Hadoop Installation

Prepared by: Dev Jethva (23MDS003)

Branch: MTech (Data Science)

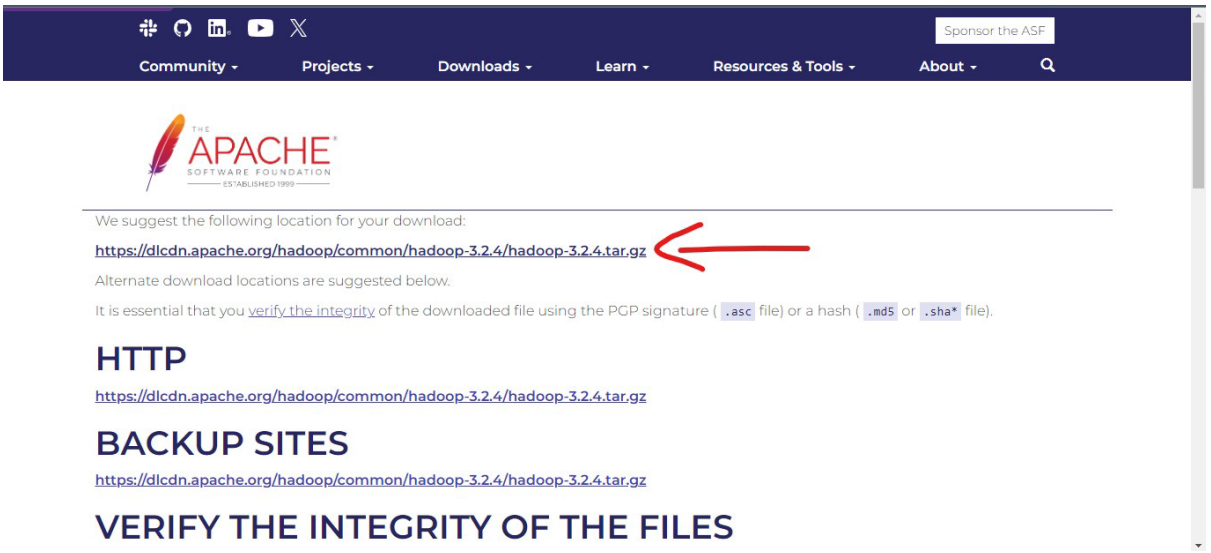
❖ Prerequisites

1. Java 8 runtime environment (JRE): Hadoop 3 requires a Java 8 installation.
2. Java 8 development Kit (JDK)
3. To unzip downloaded Hadoop binaries, we should install 7zip.

❖ Download Hadoop binaries

The first step is to download Hadoop binaries from the official website. we need to install Hadoop 3.2.4 The binary package size is about 470 MB.

https://www.apache.org/dyn/closer.cgi/hadoop/common/hadoop-3.2.4/hadoop-3.2.4.tar.gz

