MScIT Semester 2
Big Data Analytics
Hadoop Installation

Mumbai University

Install, configure and run Hadoop and HDFS and explore HDFS on Windows

Steps to Install Hadoop

1. Install Java JDK 1.8
2. Download Hadoop and extract and place under C drive
3. Set Path in Environment Variables
4. Config files under Hadoop directory
5. Create folder datanode and namenode under data directory
6. Edit HDFS and YARN files
7. Set Java Home environment in Hadoop environment
8. Setup Complete. Test by executing start-all.cmd

There are two ways to install Hadoop, i.e.

9. Single node

10. Multi node

Here, we use multi node cluster.

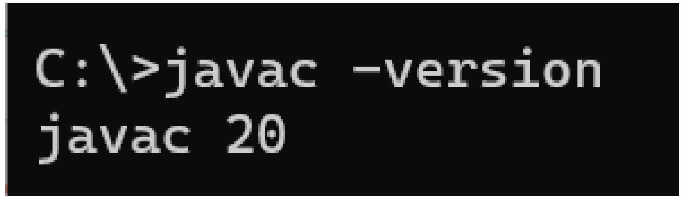
1. Install Java

11. – Java JDK Link to download

https://www.oracle.com/java/technologies/javase-jdk8-downloads.html

12. – extract and install Java in C:\Java

13. – open cmd and type -> javac -version



```
C:\>javac -version
javac 20
```

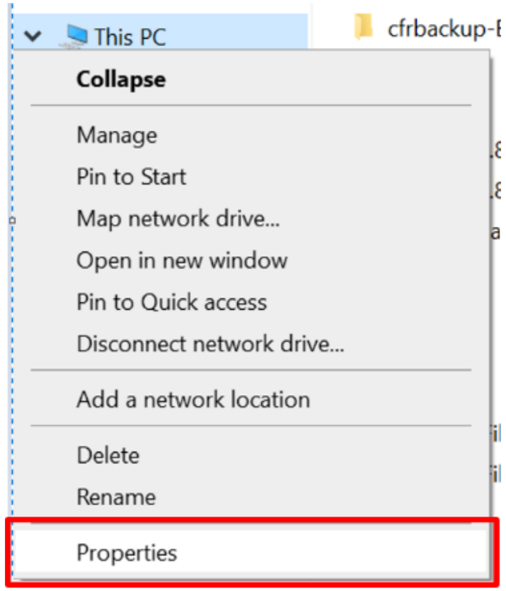
2. Download Hadoop

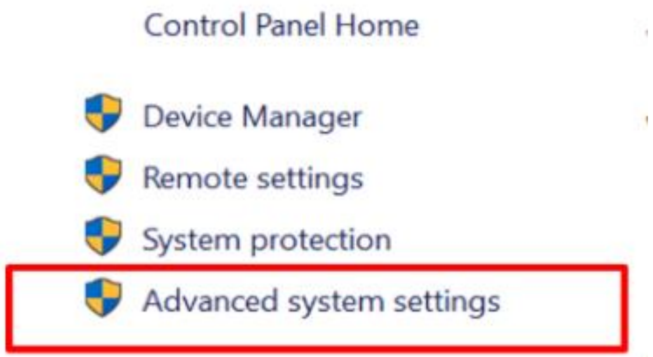
https://www.apache.org/dyn/closer.cgi/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz

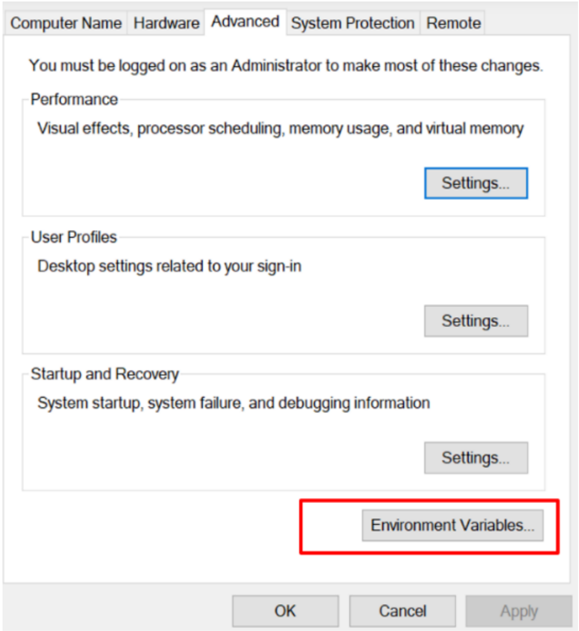
- right click .rar.gz file -> show more options -> 7-zip->and extract to C:\Hadoop-3.3.0\

3. Set the path JAVA_HOME Environment variable

4. Set the path HADOOP_HOME Environment variable







You must be logged on as an Administrator to make most of these changes.

