

**Hadoop cluster.**

As a matter of fact, Facebook handles 105 petabytes of data every hour, which is roughly 100 million gigabytes.

So, how this huge bulk of data is getting managed. Is it because there is massive hard disk available to store data?

The above problem is known as **#BIGDATA**.

So, bigdata is not a technology. It's a problem we have to overcome with technology.

And some sub-problems of BIGDATA are: VOLUME and VELOCITY:

Let's discuss a scenario:

You know there are companies like dell EMC, who have the technology to make storage drives in respect of exabytes. Although it will be costly but if neglect the matter of cost. Will it be viable to use? The answer is a big NO. Generally speaking, if it takes almost 1 min to copy 1 GB data in a typical SATA hard drive. So, storing petabytes of data in a single storage device will be a calamity. It will take several days.

## Why not use SSD?

The transfer of data will be fast. Yes, indeed. But the concept here is not about deciding which type of storage to be used. It's about the how efficiently we can use the storage! The business of google, Facebook runs on the instant access of data. And we people i.e. including me too, don't want to wait even a second. Forget about a day.

VOLUME represent the problem of how to store the data in an instant.

And VELOCITY represent the problem of how to read that data it an instant and successfully transfer to the requested user.

**The solution is DISTRIBUTED STORAGE: DISCOVERY OF SUPER COMPUTERS.**

We split the incoming data into many smaller parts and parallelly store them in many storage units. This configuration is based on the topology of MASTER-SLAVE MODEL. The setup in which multiple computers are working together is known as CLUSTER. In this design, one module is the master unit, which handles the splitting of incoming and outgoing data from the server and all other units are slaves as storage units.

•

### Requirements:

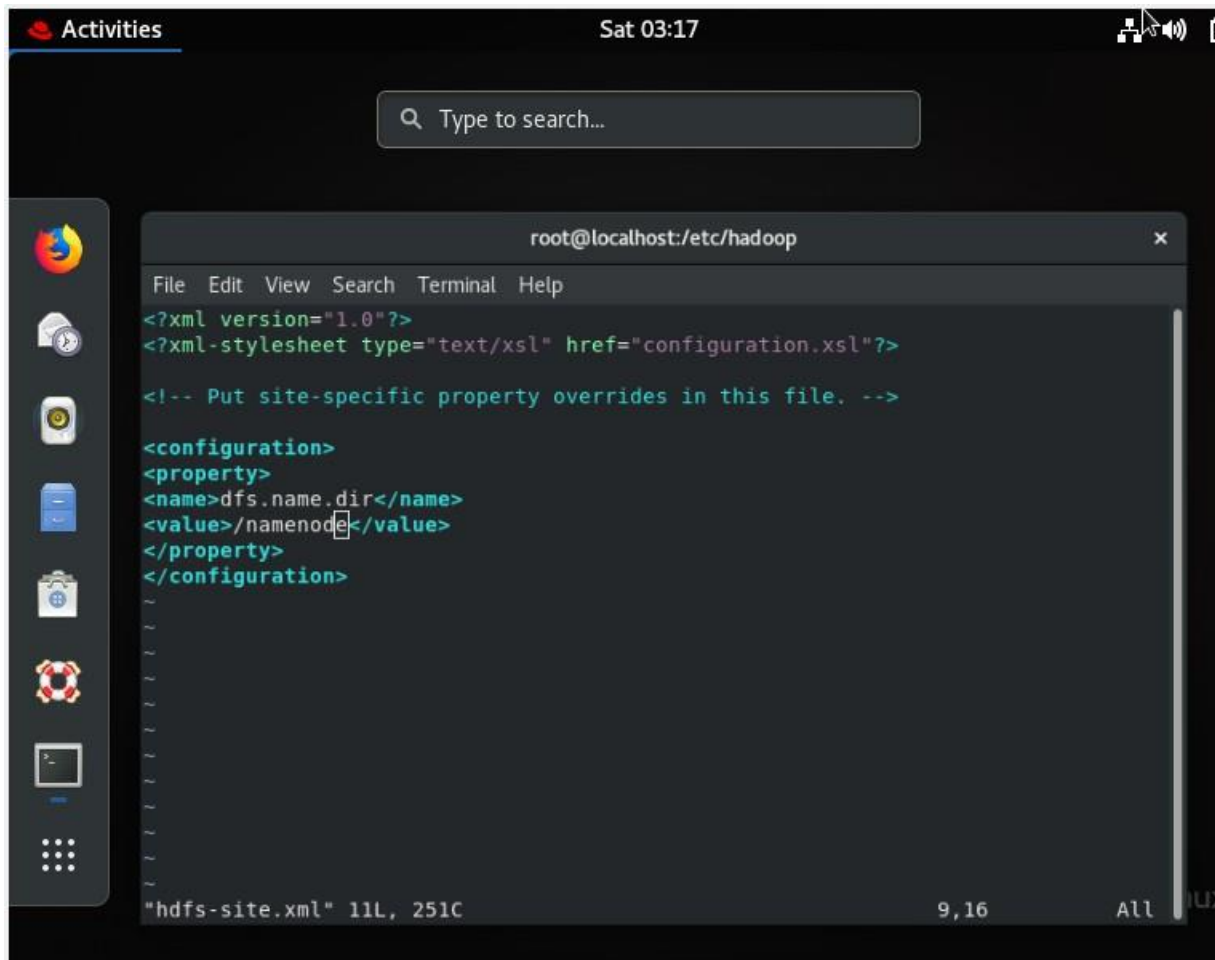
For completing this project, at least 2 laptops or PCs are required. However, I used AWS Instance which offers free 720 hours every month of computation on the basis of PAY AS WE GO MODEL.