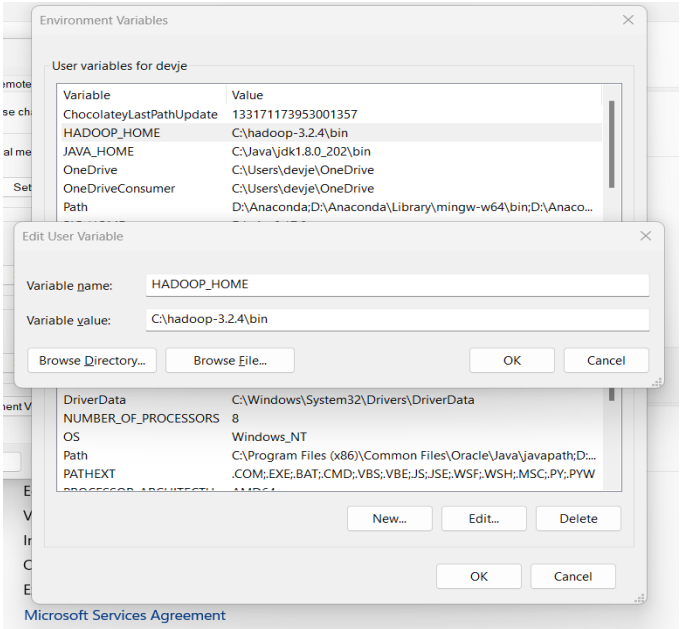


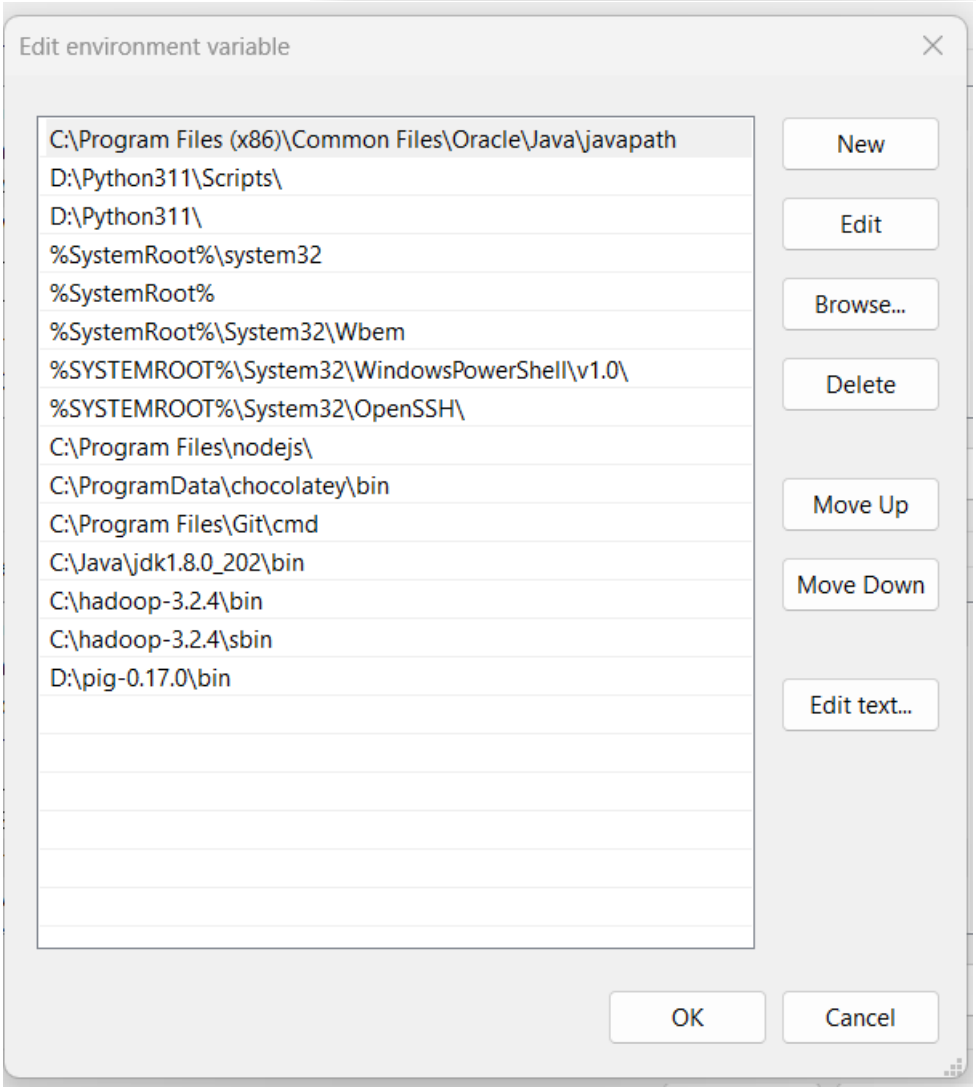
After finishing the file download, we should unpack the package using 7zip. First, we should extract the hadoop-3.2.1.tar.gz library, and then, we should unpack the extracted tar file.

❖ Setting up environment variables

After installing Hadoop and its prerequisites, we should configure the environment variables to define Hadoop and Java default paths.

To edit environment variables, go to Control Panel > System and Security > System (or right-click > properties on My Computer icon) and click on the “Advanced system settings” link.