Performance

Visual effects, processor scheduling, memory usage, and virtual memory

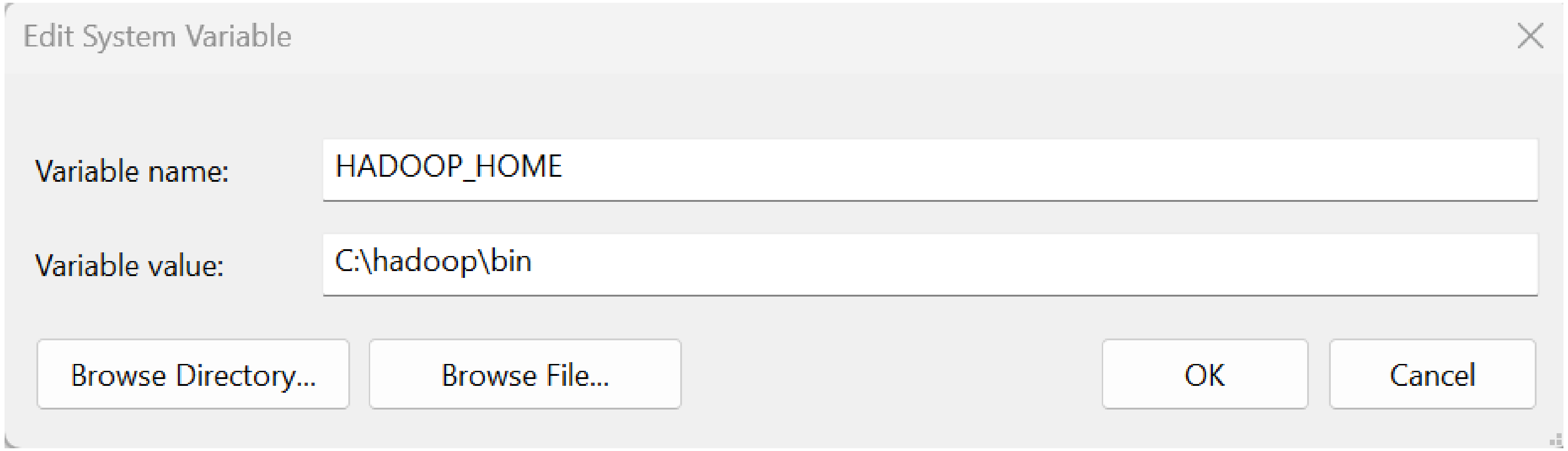
User Profiles

Desktop settings related to your sign-in

Startup and Recovery

System startup, system failure, and debugging information

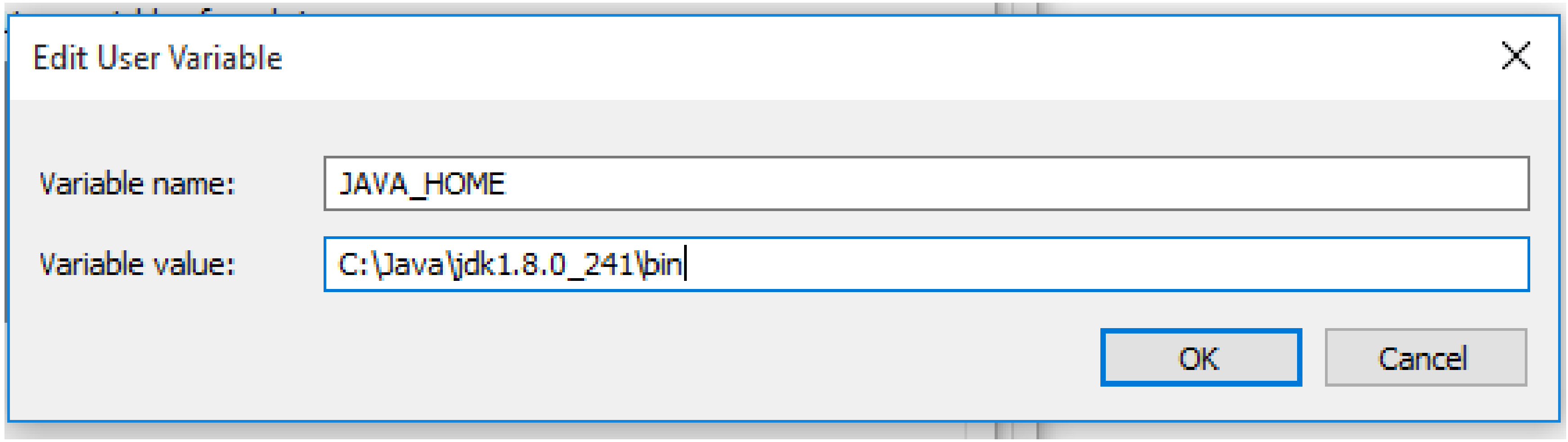
Click on New to both user variables and system variables.



