

PRACTICAL NO 1

Aim: Install, configure and run Hadoop and HDFS and explore HDFS

Steps:

1) To install Hadoop, you should have Java version in your system Check your java version through the below given command in command prompt

(Link: <https://www.oracle.com/in/java/technologies/javase/javase8-archive-downloads.html>)

Command:

```
java -version
```

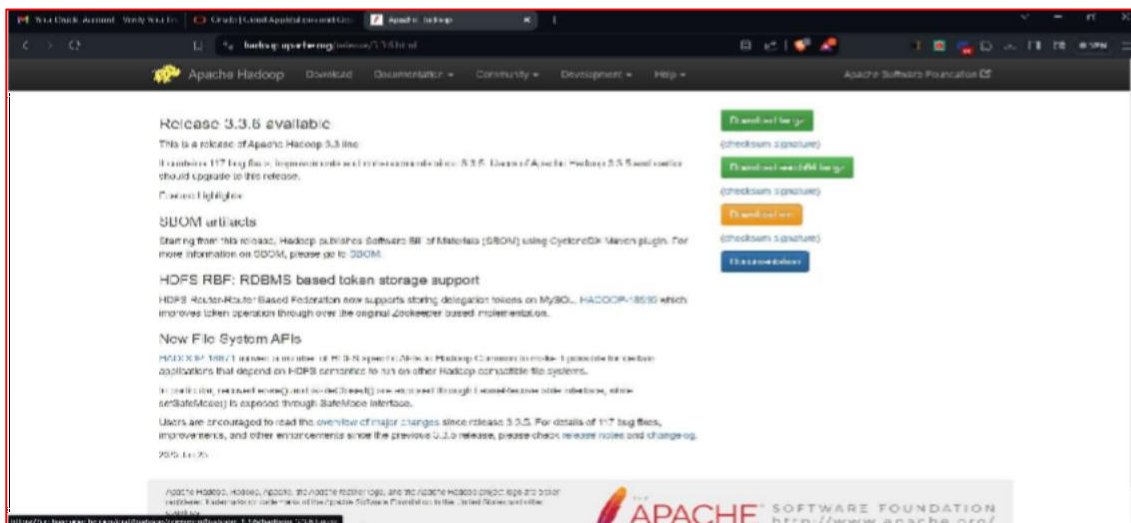
Output:

```
Microsoft Windows [Version 10.0.22621.3374]
(c) Microsoft Corporation. All rights reserved.

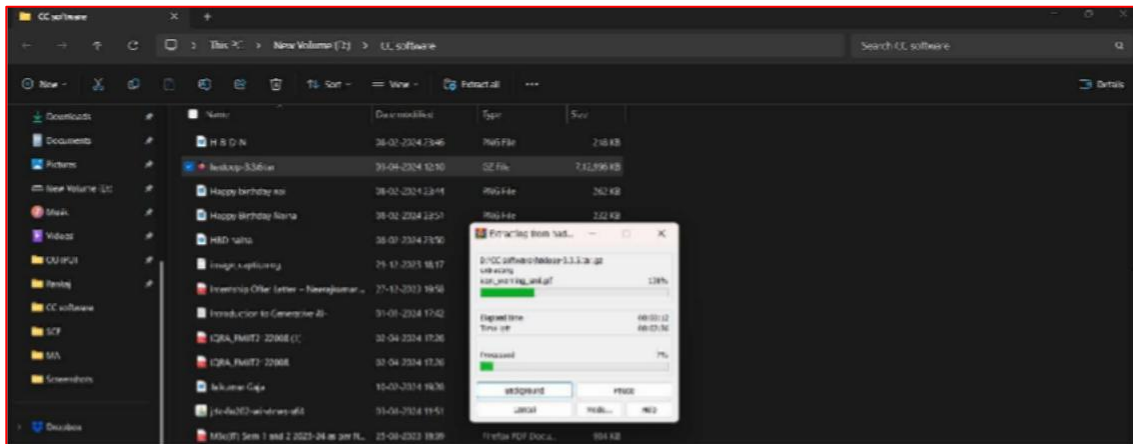
C:\Users\acer>java -version
java version "1.8.0_202"
Java(TM) SE Runtime Environment (build 1.8.0_202-b08)
Java HotSpot(TM) 64-Bit Server VM (build 25.202-b08, mixed mode)

C:\Users\acer>
```

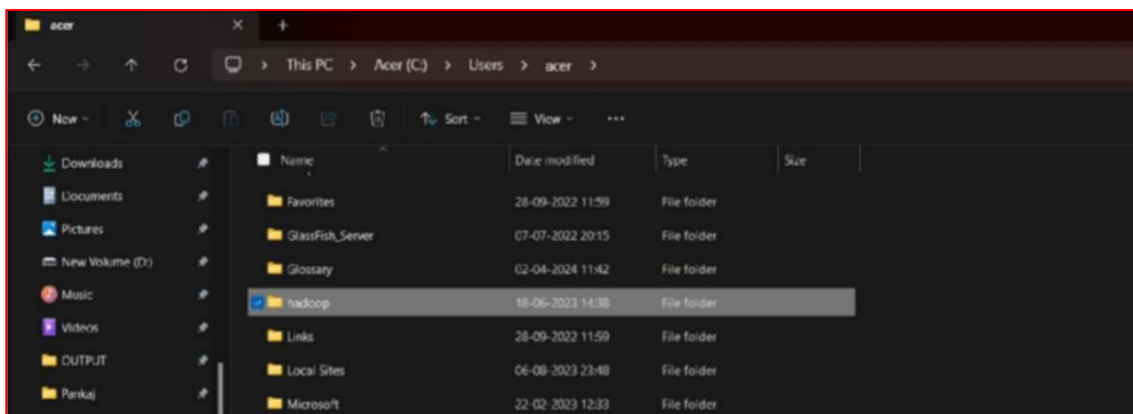
2) After downloading java version 1.8, download Hadoop version 3.3.6 from the given link. (Link: <https://hadoop.apache.org/release/3.3.6.html>)



3) Extract Hadoop to a local drive



4) Rename it “hadoop”.