Putty.exe to access the instance in your pc.

➤ **Creating Master:**

Select one OS to be your namenode. Check whether the java and Hadoop versions are compatible and successfully running. (java -version  
hadoop version)

|                    |   |                |
|--------------------|---|----------------|
| cd /etc/hadoop     | //get into the hadoop directory                                       |                |
| mkdir /namenode    | //created a directory for the namenode. cd/                           |                |
| pwd/               |   |                |
| ls                 | //check if the namenode directory you created is visible. cd namenode | //get into the |
| namenode directory |   |                |
| vim hdfs-site.xml  | //here we shall set up the namenode as follows:                       |                |



### ➤ Creating Slave:

Go to the OS which will act as the slave. Follow some similar steps to create the datanode.

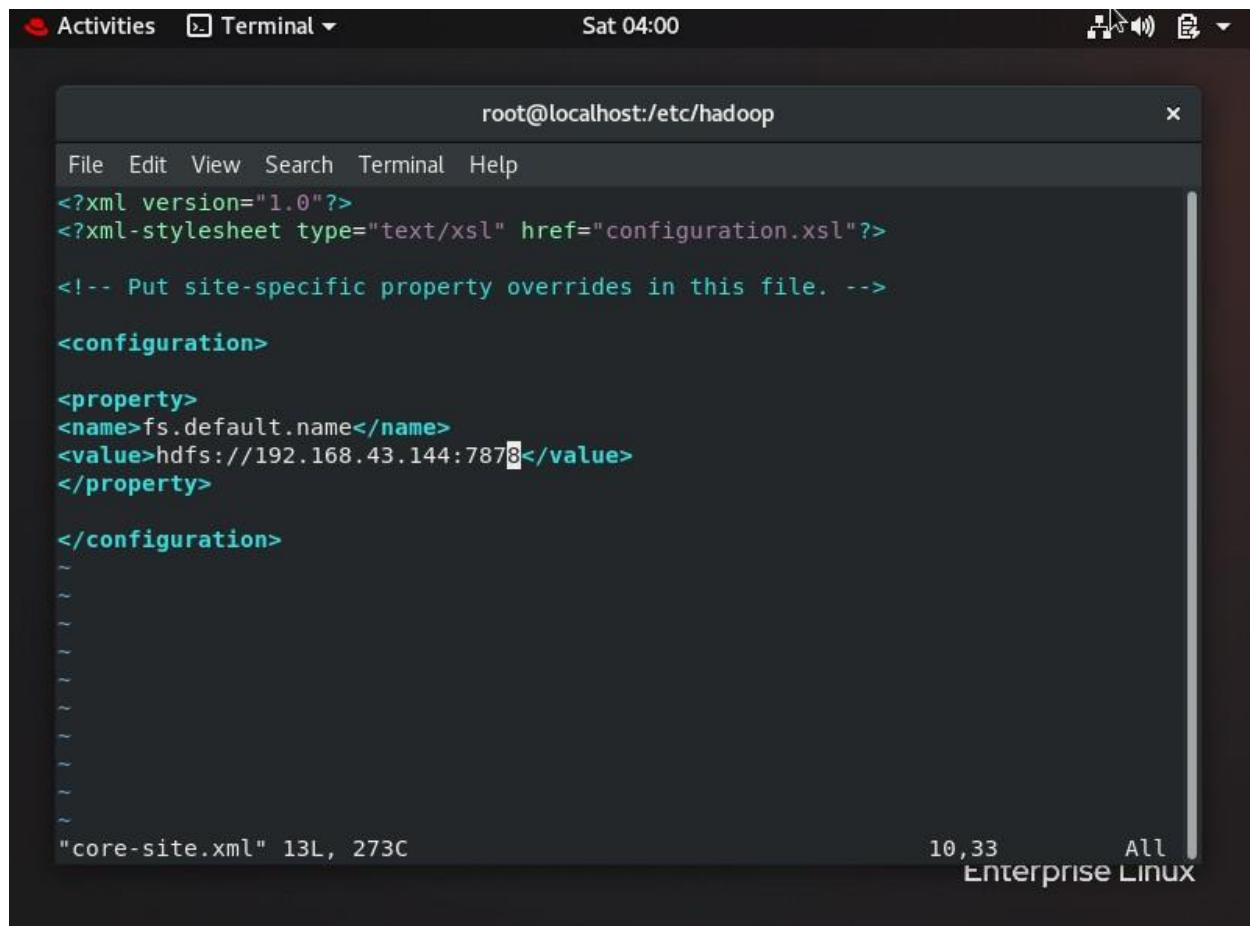
```
mkdir /s1 cd /
```

```
ls
```

```
cd s1/
```

```
vim hdfs-site.xml //be careful while putting the name and value
```





```
root@localhost:/etc/hadoop
File Edit View Search Terminal Help
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<!-- Put site-specific property overrides in this file. -->

<configuration>

  <property>
    <name>fs.default.name</name>
    <value>hdfs://192.168.43.144:7878</value>
  </property>

</configuration>
~
~
~
~
~
~
~
~
~
~
"core-site.xml" 13L, 273C 10,33 All Enterprise Linux
```

Here hdfs is the protocol used to establish connection between master and slave. Ip provided is the ip of the master. 7878 is the port number. Avoid indentation. The configuration of the master node is completed. Before activating it, we need to

format it ones.

Follow the mentioned steps to format and start the namenode. cd

```
hadoop namenode -format
```

```
hadoop-daemon.sh start namenode //namenode is started
```

```
jps //it shows that the namenode of port number 7878 is active. systemctl stop firewalld
```