Edit System Variable

Variable name: HADOOP_HOME

Variable value: C:\hadoop\bin

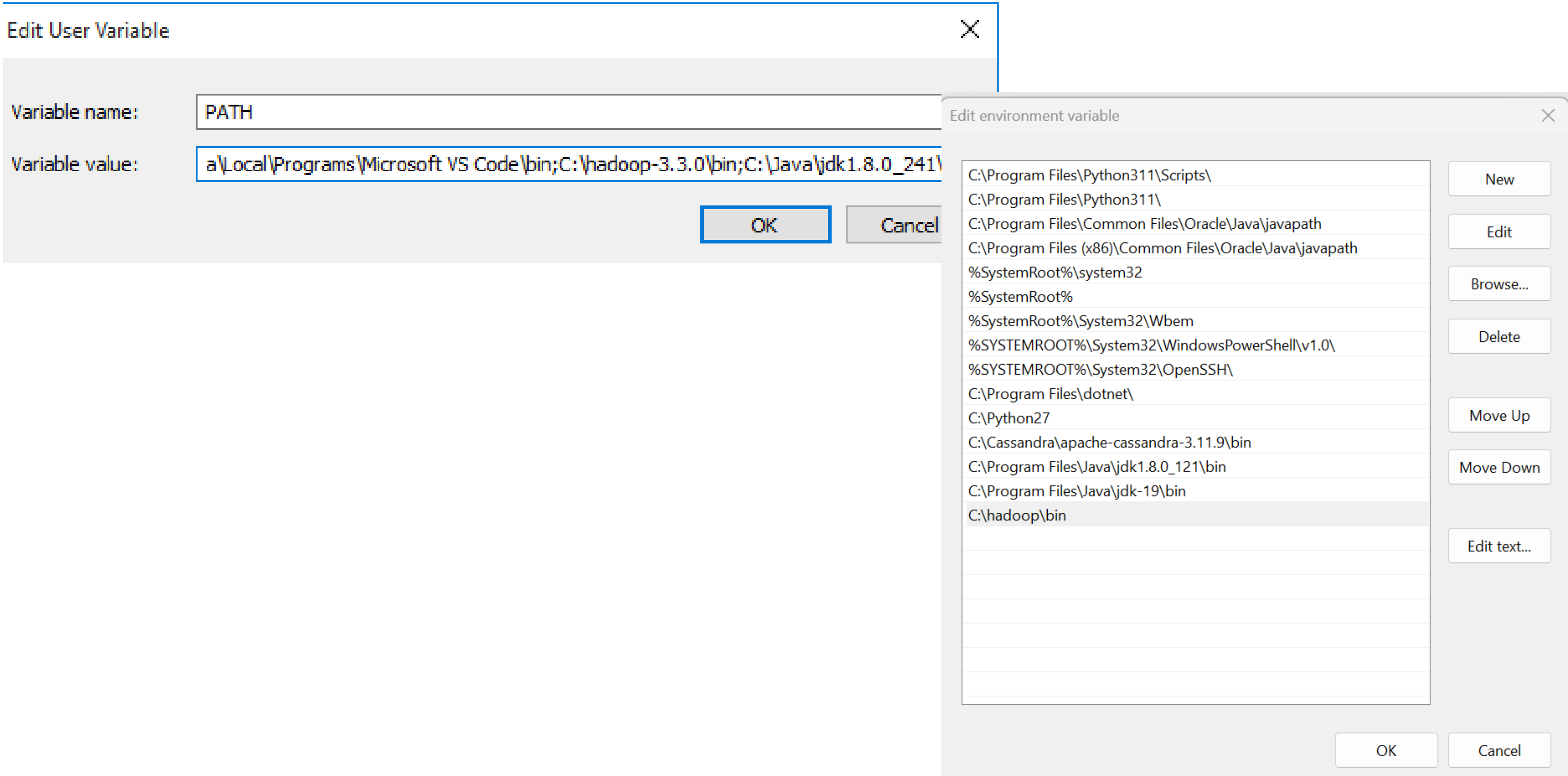


Edit User Variable

Variable name: JAVA_HOME

Variable value: C:\Java\jdk1.8.0_241\bin

- Click on user variable -> path -> edit-> add path for Hadoop and java upto ‘bin’



Edit User Variable

Variable name: PATH

Variable value: a\Local\Programs\Microsoft VS Code\bin;C:\hadoop-3.3.0\bin;C:\Java\jdk1.8.0_241\

Edit environment variable

```
C:\Program Files\Python311\Scripts\
C:\Program Files\Python311\
C:\Program Files\Common Files\Oracle\Java\javapath
C:\Program Files (x86)\Common Files\Oracle\Java\javapath
%SystemRoot%\system32
%SystemRoot%
%SystemRoot%\System32\Wbem
%SYSTEMROOT%\System32\WindowsPowerShell\v1.0\
%SYSTEMROOT%\System32\OpenSSH\
C:\Program Files\dotnet\
C:\Python27
C:\Cassandra\apache-cassandra-3.11.9\bin
C:\Program Files\Java\jdk1.8.0_121\bin
C:\Program Files\Java\jdk-19\bin
C:\hadoop\bin
```

5. Configurations

Edit file C:\hadoop\etc\hadoop\core-site.xml,

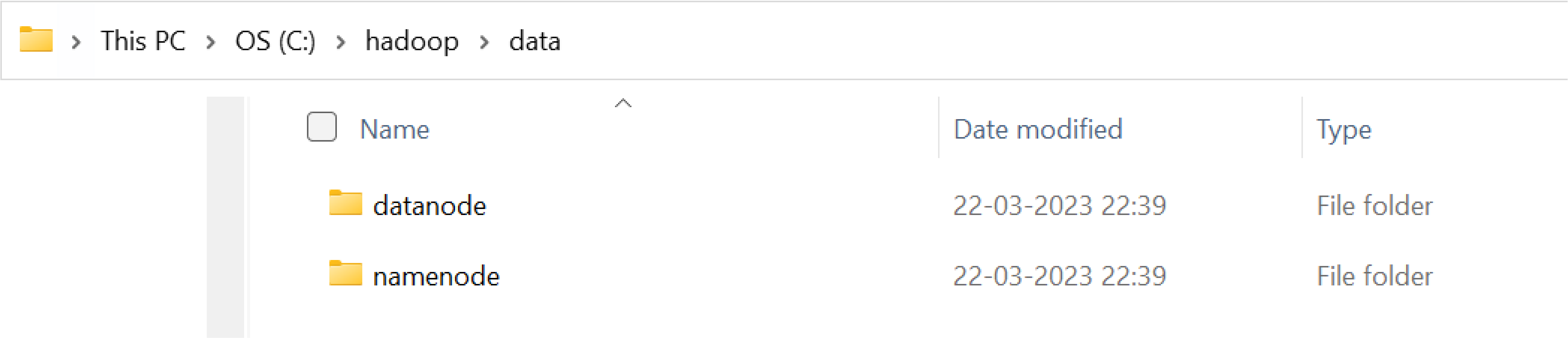
paste the xml code in folder and save

```
<configuration>

<property>

    <name>fs.defaultFS</name>

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

 </property>

</configuration>
```

Rename “mapred-site.xml.template” to “mapred-site.xml” and edit this file C:/hadoop/etc/hadoop/mapred-site.xml, paste xml code and save this file

```
<configuration>

  <property>

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

     <value>yarn</value>

  </property>

</configuration>
```

- Create folder “data” under “C:\Hadoop-3.3.0”
- Create folder “datanode” under “C:\Hadoop-3.3.0\data”
- Create folder “namenode” under “C:\Hadoop-3.3.0\data”



› This PC › OS (C:) › hadoop › data

□ Name	Date modified	Type
datanode	22-03-2023 22:39	File folder
namenode	22-03-2023 22:39	File folder

Edit file C:\Hadoop-3.3.0/etc/hadoop/hdfs-site.xml,

paste xml code and save this file.