Checking java and hadoop installation :

Command Prompt

```
Microsoft Windows [Version 10.0.22621.2715]
(c) Microsoft Corporation. All rights reserved.

C:\Users\devje>hadoop version
Hadoop 3.2.4
Source code repository Unknown -r 7e5d9983b388e372fe640f21f048f2f2ae6e9eba
Compiled by ubuntu on 2022-07-12T11:58Z
Compiled with protoc 2.5.0
From source with checksum ee031c16fe785bbb35252c749418712
This command was run using /C:/hadoop-3.2.4/share/hadoop/common/hadoop-common-3.2.4.jar

C:\Users\devje>
```

❖ **Configuring Hadoop cluster**

There are four files we should alter to configure Hadoop cluster:

1. %HADOOP_HOME%\etc\hadoop\hdfs-site.xml

   As we know, Hadoop is built using a master-slave paradigm. Before altering the HDFS configuration file, we should create a directory to store all master node (name node) data and another one to store data (data node). In this example, we created the following directories:

   D:\hadoop\hadoop-3.2.4\data\namenode

   D:\hadoop\hadoop-3.2.4\data\datanode

   Now, let's open "hdfs-site.xml" file located in "%HADOOP_HOME%\etc\hadoop" directory, and we should add the following properties within the <configuration></configuration> element:

```
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
<property>
<name>dfs.namenode.name.dir</name>
<value> file:/// D:\hadoop\hadoop-3.2.4\data\namenode</value>
</property>
<property>
<name>dfs.datanode.data.dir</name>
<value> file:/// D:\hadoop\hadoop-3.2.4\data\datanode</value>
</property>
```

   Note that we have set the replication factor to 1 since we are creating a single node cluster.

2. %HADOOP_HOME%\etc\hadoop\core-site.xml

   Now, we should configure the name node URL adding the following XML code into the <configuration></configuration> element within "core-site.xml":

```
<property>
<name>fs.default.name</name>
```

```
<value>hdfs://localhost:9820</value>
</property>
```

3. %HADOOP_HOME%\etc\hadoop\mapred-site.xml

   Now, we should add the following XML code into the <configuration></configuration> element within “mapred-site.xml”

```
<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
<description>MapReduce framework name</description>
</property>
```

4. %HADOOP_HOME%\etc\hadoop\yarn-site.xml

   Now, we should add the following XML code into the <configuration></configuration> element within “yarn-site.xml”

```
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
<description>Yarn Node Manager Aux Service</description>
</property>
```