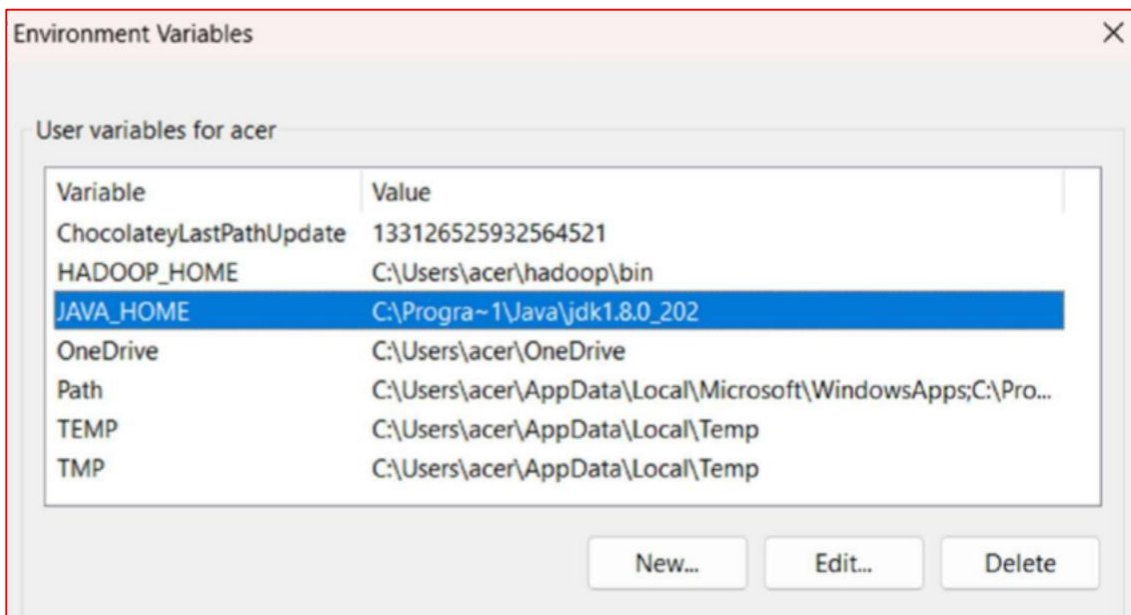
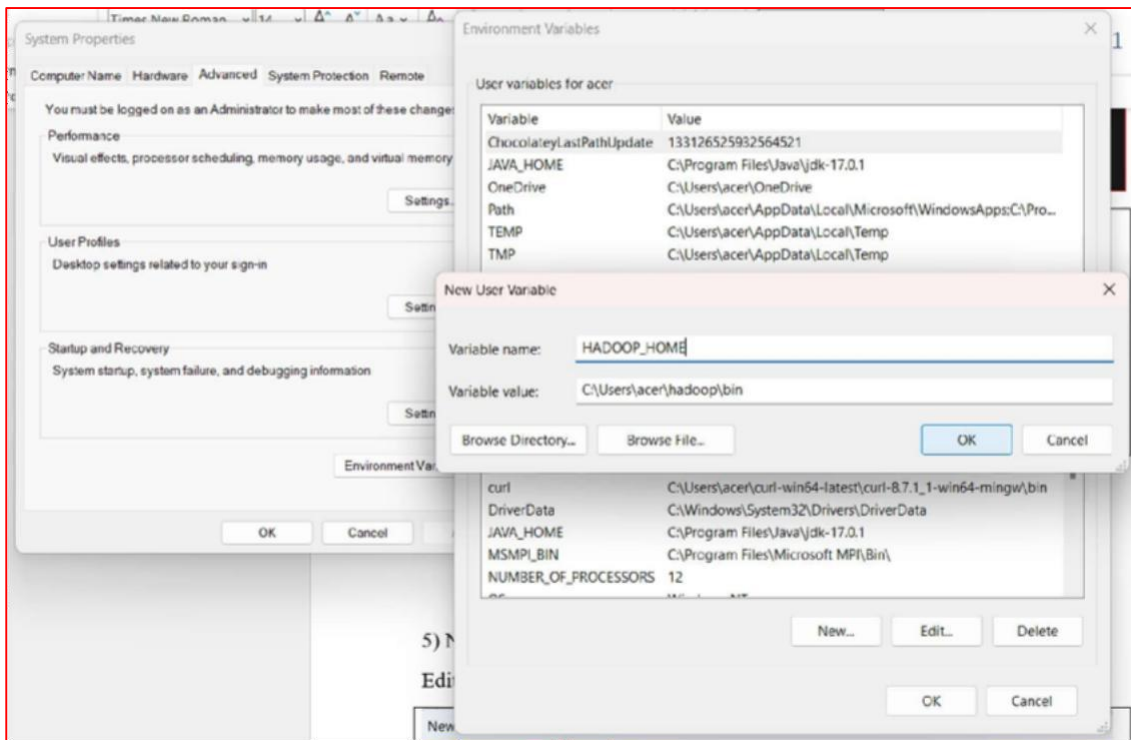


5) Now set the path open environment variables

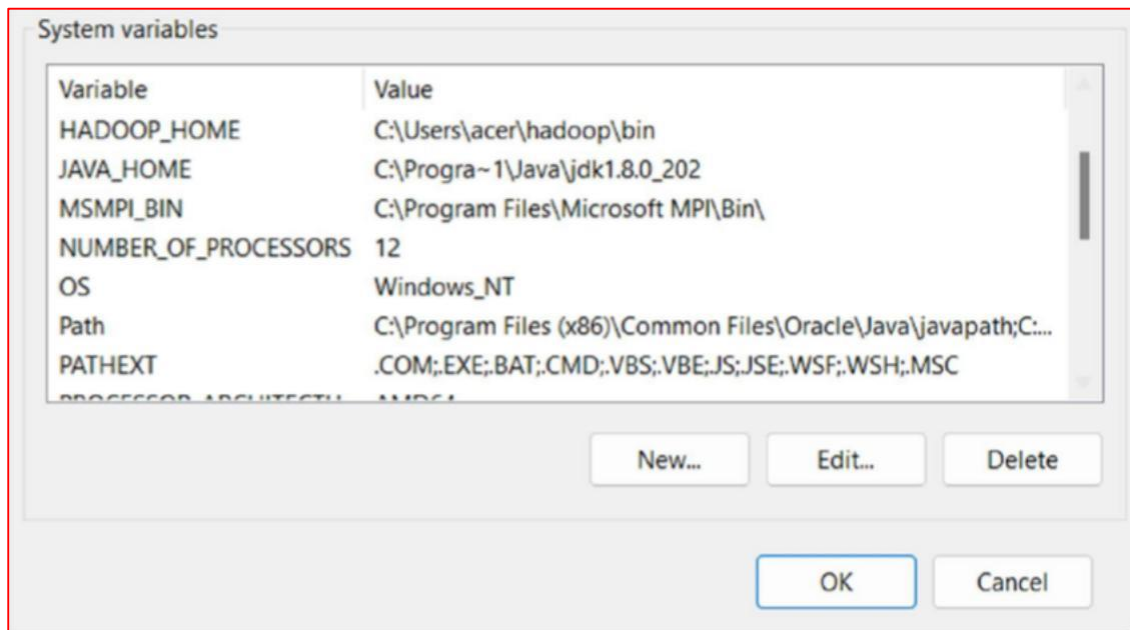
Search Edit the system environment variable on windows and click on it >>
Click on Environment Variable >> Click on New >> Add name as
HADOOP_HOME and paste the location as (C:\Users\acer\hadoop\bin) value
Do the same for both user variable as well as system variable and add both
HADOOP_HOME and JAVA_HOME.

- C:\Users\acer\hadoop\bin (also add in system variable path)
- C:\Users\acer\hadoop\sbin (also add in system variable path in order to run the start-dfs.cmd command properly)
- C:\Progra~1\Java\jdk1.8.0_202

Note:- for JAVA_HOME set the value as C:\Progra~1\Java\jdk1.8.0_202 (i.e. Program Files is replaced as Progra~1) otherwise it will give you error in command prompt when you will try to execute hadoop version.