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

```
<property>
     <name>dfs.namenode.name.dir</name>
     <value>/hadoop-3.3.0/data/namenode</value>
  </property>
  <property>
     <name>dfs.datanode.data.dir</name>
     <value>/hadoop-3.3.0/data/datanode</value>
  </property>
 </configuration>
```

Edit file C:/Hadoop-3.3.0/etc/hadoop/yarn-site.xml,

paste xml code and save this file.

```
<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>
   <property>
                                                <name>yarn.resourcemanager.address</name>
                                                <value>127.0.0.1:8032</value>
   </property>
   <property>
                                                <name>yarn.resourcemanager.scheduler.address</name>
                                                <value>127.0.0.1:8030</value>
    </property>
    <property>
                                                <name>yarn.resourcemanager.resource-tracker.address</name>
                                                <value>127.0.0.1:8031</value>
     </property>
</configuration>
```

6. Edit file C:/Hadoop-3.3.0/etc/hadoop/hadoop-env.cmd

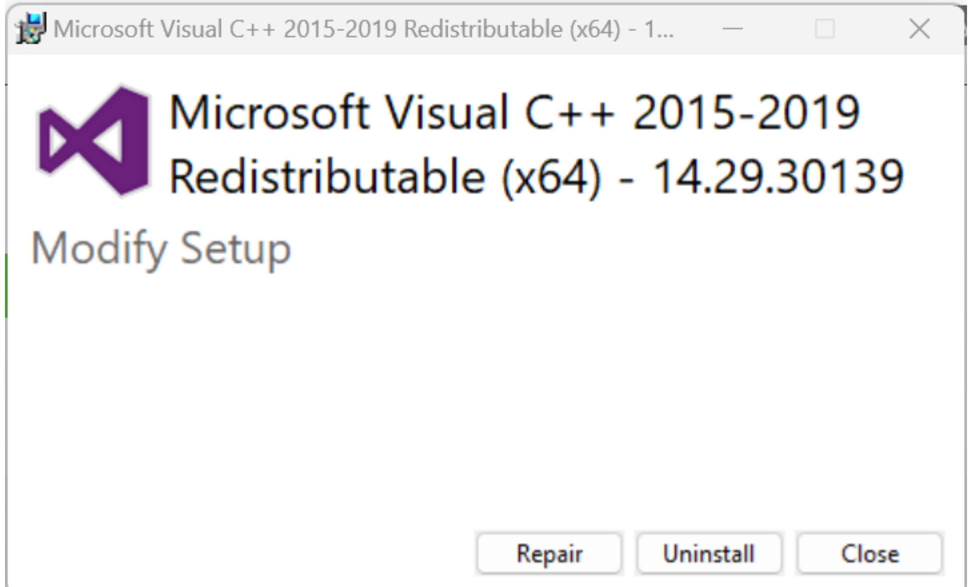
Find “JAVA_HOME=%JAVA_HOME%” and replace it as

```
set JAVA_HOME="C:\java\jdk1.8.0_121"
```

7. Download “redistributable” package

Download and run VC_redist.x64.exe

This is a “redistributable” package of the Visual C runtime code for 64-bit applications, from Microsoft. It contains certain shared code that every application written with Visual C expects to have available on the Windows computer it runs on.



Microsoft Visual C++ 2015-2019 Redistributable (x64) - 1...

Microsoft Visual C++ 2015-2019 Redistributable (x64) - 14.29.30139

Modify Setup

8. Hadoop Configurations

Download bin folder from

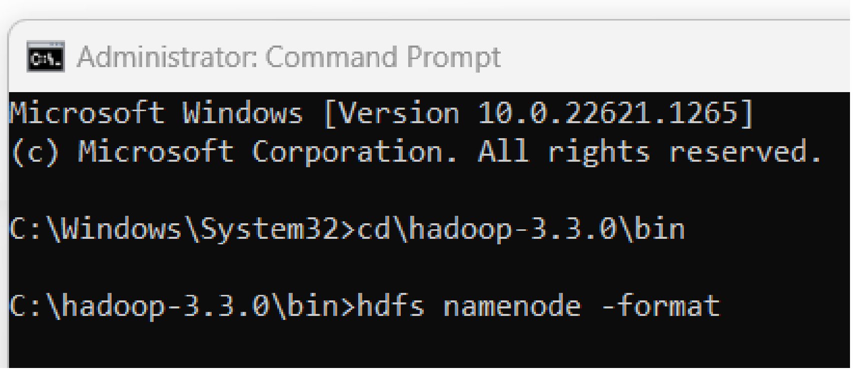
https://github.com/s911415/apache-hadoop-3.1.0-winutils

– Copy the bin folder to c:\hadoop-3.3.0. Replace the existing bin folder.