❖ **Formatting Name node**

After finishing the configuration, let’s try to format the name node using the following command:

```
hdfs namenode -format
```

```
Command Prompt - hdfs nan
Microsoft Windows [Version 10.0.22621.2715]
(c) Microsoft Corporation. All rights reserved.

C:\Users\devje>hadoop version
Hadoop 3.2.4
Source code repository Unknown -r 7e5d9983b388e372fe640f21f048f2f2ae6e9eba
Compiled by ubuntu on 2022-07-12T11:58Z
Compiled with protoc 2.5.0
From source with checksum ee031c16fe785bbb35252c749418712
This command was run using /C:/hadoop-3.2.4/share/hadoop/common/hadoop-common-3.2.4.jar

C:\Users\devje>hdfs namenode -format
2023-11-22 19:36:02,734 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = DEV/192.168.29.216
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 3.2.4
STARTUP_MSG:   classpath = C:\hadoop-3.2.4\etc\hadoop;C:\hadoop-3.2.4\share\hadoop\common;C:\hadoop-3.2.4\share\hadoop\common\lib\accessors-smart-2.4.7.jar;
C:\hadoop-3.2.4\share\hadoop\common\lib\animal-sniffer-annotations-1.17.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\asm-5.0.4.jar;C:\hadoop-3.2.4\share\hado
op\common\lib\audience-annotations-0.5.0.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\avro-1.7.7.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\checker-qual-2.5
.2.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\commons-beanutils-1.9.4.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\commons-cli-1.2.jar;C:\hadoop-3.2.4\share
\hadoop\common\lib\commons-codec-1.11.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\commons-collections-3.2.2.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\comm
ons-compress-1.21.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\commons-configuration2-2.1.1.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\commons-io-2.8.0.jar;
C:\hadoop-3.2.4\share\hadoop\common\lib\commons-lang3-3.7.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\commons-logging-1.1.3.jar;C:\hadoop-3.2.4\share\hadoop
\common\lib\commons-math3-3.1.1.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\commons-net-3.6.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\commons-text-1.4.jar
;C:\hadoop-3.2.4\share\hadoop\common\lib\curator-client-2.13.0.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\curator-framework-2.13.0.jar;C:\hadoop-3.2.4\shar
e\hadoop\common\lib\curator-recipes-2.13.0.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\dnsjava-2.1.7.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\error_prone
_annotations-2.2.0.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\failureaccess-1.0.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\gson-2.9.0.jar;C:\hadoop-3.2.4\
share\hadoop\common\lib\guava-27.0-jre.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\hadoop-annotations-3.2.4.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\hado
op-auth-3.2.4.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\htrace-core4-4.1.0-incubating.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\httpclient-4.5.13.jar;C:
\hadoop-3.2.4\share\hadoop\common\lib\httpcore-4.4.13.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\j2objc-annotations-1.1.jar;C:\hadoop-3.2.4\share\hadoop\co
mmon\lib\jackson-annotations-2.10.5.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\jackson-core-2.10.5.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\jackson-core
-asl-1.9.13.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\jackson-databind-2.10.5.1.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\jackson-jaxrs-1.9.13.jar;C:\ha
doop-3.2.4\share\hadoop\common\lib\jackson-mapper-asl-1.9.13.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\jackson-xc-1.9.13.jar;C:\hadoop-3.2.4\share\hadoop\
common\lib\javax.activation-api-1.2.0.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\javax.servlet-api-3.1.0.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\jaxb-a
pi-2.2.11.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\jaxb-impl-2.2.3-1.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\jcip-annotations-1.0-1.jar;C:\hadoop-3.2
.4\share\hadoop\common\lib\jersey-core-1.19.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\jersey-json-1.19.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\jersey-
server-1.19.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\jersey-servlet-1.19.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\jettison-1.1.jar;C:\hadoop-3.2.4\sha
re\hadoop\common\lib\jetty-http-9.4.43.v20210629.jar;C:\hadoop-3.2.4\share\hadoop\common\lib\jetty-io-9.4.43.v20210629.jar;C:\hadoop-3.2.4\share\hadoop\comm
```

❖ Starting Hadoop services

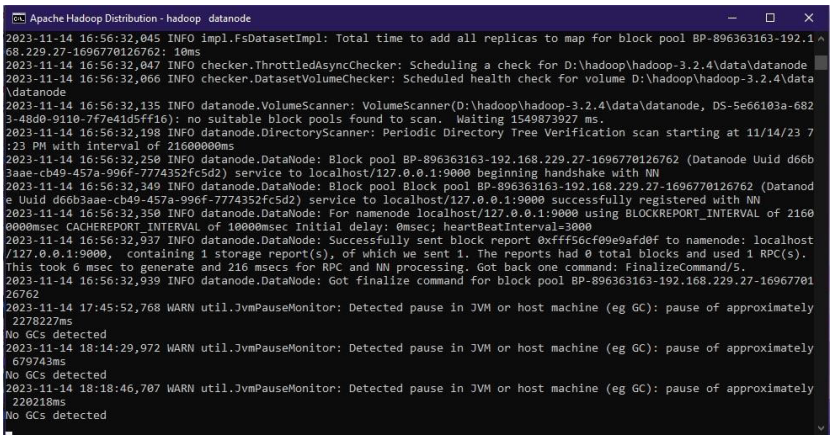
Now, we will open PowerShell, and navigate to "%HADOOP_HOME%\sbin" directory or just open cmd as admin. Then we will run the following command to start the Hadoop nodes:

```
start-all
```

