:	0

11) Submitting the job

Full screen with speaker view



```
ubuntu@ip-172-31-23-93: /usr/local/hadoop/share/hadoop/mapreduce
hadoop-mapreduce-examples-2.5.0.jar
lib
lib-examples
sources
ubuntu@ip-172-31-23-93:/usr/local/hadoop/share/hadoop/mapreduce$ ls
hadoop-mapreduce-client-app-2.5.0.jar
hadoop-mapreduce-client-common-2.5.0.jar
hadoop-mapreduce-client-core-2.5.0.jar
hadoop-mapreduce-client-hs-2.5.0.jar
hadoop-mapreduce-client-hs-plugins-2.5.0.jar
hadoop-mapreduce-client-jobclient-2.5.0-tests.jar
hadoop-mapreduce-client-jobclient-2.5.0.jar
hadoop-mapreduce-client-shuffle-2.5.0.jar
hadoop-mapreduce-examples-2.5.0.jar
lib
lib-examples
sources
ubuntu@ip-172-31-23-93:/usr/local/hadoop/share/hadoop/mapreduce$ hadoop jar hadoop-mapreduce-examples-2.5.0.jar pi 15 10000
Number of Maps = 15
Samples per Map = 10000
Wrote input for Map #0
Wrote input for Map #1
Wrote input for Map #2
```

12) Terasort


```
ubuntu@ip-172-31-23-93: /usr/local/hadoop/share/hadoop/mapreduce
    Total committed heap usage (bytes)=28114944
org.apache.hadoop.examples.terasort.TeraGen$Counters
    CHECKSUM=74183521510228
File Input Format Counters
    Bytes Read=0
File Output Format Counters
    Bytes Written=3456700
ubuntu@ip-172-31-23-93:/usr/local/hadoop/share/hadoop/mapreduce$ hadoop dfs -ls
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Found 1 items
drwxr-xr-x  - ubuntu supergroup          0 2023-10-05 08:12 input
ubuntu@ip-172-31-23-93:/usr/local/hadoop/share/hadoop/mapreduce$ hadoop jar hadop
op-mapreduce-examples-2.5.0.jar terasort input output
23/10/05 08:15:41 INFO terasort.TeraSort: starting
23/10/05 08:15:43 INFO input.FileInputFormat: Total input paths to process : 1
Spent 163ms computing base-splits.
Spent 3ms computing TeraScheduler splits.
Computing input splits took 172ms
Sampling 1 splits of 1
Making 1 from 34567 sampled records
Computing paritions took 369ms
Spent 542ms computing partitions.
```