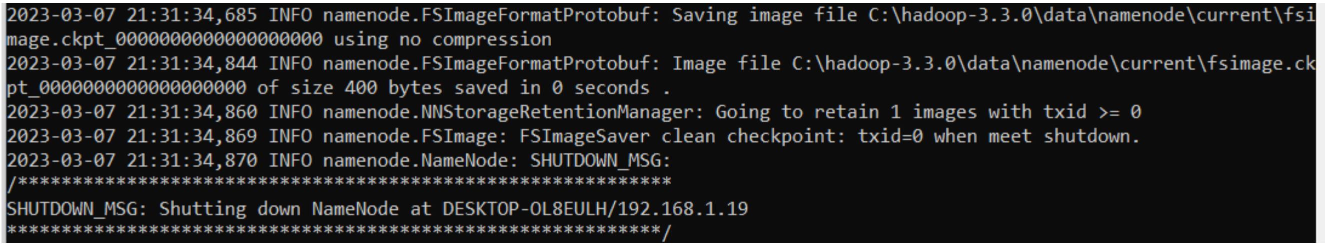
9. copy "hadoop-yarn-server-timelineservice-3.0.3.jar" from ~\hadoop-3.0.3\share\hadoop\yarn\timelineservice to ~\hadoop-3.0.3\share\hadoop\yarn folder.

Format the NameNode

– Open cmd ‘Run as Administrator’ and type command “hdfs namenode –format”



```
Microsoft Windows [Version 10.0.22621.1265]
(c) Microsoft Corporation. All rights reserved.

C:\Windows\System32>cd\hadoop-3.3.0\bin

C:\hadoop-3.3.0\bin>hdfs namenode -format
```



```
2023-03-07 21:31:34,685 INFO namenode.FSImageFormatProtobuf: Saving image file C:\hadoop-3.3.0\data\namenode\current\fsi
mage.ckpt_0000000000000000000 using no compression
2023-03-07 21:31:34,844 INFO namenode.FSImageFormatProtobuf: Image file C:\hadoop-3.3.0\data\namenode\current\fsimage.ck
pt_0000000000000000000 of size 400 bytes saved in 0 seconds .
2023-03-07 21:31:34,860 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2023-03-07 21:31:34,869 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2023-03-07 21:31:34,870 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at DESKTOP-OL8EULH/192.168.1.19
************************************************************/
```

. Testing

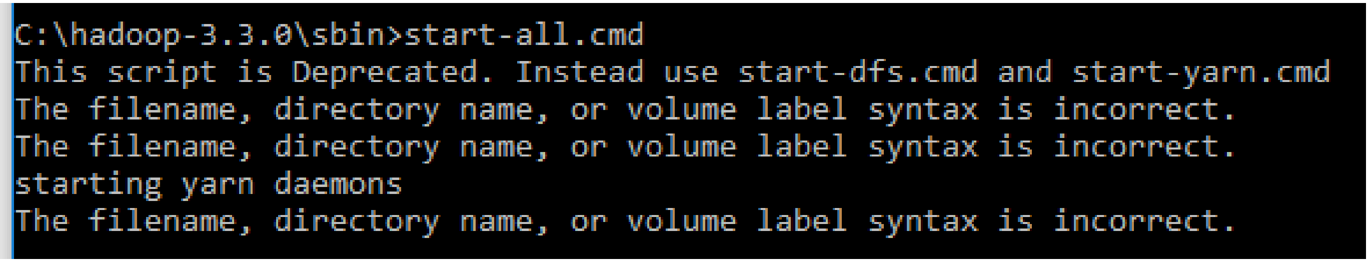
– Open cmd ‘Run as Administrator’ and change directory to C:\Hadoop-3.3.0\sbin