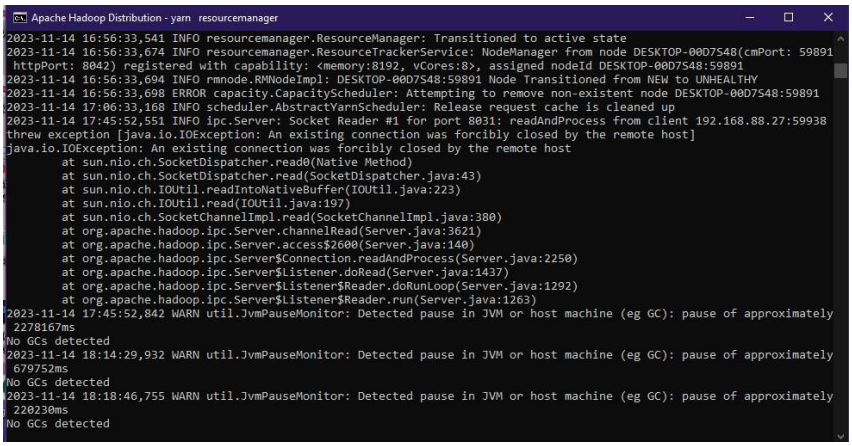
This will run both dfs and yarn, must have to run all 4 terminal , no one have to shutdown , than installation was successful also check this with 'jps' it display all running services.

```
Administrator: Command Prompt
Microsoft Windows [Version 10.0.19045.3570]
(c) Microsoft Corporation. All rights reserved.

C:\WINDOWS\system32>start-all
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons

C:\WINDOWS\system32>jps
8208 Jps
8244 NameNode
3416 DataNode
4408 ResourceManager
6072 NodeManager

C:\WINDOWS\system32>
```