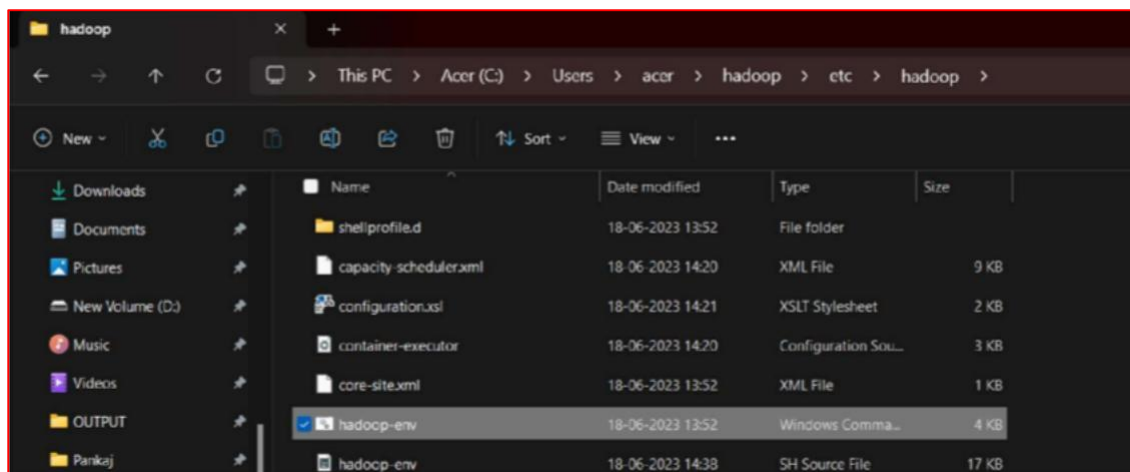


6) Now set the system variables

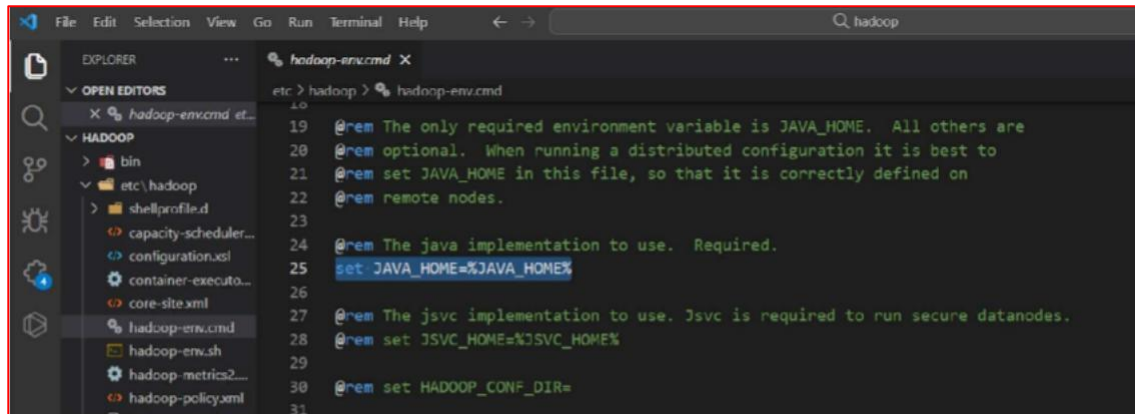


7) Now go to “**hadoop**” folder open “**etc**” folder in it then open “**hadoop**” folder inside it and then select the “**hadoop-env**” windows command open it and make the in changes according to below given code.

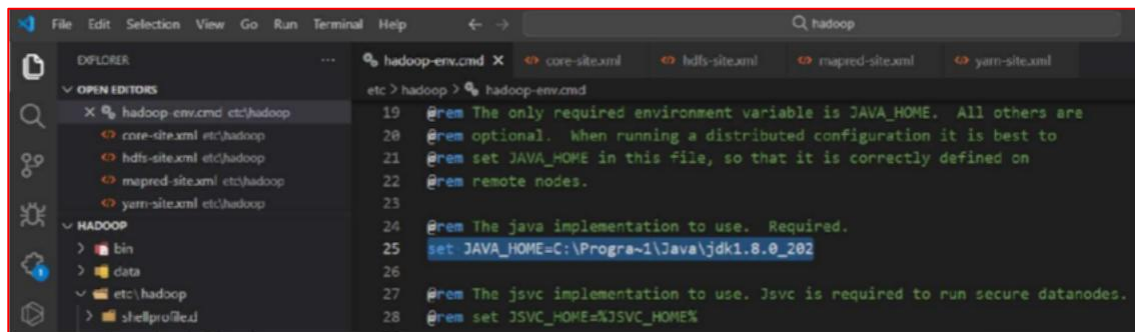
Note:- Open the file using notepad or any other editor here we have used visual studio code.



8) Edit **hadoop-env.cmd** and replace **%JAVA_HOME%** with the path of the java folder where your jdk 1.8 is installed i.e. **C:\Program Files\Java\jdk1.8.0_202**.



```
etc > hadoop > hadoop-env.cmd
19 @rem The only required environment variable is JAVA_HOME. All others are
20 @rem optional. When running a distributed configuration it is best to
21 @rem set JAVA_HOME in this file, so that it is correctly defined on
22 @rem remote nodes.
23
24 @rem The java implementation to use. Required.
25 set JAVA_HOME=%JAVA_HOME%
26
27 @rem The jsvc implementation to use. Jsvc is required to run secure datanodes.
28 @rem set JSVC_HOME=%JSVC_HOME%
29
30 @rem set HADOOP_CONF_DIR=
31
```



```
etc > hadoop > hadoop-env.cmd
19 @rem The only required environment variable is JAVA_HOME. All others are
20 @rem optional. When running a distributed configuration it is best to
21 @rem set JAVA_HOME in this file, so that it is correctly defined on
22 @rem remote nodes.
23
24 @rem The java implementation to use. Required.
25 set JAVA_HOME=C:\Program Files\Java\jdk1.8.0_202
26
27 @rem The jsvc implementation to use. Jsvc is required to run secure datanodes.
28 @rem set JSVC_HOME=%JSVC_HOME%

```

9) Make the changes in code of below given files by going to the “etc” folder in “hadoop”

- i. core-site.xml
- ii. hdfs-site.xml
- iii. mapred-site.xml
- iv. yarn-site.xml

i) For core-site.xml

Code:

<configuration>

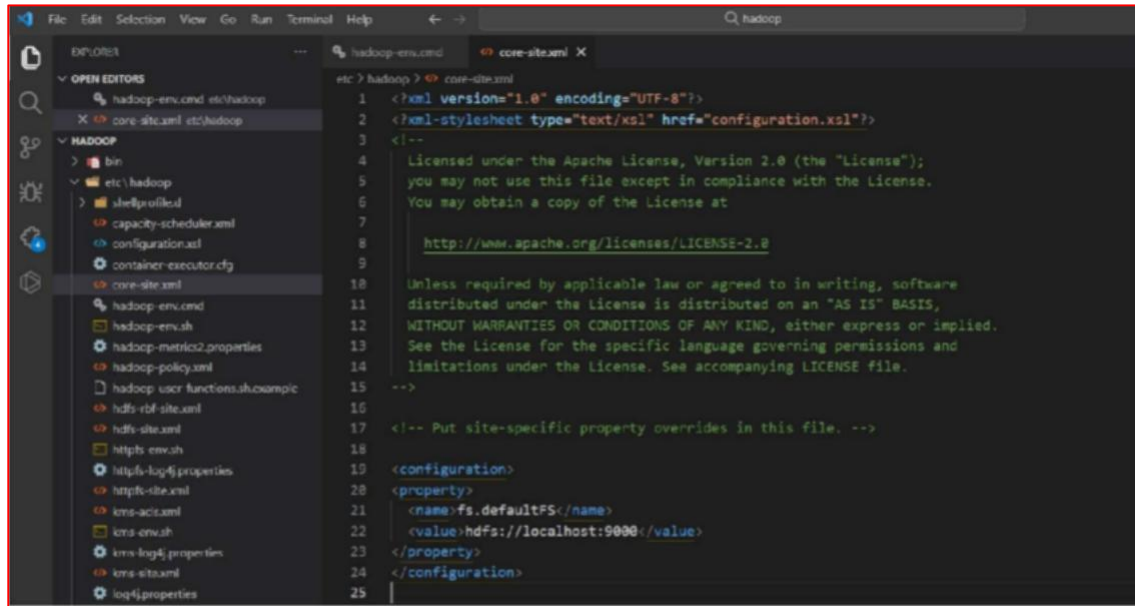
<property>

<name>fs.defaultFS</name>

<value>hdfs://localhost:9000</value>

</property>

</configuration>



ii) For hdfs-site.xml

Code:

<configuration>

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

<property>

<name>dfs.namenode.name.dir</name>

<value>C:\hadoop\data\namenode</value>

</property>

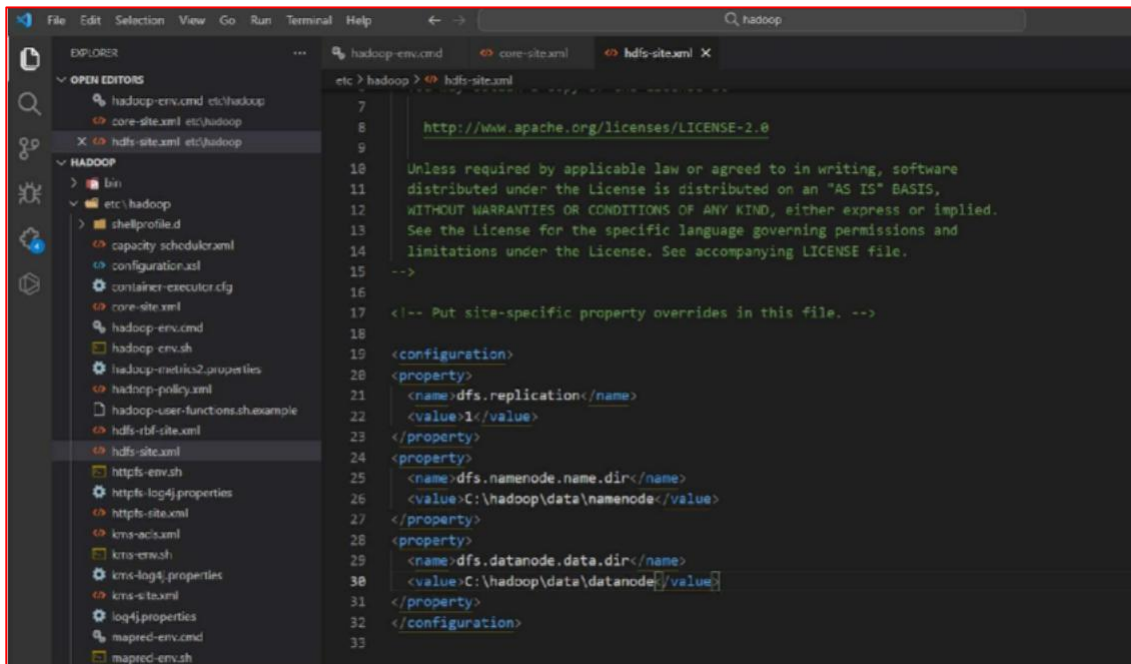
<property>

<name>dfs.datanode.data.dir</name>

<value>C:\hadoop\data\datanode</value>

```
</property>
```

```
</configuration>
```



iii) For mapred-site.xml

Code:

```
<configuration>
```

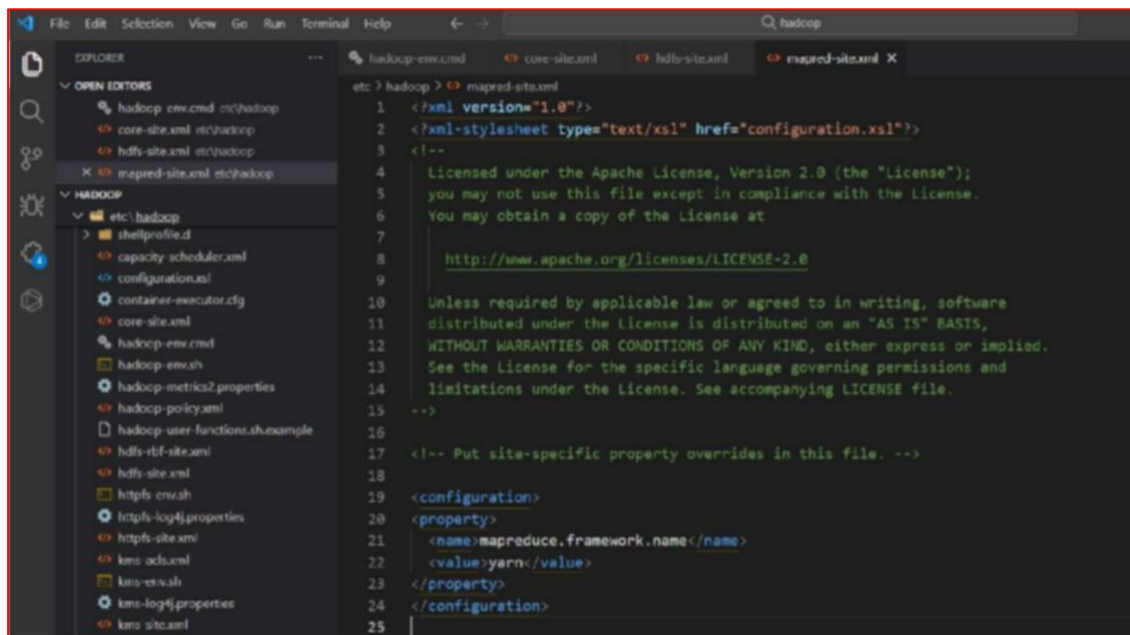
```
<property>
```

```
  <name>mapreduce.framework.name</name>
```

```
  <value>yarn</value>
```

```
</property>
```

```
</configuration>
```

iv) For yarn-site.xml

Code:

```
<configuration>
```

```
<property>
```

```
  <name>yarn.nodemanager.aux-services</name>
```

```
  <value>mapreduce_shuffle</value>
```

```
</property>
```

```
<property>
```