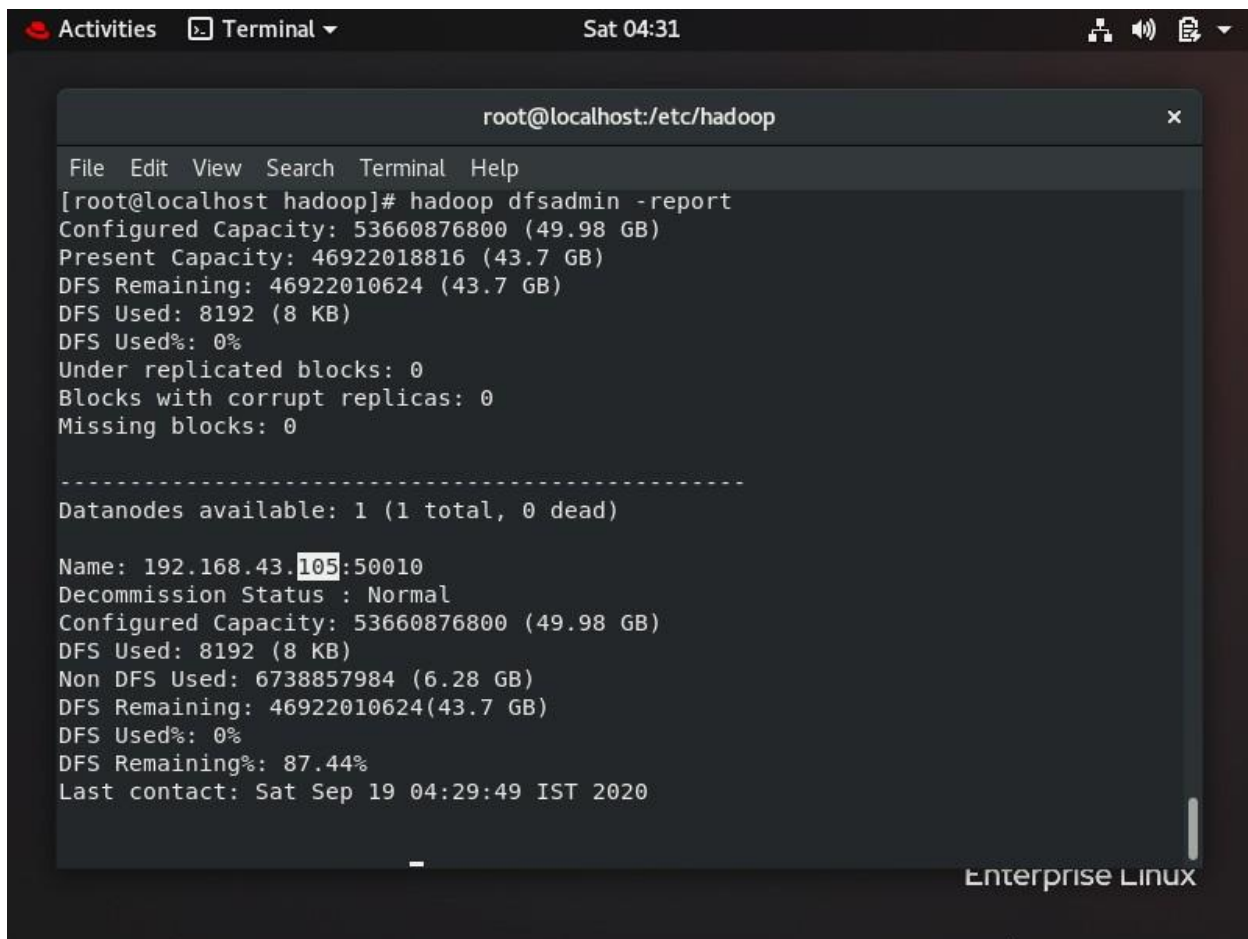
```
//disables firewall to enable slave connection.
```

### ➤ **Activate the slave**

```
cd /etc/hadoop vim core-site.xml
```



hadoop dfsadmin -report



```
root@localhost:/etc/hadoop
File Edit View Search Terminal Help
[root@localhost hadoop]# hadoop dfsadmin -report
Configured Capacity: 53660876800 (49.98 GB)
Present Capacity: 46922018816 (43.7 GB)
DFS Remaining: 46922010624 (43.7 GB)
DFS Used: 8192 (8 KB)
DFS Used%: 0%
Under replicated blocks: 0
Blocks with corrupt replicas: 0
Missing blocks: 0

-----
Datanodes available: 1 (1 total, 0 dead)

Name: 192.168.43.105:50010
Decommission Status : Normal
Configured Capacity: 53660876800 (49.98 GB)
DFS Used: 8192 (8 KB)
Non DFS Used: 6738857984 (6.28 GB)
DFS Remaining: 46922010624(43.7 GB)
DFS Used%: 0%
DFS Remaining%: 87.44%
Last contact: Sat Sep 19 04:29:49 IST 2020
```

It shows that the datanode available is 1. It also shows the ip of the datanode. Similarly, you can add many slaves to the master and increase its efficiency. This makes the implementation of hadoop distributive storage cluster successful.