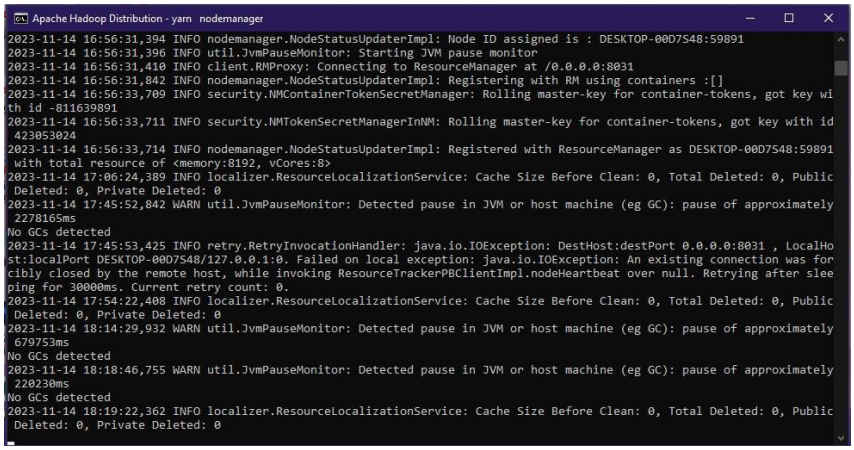
```
Apache Hadoop Distribution - hadoop  namenode
2023-11-14 16:56:29,574 INFO blockmanagement.CacheReplicationMonitor: Starting CacheReplicationMonitor with interval 300
00 milliseconds
2023-11-14 16:56:32,326 INFO hdfs.StateChange: BLOCK* registerDatanode: from DatanodeRegistration(127.0.0.1:9866, datano
deUuid=d66b3aae-cb49-457a-996f-7774352fc5d2, infoPort=9864, infoSecurePort=0, ipcPort=9867, storageInfo=lv=-57;cid=CID-5
cd3e39f-0512-4483-8d07-4828cbb2c942;nsid=1557171817;c=1696770126762) storage d66b3aae-cb49-457a-996f-7774352fc5d2
2023-11-14 16:56:32,329 INFO net.NetworkTopology: Adding a new node: /default-rack/127.0.0.1:9866
2023-11-14 16:56:32,330 INFO blockmanagement.BlockReportLeaseManager: Registered DN d66b3aae-cb49-457a-996f-7774352fc5d2
 (127.0.0.1:9866).
2023-11-14 16:56:32,681 INFO blockmanagement.DatanodeDescriptor: Adding new storage ID DS-5e66103a-6823-48d0-9110-7f7e41
d5ff16 for DN 127.0.0.1:9866
2023-11-14 16:56:32,811 INFO BlockStateChange: BLOCK* processReport 0xfff56cf09e9afd0f: Processing first storage report
for DS-5e66103a-6823-48d0-9110-7f7e41d5ff16 from datanode DatanodeRegistration(127.0.0.1:9866, datanodeUuid=d66b3aae-cb4
9-457a-996f-7774352fc5d2, infoPort=9864, infoSecurePort=0, ipcPort=9867, storageInfo=lv=-57;cid=CID-5cd3e39f-0512-4483-8
d07-4828cbb2c942;nsid=1557171817;c=1696770126762)
2023-11-14 16:56:32,815 INFO BlockStateChange: BLOCK* processReport 0xfff56cf09e9afd0f: from storage DS-5e66103a-6823-48
d0-9110-7f7e41d5ff16 node DatanodeRegistration(127.0.0.1:9866, datanodeUuid=d66b3aae-cb49-457a-996f-7774352fc5d2, infoPo
rt=9864, infoSecurePort=0, ipcPort=9867, storageInfo=lv=-57;cid=CID-5cd3e39f-0512-4483-8d07-4828cbb2c942;nsid=1557171817
;c=1696770126762), blocks: 0, hasStaleStorage: false, processing time: 3 msecs, invalidatedBlocks: 0
2023-11-14 17:45:52,841 WARN util.JvmPauseMonitor: Detected pause in JVM or host machine (eg GC): pause of approximately
 2278164ms
No GCs detected
2023-11-14 17:45:53,562 WARN blockmanagement.HeartbeatManager: Skipping next heartbeat scan due to excessive pause
2023-11-14 18:14:29,932 WARN util.JvmPauseMonitor: Detected pause in JVM or host machine (eg GC): pause of approximately
 679752ms
No GCs detected
2023-11-14 18:14:29,996 WARN blockmanagement.HeartbeatManager: Skipping next heartbeat scan due to excessive pause
2023-11-14 18:18:46,755 WARN util.JvmPauseMonitor: Detected pause in JVM or host machine (eg GC): pause of approximately
 220230ms
No GCs detected
```



```
Apache Hadoop Distribution - hadoop  datanode
2023-11-14 16:56:32,045 INFO impl.FsDatasetImpl: Total time to add all replicas to map for block pool BP-896363163-192.1
68.229.27-1696770126762: 10ms
2023-11-14 16:56:32,047 INFO checker.ThrottledAsyncChecker: Scheduling a check for D:\hadoop\hadoop-3.2.4\data\datanode
2023-11-14 16:56:32,066 INFO checker.DatasetVolumeChecker: Scheduled health check for volume D:\hadoop\hadoop-3.2.4\data
\datanode
2023-11-14 16:56:32,135 INFO datanode.VolumeScanner: VolumeScanner(D:\hadoop\hadoop-3.2.4\data\datanode, DS-5e66103a-682
3-48d0-9110-7f7e41d5ff16): no suitable block pools found to scan.  Waiting 1549873927 ms.
2023-11-14 16:56:32,198 INFO datanode.DirectoryScanner: Periodic Directory Tree Verification scan starting at 11/14/23 7
:23 PM with interval of 21600000ms
2023-11-14 16:56:32,250 INFO datanode.DataNode: Block pool BP-896363163-192.168.229.27-1696770126762 (Datanode Uuid d66b
3aae-cb49-457a-996f-7774352fc5d2) service to localhost/127.0.0.1:9000 beginning handshake with NN