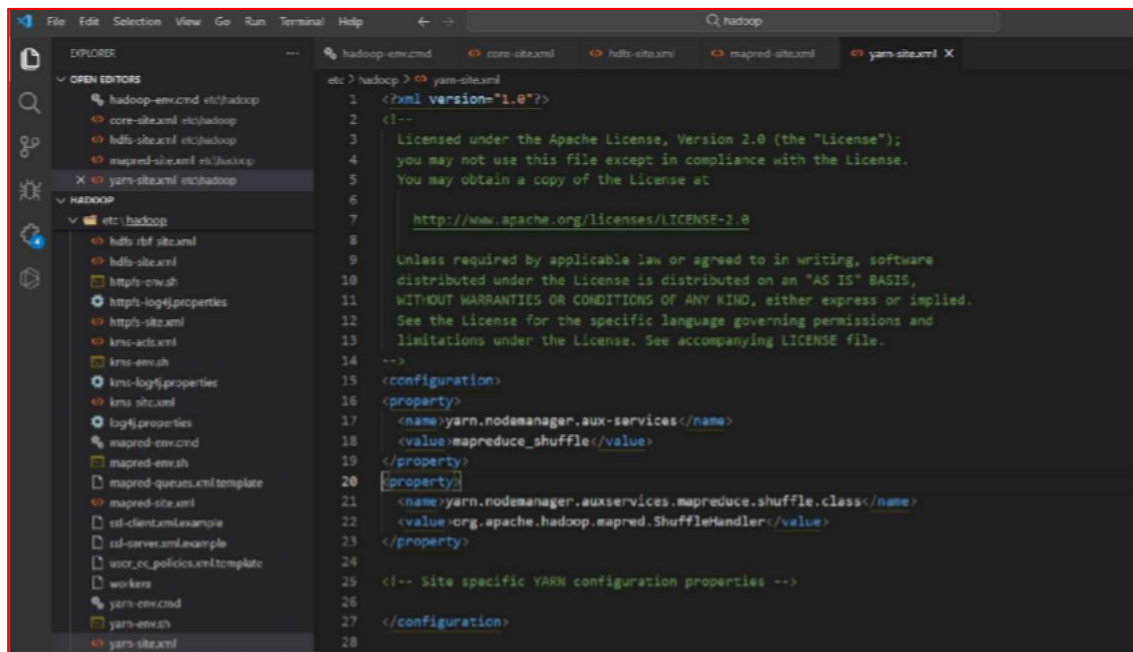
```
  <name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
```

```
  <value>org.apache.hadoop.mapred.ShuffleHandler</value>
```

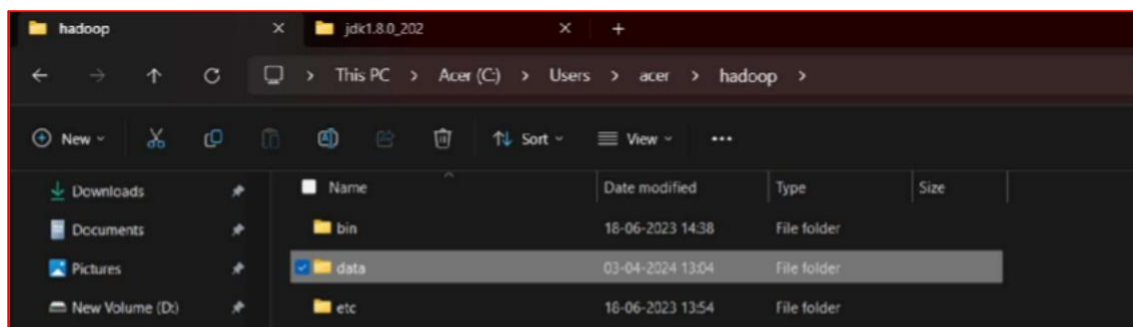
```
</property>
```

```
<!-- Site specific YARN configuration properties -->
```