– type start-all.cmd

OR

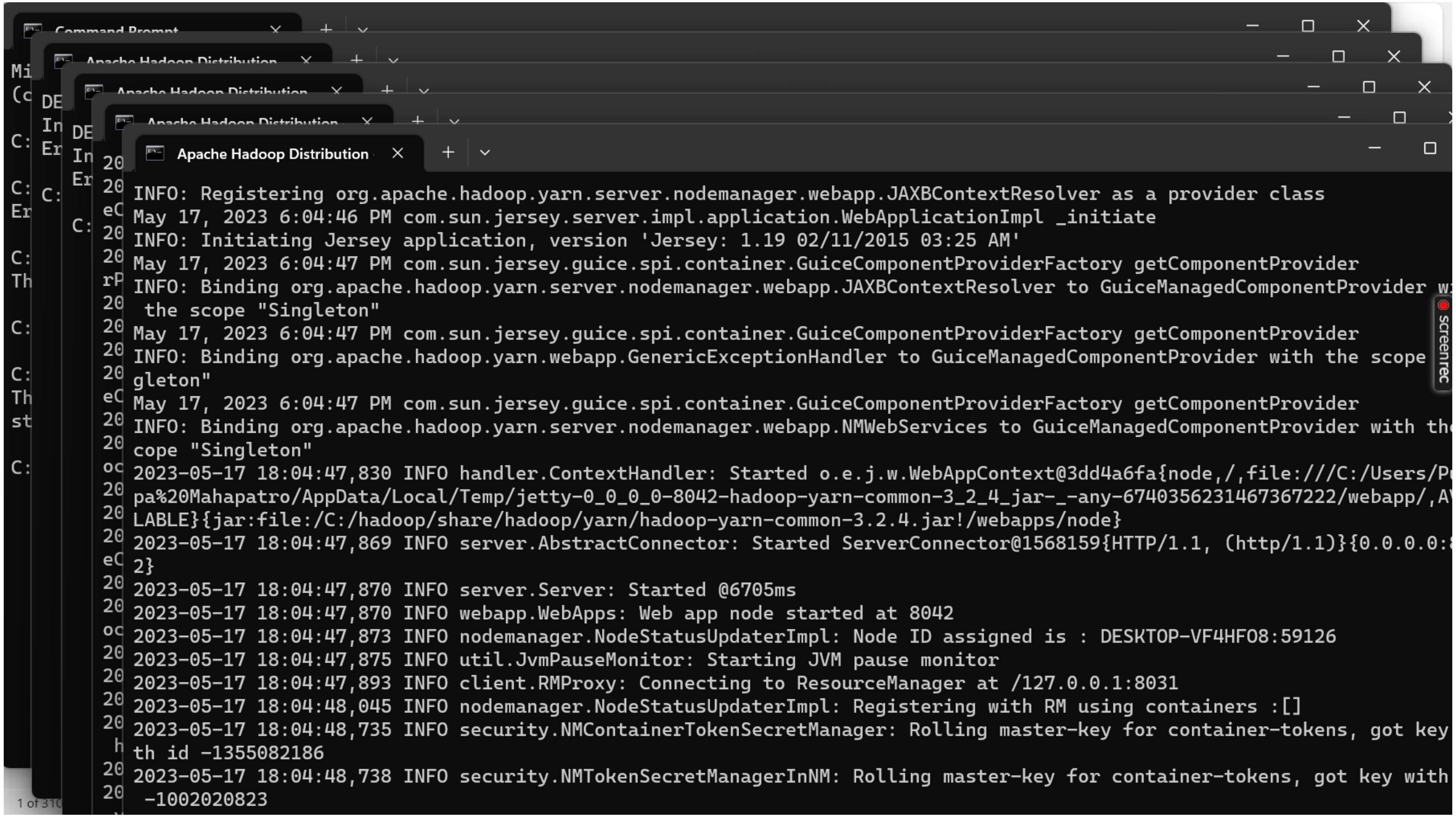
- type start-dfs.cmd

– type start-yarn.cmd



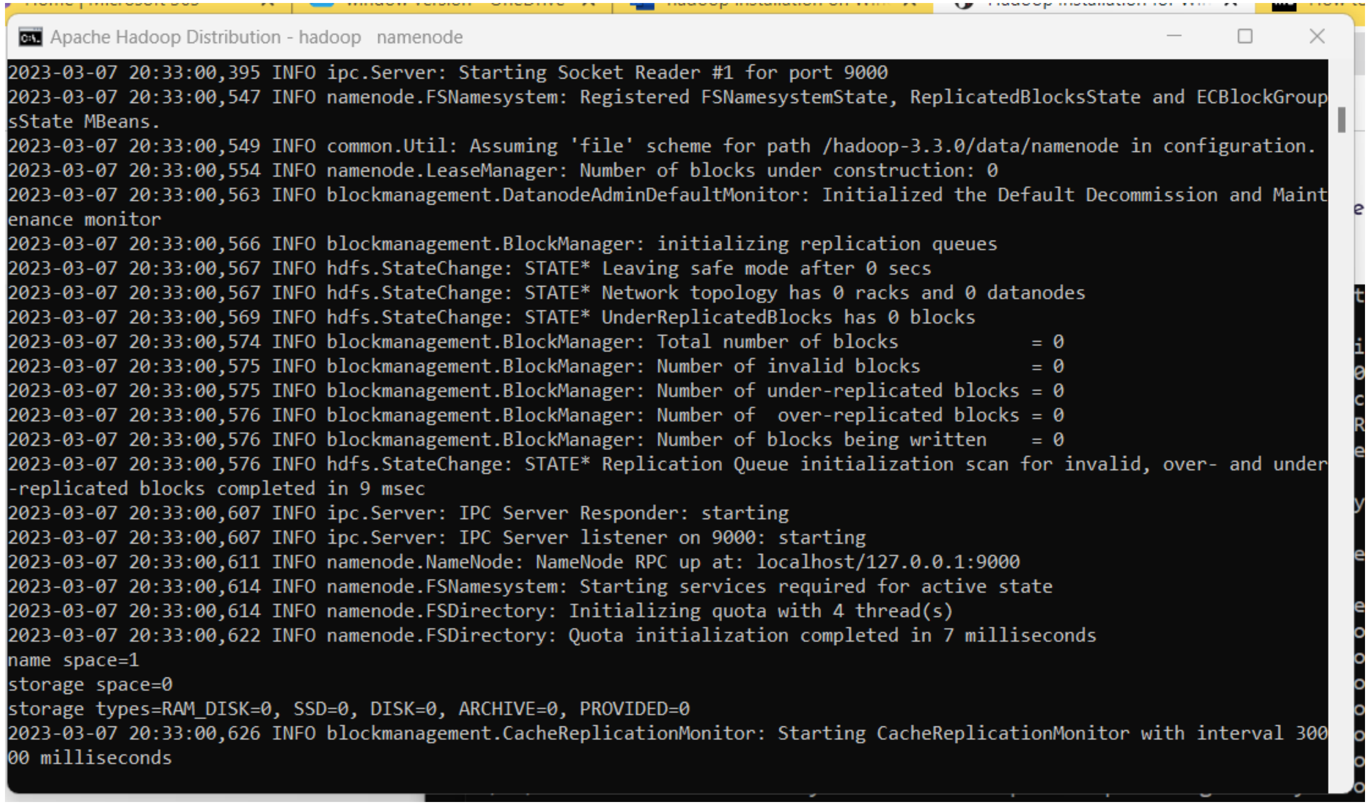
```
C:\hadoop-3.3.0\sbin>start-all.cmd
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
The filename, directory name, or volume label syntax is incorrect.
The filename, directory name, or volume label syntax is incorrect.
starting yarn daemons
The filename, directory name, or volume label syntax is incorrect.
```