2023-11-14 16:56:32,349 INFO datanode.DataNode: Block pool Block pool BP-896363163-192.168.229.27-1696770126762 (Datanod
e Uuid d66b3aae-cb49-457a-996f-7774352fc5d2) service to localhost/127.0.0.1:9000 successfully registered with NN
2023-11-14 16:56:32,350 INFO datanode.DataNode: For namenode localhost/127.0.0.1:9000 using BLOCKREPORT_INTERVAL of 2160
0000msec CACHEREPORT_INTERVAL of 10000msec Initial delay: 0msec; heartBeatInterval=3000
2023-11-14 16:56:32,937 INFO datanode.DataNode: Successfully sent block report 0xfff56cf09e9afd0f 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 6 msec to generate and 216 msecs for RPC and NN processing. Got back one command: FinalizeCommand/5.
2023-11-14 16:56:32,939 INFO datanode.DataNode: Got finalize command for block pool BP-896363163-192.168.229.27-16967701
26762
2023-11-14 17:45:52,768 WARN util.JvmPauseMonitor: Detected pause in JVM or host machine (eg GC): pause of approximately
 2278227ms
No GCs detected
2023-11-14 18:14:29,972 WARN util.JvmPauseMonitor: Detected pause in JVM or host machine (eg GC): pause of approximately
 679743ms
No GCs detected
2023-11-14 18:18:46,707 WARN util.JvmPauseMonitor: Detected pause in JVM or host machine (eg GC): pause of approximately
 220218ms
No GCs detected
```



```
Apache Hadoop Distribution - yarn  resourcemanager
2023-11-14 16:56:33,541 INFO resourcemanager.ResourceManager: Transitioned to active state
2023-11-14 16:56:33,674 INFO resourcemanager.ResourceTrackerService: NodeManager from node DESKTOP-00D7S48(cmPort: 59891
 httpPort: 8042) registered with capability: <memory:8192, vCores:8>, assigned nodeId DESKTOP-00D7S48:59891
2023-11-14 16:56:33,694 INFO rmnode.RMNodeImpl: DESKTOP-00D7S48:59891 Node Transitioned from NEW to UNHEALTHY
2023-11-14 16:56:33,698 ERROR capacity.CapacityScheduler: Attempting to remove non-existent node DESKTOP-00D7S48:59891
2023-11-14 17:06:33,168 INFO scheduler.AbstractYarnScheduler: Release request cache is cleaned up
2023-11-14 17:45:52,551 INFO ipc.Server: Socket Reader #1 for port 8031: readAndProcess from client 192.168.88.27:59938
threw exception [java.io.IOException: An existing connection was forcibly closed by the remote host]
java.io.IOException: An existing connection was forcibly closed by the remote host
        at sun.nio.ch.SocketDispatcher.read0(Native Method)
        at sun.nio.ch.SocketDispatcher.read(SocketDispatcher.java:43)
        at sun.nio.ch.IOUtil.readIntoNativeBuffer(IOUtil.java:223)
        at sun.nio.ch.IOUtil.read(IOUtil.java:197)
        at sun.nio.ch.SocketChannelImpl.read(SocketChannelImpl.java:380)
        at org.apache.hadoop.ipc.Server.channelRead(Server.java:3621)
        at org.apache.hadoop.ipc.Server.access$2600(Server.java:140)
        at org.apache.hadoop.ipc.Server$Connection.readAndProcess(Server.java:2250)
        at org.apache.hadoop.ipc.Server$Listener.doRead(Server.java:1437)
        at org.apache.hadoop.ipc.Server$Listener$Reader.doRunLoop(Server.java:1292)
        at org.apache.hadoop.ipc.Server$Listener$Reader.run(Server.java:1263)
2023-11-14 17:45:52,842 WARN util.JvmPauseMonitor: Detected pause in JVM or host machine (eg GC): pause of approximately
 2278167ms
No GCs detected
2023-11-14 18:14:29,932 WARN util.JvmPauseMonitor: Detected pause in JVM or host machine (eg GC): pause of approximately
 679752ms
No GCs detected
2023-11-14 18:18:46,755 WARN util.JvmPauseMonitor: Detected pause in JVM or host machine (eg GC): pause of approximately
 220230ms
No GCs detected
```