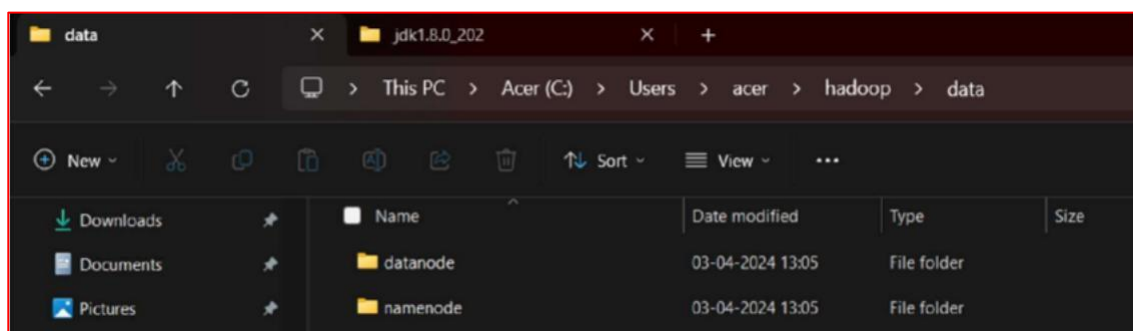
```
</configuration>
```

10) Create a folder “**data**” in hadoop directory

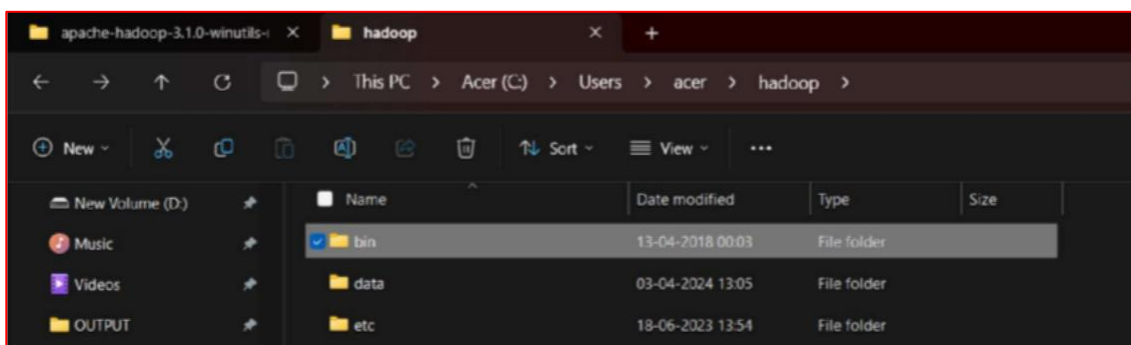
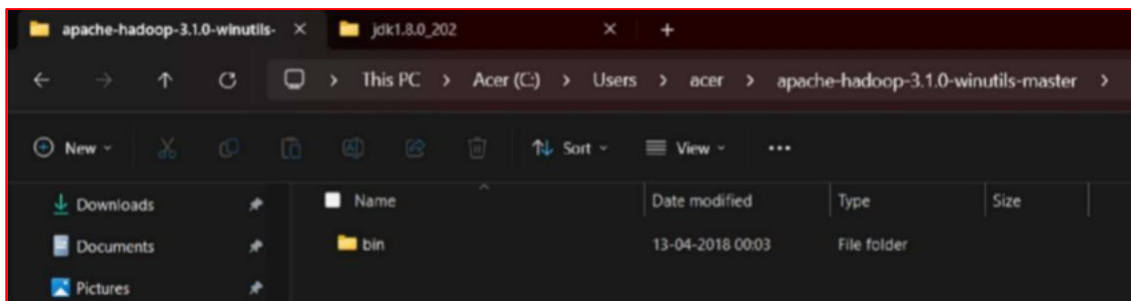
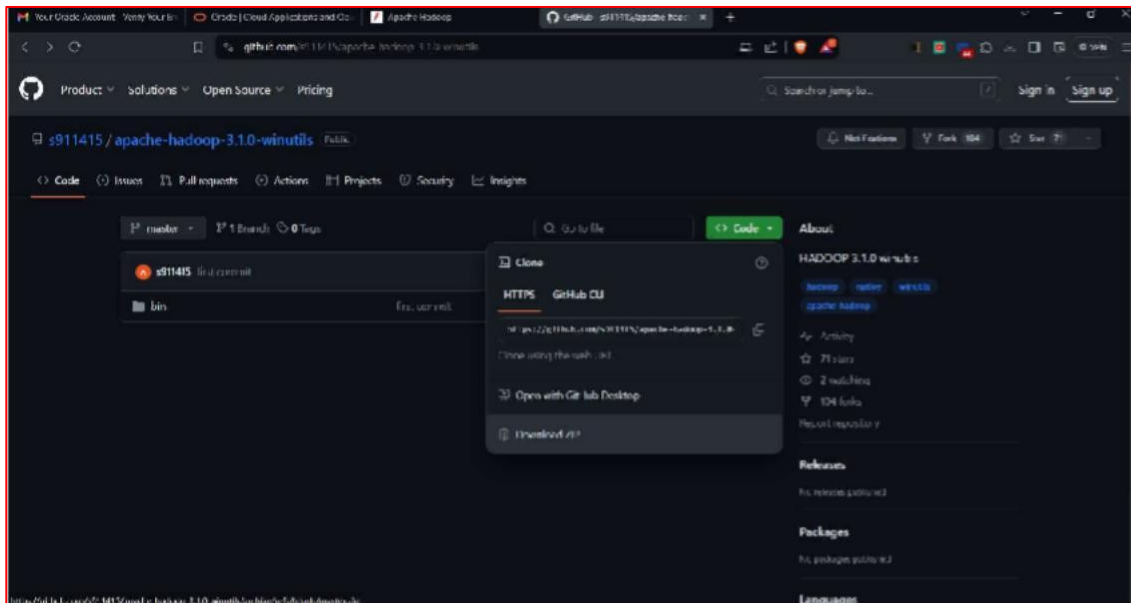


11) Create a folder with name “**datanode**” and folder “**namenode**” in the **data** folder



12) Download the zip file using the below link, Extract it and copy the bin folder in hadoop directory and replace it

(Link: <https://github.com/s911415/apache-hadoop-3.1.0-winutils>)

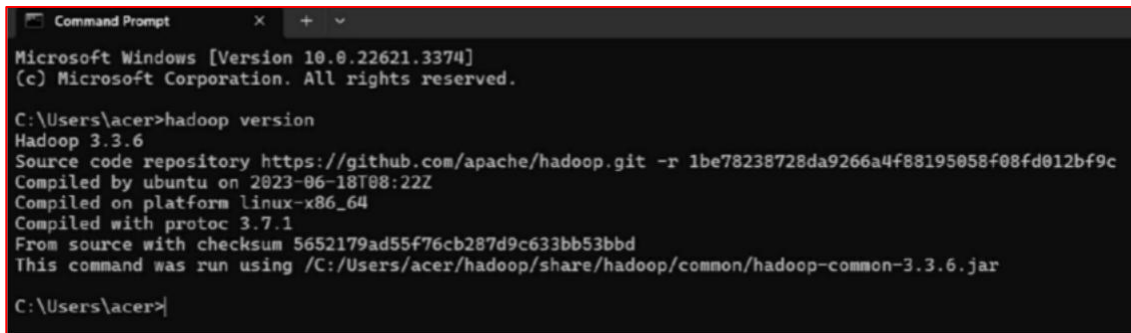


13) Check whether hadoop is successfully installed by running the below command

Command:

hadoop version

Output:



```

Microsoft Windows [Version 10.0.22621.3374]
(c) Microsoft Corporation. All rights reserved.

C:\Users\acer>hadoop version
Hadoop 3.3.6
Source code repository https://github.com/apache/hadoop.git -r 1be78238728da9266a4f88195058f08fd012bf9c
Compiled by ubuntu on 2023-06-18T08:22Z
Compiled on platform linux-x86_64
Compiled with protoc 3.7.1
From source with checksum 5652179ad55f76cb287d9c633bb53bbd
This command was run using /C:/Users/acer/hadoop/share/hadoop/common/hadoop-common-3.3.6.jar
C:\Users\acer>
  
```

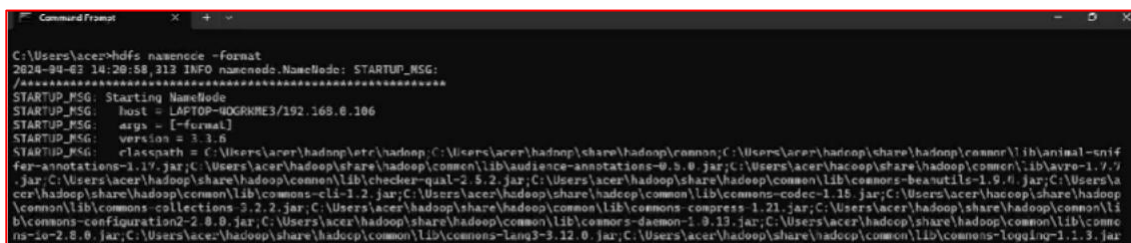
14) Format the NameNode

Formatting the NameNode is done once when hadoop is installed and not for running hadoop filesystem, else it will delete all the data inside HDFS. Run this

Command:

hdfs namenode -format

Output:

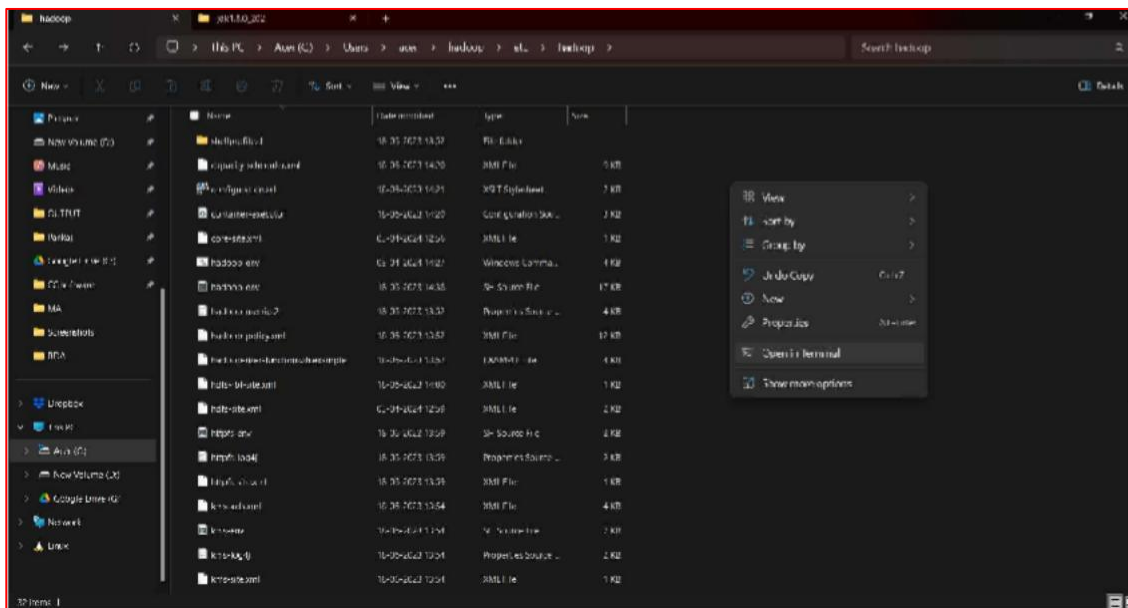


```

C:\Users\acer>hdfs namenode -format
2024-09-03 14:20:58,313 INFO namenode.NameNode: STARTUP_MSG:
/*****
STARTUP_MSG: Starting NameNode
STARTUP_MSG: host = LAPTOP-4OGRKNE3/192.168.0.106
STARTUP_MSG: aux = [-format]
STARTUP_MSG: version = 3.3.6
STARTUP_MSG: classpath = C:\Users\acer\hadoop\etc\hadoop;C:\Users\acer\hadoop\share\hadoop\common;C:\Users\acer\hadoop\share\hadoop\common\lib\animal-sniff-
fer-annotations-1.19.jar;C:\Users\acer\hadoop\share\hadoop\common\lib\audience-annotations-0.5.0.jar;C:\Users\acer\hadoop\share\hadoop\common\lib\avro-1.9.7
.jar;C:\Users\acer\hadoop\share\hadoop\common\lib\checker-qual-2.5.2.jar;C:\Users\acer\hadoop\share\hadoop\common\lib\commons-beanutils-1.9.4.jar;C:\Users\ac
er\hadoop\share\hadoop\common\lib\commons-cli-1.2.jar;C:\Users\acer\hadoop\share\hadoop\common\lib\commons-codec-1.15.jar;C:\Users\acer\hadoop\share\hadoop
\common\lib\commons-collections-3.2.2.jar;C:\Users\acer\hadoop\share\hadoop\common\lib\commons-compress-1.21.jar;C:\Users\acer\hadoop\share\hadoop\common\li
b\commons-configuration2-2.8.0.jar;C:\Users\acer\hadoop\share\hadoop\common\lib\commons-daemon-1.0.13.jar;C:\Users\acer\hadoop\share\hadoop\common\lib\commo
ns-io-2.8.0.jar;C:\Users\acer\hadoop\share\hadoop\common\lib\commons-lang3-3.12.0.jar;C:\Users\acer\hadoop\share\hadoop\common\lib\commons-logging-1.1.3.jar
*****/
  
```

15) Go to hadoop inside that go to etc folder and then open hadoop directory and right click choose the option open in terminal.

C:\Users\acer\hadoop\etc\hadoop



Command:

start-dfs.cmd

start-yarn cmd

Output:

```

Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! https://aka.ms/PSWindows

PS C:\Users\acer\hadoop\etc\hadoop> start-dfs.cmd
PS C:\Users\acer\hadoop\etc\hadoop> start-yarn.cmd
starting yarn daemons
PS C:\Users\acer\hadoop\etc\hadoop> |
  
```

Note:- ensure that the **sbin** directory is included in your system's **PATH** variable

start-dfs.cmd - output

```

2024-04-03 15:31:50,102 INFO impl.FsDatasetImpl: Total time to add all replicas to map for block pool BP-109928413-192.168.0.106-171211702475: 6ms
2024-04-03 15:31:50,102 INFO checker.ThrottledSyncChecker: Scheduling a check for C:\hadoop\data\catanode
2024-04-03 15:31:50,129 INFO checker.DatasetVolumeChecker: Scheduled health check for volume C:\hadoop\data\catanode
2024-04-03 15:31:50,129 INFO datanode.VolumesScanner: Now scanning bpid BP-109928413-192.168.0.106-171211702475 on volume C:\hadoop\data\catanode
2024-04-03 15:31:50,132 INFO datanode.VolumesScanner: VolumesScanner[C:\hadoop\data\catanode, 05-bbdc6882-673c-848d-bc23-2e6211b8b0a2]: finished scanning block pool BP-109928413-192.168.0.106-171211702475
2024-04-03 15:31:50,139 WARN datanode.DirectoryScanner: dfs.datanode.directoryscan.throttle.limit.ms.per.sec set to value above 10000 ms/sec. Assuming default value of -1
2024-04-03 15:31:50,139 INFO datanode.DirectoryScanner: Periodic Directory Tree Verification scan starting at 7693775ms with interval of 21600000ms and throttle limit of -1ms/s
2024-04-03 15:31:50,155 INFO datanode.DataNode: Block pool BP-109928413-192.168.0.106-171211702475 (Datanode Uuid bdc9a96b-08d3-4fee-a5ed-840a1cd0804b): service to localhost/127.0.0.1:9000 beginning handshake with NN
2024-04-03 15:31:50,169 INFO datanode.DataNode: VolumesScanner[C:\hadoop\data\catanode, 05-bbdc6882-673c-848d-bc23-2e6211b8b0a2]: no suitable block pools found to scan. Waiting 1234399969 ms.
2024-04-03 15:31:50,350 INFO datanode.DataNode: Block pool BP-109928413-192.168.0.106-171211702475 (Datanode Uuid bdc9a96b-08d3-4fee-a5ed-840a1cd0804b): service to localhost/127.0.0.1:9000 successfully registered with NN
2024-04-03 15:31:50,359 INFO datanode.DataNode: For namenode localhost/127.0.0.1:9000 using BLOCKREPORT_INTERVAL of 21600000secs CACHEREPORT_INTERVAL of 190000secs Initial delay: 0secs; heartbeatInterval:3000
2024-04-03 15:31:50,359 INFO datanode.DataNode: Starting IIS Task Handler.
2024-04-03 15:31:50,519 INFO datanode.DataNode: After receiving heartbeat response, updating state of namenode localhost:9000 to active
2024-04-03 15:31:50,512 INFO datanode.DataNode: Successfully sent block report 0x7eedb1d5388a0c1 with lease ID 0x5fcc2d78674035ub to namenode: localhost/127.0.0.1:9000, containing 1 storage report(s), of which we sent 1. The reports had 0 total blocks and used 1 RPC(s). This took 8 msecs to generate and 02 msecs for RPC and NN processing. Get back one command: FinalizeCommand/S.
2024-04-03 15:31:50,614 INFO datanode.DataNode: Get finalize command for block pool BP-109928413-192.168.0.106-171211702475
  