Type start-all.cmd


Command Prompt
Apache Hadoop Distribution
Apache Hadoop Distribution
Apache Hadoop Distribution
Apache Hadoop Distribution
screenrec
1 of 310

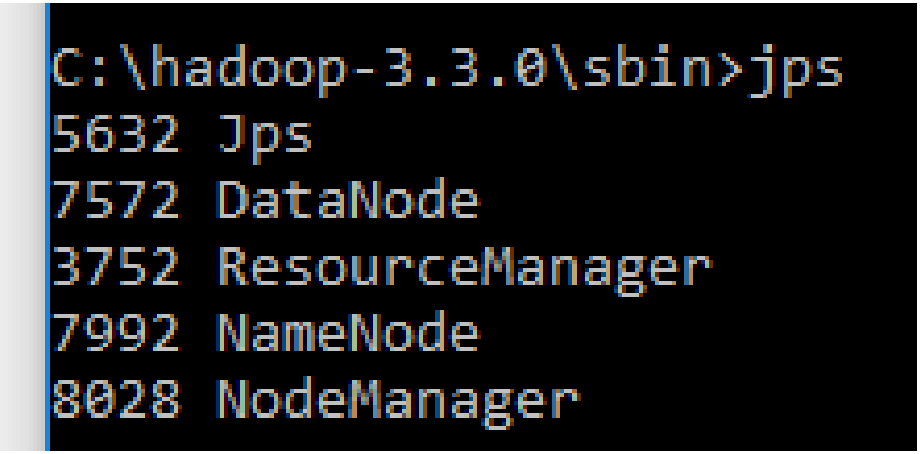
```
INFO: Registering org.apache.hadoop.yarn.server.nodemanager.webapp.JAXBContextResolver as a provider class
May 17, 2023 6:04:46 PM com.sun.jersey.server.impl.application.WebApplicationImpl _initiate
INFO: Initiating Jersey application, version 'Jersey: 1.19 02/11/2015 03:25 AM'
May 17, 2023 6:04:47 PM com.sun.jersey.guice.spi.container.GuiceComponentProviderFactory getComponentProvider
INFO: Binding org.apache.hadoop.yarn.server.nodemanager.webapp.JAXBContextResolver to GuiceManagedComponentProvider wi
 the scope "Singleton"
May 17, 2023 6:04:47 PM com.sun.jersey.guice.spi.container.GuiceComponentProviderFactory getComponentProvider
INFO: Binding org.apache.hadoop.yarn.webapp.GenericExceptionHandler to GuiceManagedComponentProvider with the scope
gleton"
May 17, 2023 6:04:47 PM com.sun.jersey.guice.spi.container.GuiceComponentProviderFactory getComponentProvider
INFO: Binding org.apache.hadoop.yarn.server.nodemanager.webapp.NMWebServices to GuiceManagedComponentProvider with the
cope "Singleton"
2023-05-17 18:04:47,830 INFO handler.ContextHandler: Started o.e.j.w.WebAppContext@3dd4a6fa{node,/,file:///C:/Users/Pu
pa%20Mahapatro/AppData/Local/Temp/jetty-0_0_0_0-8042-hadoop-yarn-common-3_2_4_jar-_-any-6740356231467367222/webapp/,AV
LABLE}{jar:file:/C:/hadoop/share/hadoop/yarn/hadoop-yarn-common-3.2.4.jar!/webapps/node}
2023-05-17 18:04:47,869 INFO server.AbstractConnector: Started ServerConnector@1568159{HTTP/1.1, (http/1.1)}{0.0.0.0:8
2}
2023-05-17 18:04:47,870 INFO server.Server: Started @6705ms
2023-05-17 18:04:47,870 INFO webapp.WebApps: Web app node started at 8042
2023-05-17 18:04:47,873 INFO nodemanager.NodeStatusUpdaterImpl: Node ID assigned is : DESKTOP-VF4HFO8:59126
2023-05-17 18:04:47,875 INFO util.JvmPauseMonitor: Starting JVM pause monitor
2023-05-17 18:04:47,893 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8031
2023-05-17 18:04:48,045 INFO nodemanager.NodeStatusUpdaterImpl: Registering with RM using containers :[]
2023-05-17 18:04:48,735 INFO security.NMContainerTokenSecretManager: Rolling master-key for container-tokens, got key
th id -1355082186
2023-05-17 18:04:48,738 INFO security.NMTokenSecretManagerInNM: Rolling master-key for container-tokens, got key with
 -1002020823
```

- You will get 4 more running threads for Datanode, namenode, resouce manager and node manager**You will get 4 more running threads for Datanode, namenode, resouce manager and node manager**