```

start-yarn.cmd - output

```

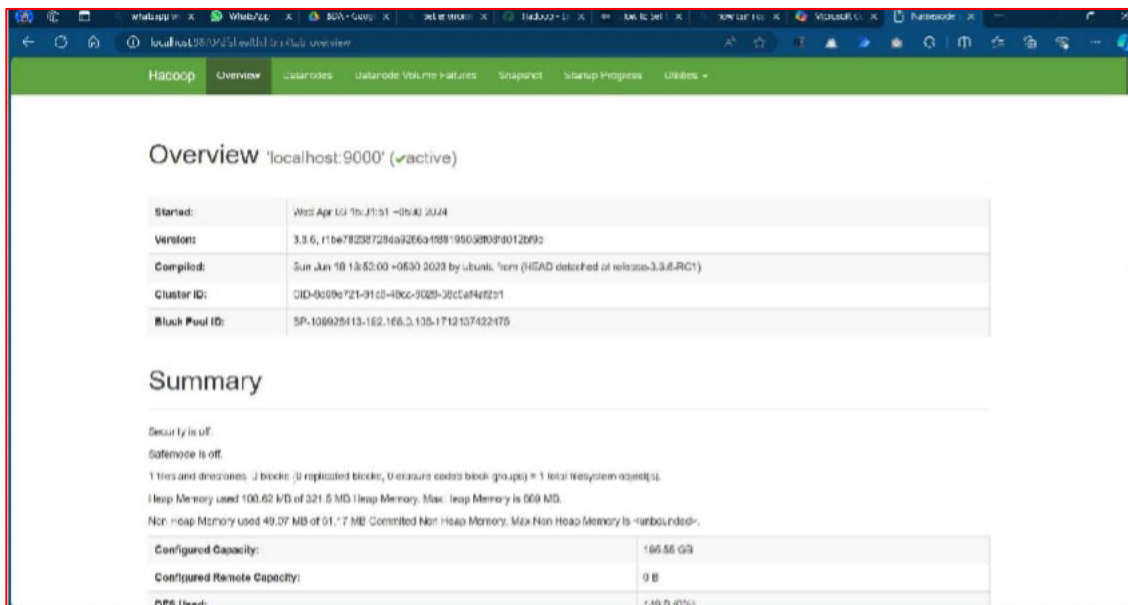
2024-04-03 15:33:07,023 INFO containermanager.ContainerManagerImpl: AMRMProxyService is disabled
2024-04-03 15:33:07,023 INFO localizer.ResourceLocalizerService: per directory file limit = 8192
2024-04-03 15:33:07,708 WARN nativeio.NativeIO: NativeIO.getStat error (3): The system cannot find the path specified.
-- file path: ttp:\hadoop-acer\m-local-dir\mPrivate
2024-04-03 15:33:07,778 WARN nativeio.NativeIO: NativeIO.getStat error (3): The system cannot find the path specified.
-- file path: ttp:\hadoop-acer\m-local-dir\filecache
2024-04-03 15:33:07,853 WARN nativeio.NativeIO: NativeIO.getStat error (3): The system cannot find the path specified.
-- file path: ttp:\hadoop-acer\m-local-dir\usercache
2024-04-03 15:33:07,906 INFO event.AsyncDispatcher: Registering place org.apache.hadoop.yarn.server.nodemanager.containermanager.localizer.event.LocalizerEventType for class org.apache.hadoop.yarn.server.nodemanager.containermanager.localizer.ResourceLocalizationServiceLocalizerTracker
2024-04-03 15:33:07,961 INFO resources.ResourceManagerModule: Using traffic control bandwidth handler
2024-04-03 15:33:08,214 INFO containermanager.AuxServices: Initialized auxiliary service mapreduce_shuffle
2024-04-03 15:33:08,216 INFO containermanager.AuxServices: Adding auxiliary service mapreduce_shuffle version null.
2024-04-03 15:33:08,216 INFO monitor.ContainersMonitorImpl: Using ResourceCalculatorPlugin: org.apache.hadoop.yarn.util.ResourceCalculatorPlugin:0.1100060
2024-04-03 15:33:08,216 INFO monitor.ContainersMonitorImpl: Using ResourceCalculatorPreprocessTree: null
2024-04-03 15:33:08,246 INFO conf.Configuration: resource-types.xml not found
2024-04-03 15:33:08,246 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-04-03 15:33:08,258 INFO monitor.ContainersMonitorImpl: Setting the resources allocated to containers to <memory:8192, vCores:8>
2024-04-03 15:33:08,258 INFO monitor.ContainersMonitorImpl: Physical memory check enabled: true
2024-04-03 15:33:08,258 INFO monitor.ContainersMonitorImpl: Virtual memory check enabled: true
2024-04-03 15:33:08,258 INFO monitor.ContainersMonitorImpl: Elastic memory control enabled: false
2024-04-03 15:33:08,258 INFO monitor.ContainersMonitorImpl: Strict memory control enabled: true
2024-04-03 15:33:08,258 INFO monitor.ContainersMonitorImpl: ContainersMonitor enabled: true
2024-04-03 15:33:08,258 INFO monitor.ContainersMonitorImpl: Container log Monitor Enabled: false
  
```

16) Two more windows will open, one for yarn resource manager and one for yarn node manager

Note: Make sure all the 4 Apache Hadoop Distribution windows are up n running. If they are not running, you will see an error or a shutdown message. In that case, you need to debug the error.

To access information about resource manager current jobs, successful and failed jobs, go to this link in browser-

<http://localhost:9870>



Hadoop Overview 'localhost:9000' (✓active)

Started:	Wed Apr 03 10:41:51 -0500 2024
Version:	3.3.6, r1be78297286a9266a1989195098094012b992
Compiled:	Sun Jun 16 13:52:00 +0530 2023 by ubuntu from (HIGAD detached of release-3.3.6-RC1)
Cluster ID:	CID-9009a721-91c5-48cc-9025-35c5a74422d1
Block Pool ID:	BP-108929413-162.166.2.135-1712137432475

Summary

Security is off.
SafeMode is off.
1 files and directories, 2 blocks, 0 replicated blocks, 0 erasure coded block groups = 1 total filesystem objects.
Heap Memory used 100.62 MB of 321.5 MB Heap Memory. Max Heap Memory is 609 MB.
Non-Heap Memory used 49.07 MB of 51.7 MB Committed Non-Heap Memory. Max Non-Heap Memory is 'unbounded'.

Configured Capacity:	165.56 GB
Configured Remote Capacity:	0 B
DFS Used:	1.69 B (0%)