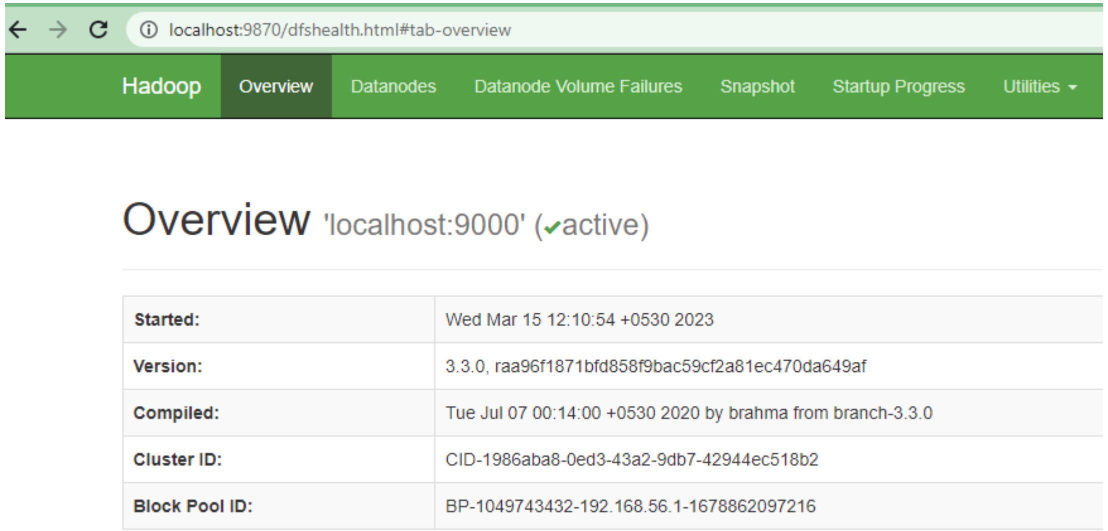
```
2023-03-07 20:33:00,395 INFO ipc.Server: Starting Socket Reader #1 for port 9000
2023-03-07 20:33:00,547 INFO namenode.FSNamesystem: Registered FSNamesystemState, ReplicatedBlocksState and ECBlockGroup
sState MBeans.
2023-03-07 20:33:00,549 INFO common.Util: Assuming 'file' scheme for path /hadoop-3.3.0/data/namenode in configuration.
2023-03-07 20:33:00,554 INFO namenode.LeaseManager: Number of blocks under construction: 0
2023-03-07 20:33:00,563 INFO blockmanagement.DatanodeAdminDefaultMonitor: Initialized the Default Decommission and Maint
enance monitor
2023-03-07 20:33:00,566 INFO blockmanagement.BlockManager: initializing replication queues
2023-03-07 20:33:00,567 INFO hdfs.StateChange: STATE* Leaving safe mode after 0 secs
2023-03-07 20:33:00,567 INFO hdfs.StateChange: STATE* Network topology has 0 racks and 0 datanodes
2023-03-07 20:33:00,569 INFO hdfs.StateChange: STATE* UnderReplicatedBlocks has 0 blocks
2023-03-07 20:33:00,574 INFO blockmanagement.BlockManager: Total number of blocks            = 0
2023-03-07 20:33:00,575 INFO blockmanagement.BlockManager: Number of invalid blocks          = 0
2023-03-07 20:33:00,575 INFO blockmanagement.BlockManager: Number of under-replicated blocks = 0
2023-03-07 20:33:00,576 INFO blockmanagement.BlockManager: Number of  over-replicated blocks = 0
2023-03-07 20:33:00,576 INFO blockmanagement.BlockManager: Number of blocks being written    = 0
2023-03-07 20:33:00,576 INFO hdfs.StateChange: STATE* Replication Queue initialization scan for invalid, over- and under
-replicated blocks completed in 9 msec
2023-03-07 20:33:00,607 INFO ipc.Server: IPC Server Responder: starting
2023-03-07 20:33:00,607 INFO ipc.Server: IPC Server listener on 9000: starting
2023-03-07 20:33:00,611 INFO namenode.NameNode: NameNode RPC up at: localhost/127.0.0.1:9000
2023-03-07 20:33:00,614 INFO namenode.FSNamesystem: Starting services required for active state
2023-03-07 20:33:00,614 INFO namenode.FSDirectory: Initializing quota with 4 thread(s)
2023-03-07 20:33:00,622 INFO namenode.FSDirectory: Quota initialization completed in 7 milliseconds
name space=1
storage space=0
storage types=RAM_DISK=0, SSD=0, DISK=0, ARCHIVE=0, PROVIDED=0
2023-03-07 20:33:00,626 INFO blockmanagement.CacheReplicationMonitor: Starting CacheReplicationMonitor with interval 300
00 milliseconds
```

Type JPS command to start-all.cmd command prompt, you will get following output.



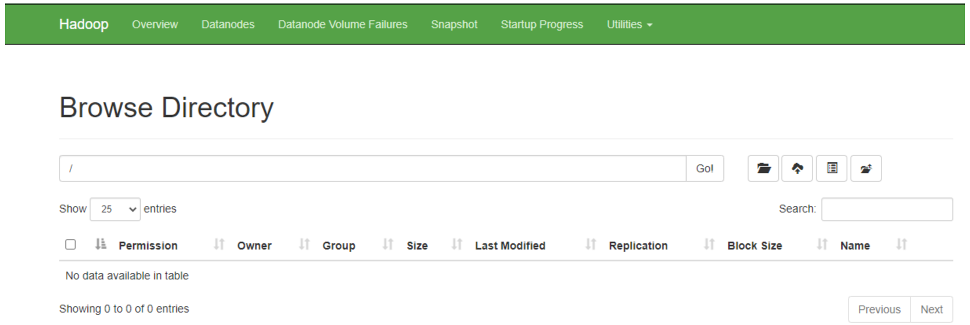
```
C:\hadoop-3.3.0\sbin>jps
5632 Jps
7572 DataNode
3752 ResourceManager
7992 NameNode
8028 NodeManager
```

- Run http://localhost:9870/ from any browser



Overview 'localhost:9000' (✓active)

Started:	Wed Mar 15 12:10:54 +0530 2023
Version:	3.3.0, raa96f1871bfd858f9bac59cf2a81ec470da649af
Compiled:	Tue Jul 07 00:14:00 +0530 2020 by brahma from branch-3.3.0
Cluster ID:	CID-1986aba8-0ed3-43a2-9db7-42944ec518b2
Block Pool ID:	BP-1049743432-192.168.56.1-1678862097216



Browse Directory

/

Show 25 entries

Search:

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
No data available in table								

Showing 0 to 0 of 0 entries