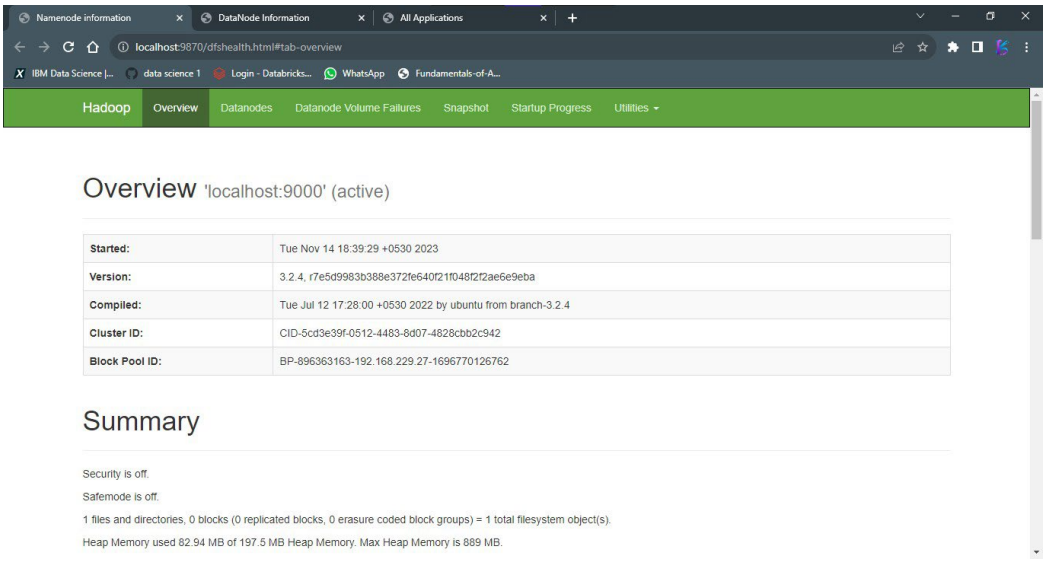


```
Apache Hadoop Distribution - yarn  nodemanager
2023-11-14 16:56:31,394 INFO nodemanager.NodeStatusUpdaterImpl: Node ID assigned is : DESKTOP-00D7S48:59891
2023-11-14 16:56:31,396 INFO util.JvmPauseMonitor: Starting JVM pause monitor
2023-11-14 16:56:31,410 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8031
2023-11-14 16:56:31,842 INFO nodemanager.NodeStatusUpdaterImpl: Registering with RM using containers :[]
2023-11-14 16:56:33,709 INFO security.NMContainerTokenSecretManager: Rolling master-key for container-tokens, got key wi
th id -811639891
2023-11-14 16:56:33,711 INFO security.NMTokenSecretManagerInNM: Rolling master-key for container-tokens, got key with id
 423053024
2023-11-14 16:56:33,714 INFO nodemanager.NodeStatusUpdaterImpl: Registered with ResourceManager as DESKTOP-00D7S48:59891
 with total resource of <memory:8192, vCores:8>
2023-11-14 17:06:24,389 INFO localizer.ResourceLocalizationService: Cache Size Before Clean: 0, Total Deleted: 0, Public
 Deleted: 0, Private Deleted: 0
2023-11-14 17:45:52,842 WARN util.JvmPauseMonitor: Detected pause in JVM or host machine (eg GC): pause of approximately
 2278165ms
No GCs detected
2023-11-14 17:45:53,425 INFO retry.RetryInvocationHandler: java.io.IOException: DestHost:destPort 0.0.0.0:8031 , LocalHo
st:localPort DESKTOP-00D7S48/127.0.0.1:0. Failed on local exception: java.io.IOException: An existing connection was for
cibly closed by the remote host, while invoking ResourceTrackerPBClientImpl.nodeHeartbeat over null. Retrying after slee
ping for 30000ms. Current retry count: 0.
2023-11-14 17:54:22,408 INFO localizer.ResourceLocalizationService: Cache Size Before Clean: 0, Total Deleted: 0, Public
 Deleted: 0, Private Deleted: 0
2023-11-14 18:14:29,932 WARN util.JvmPauseMonitor: Detected pause in JVM or host machine (eg GC): pause of approximately
 679753ms
No GCs detected
2023-11-14 18:18:46,755 WARN util.JvmPauseMonitor: Detected pause in JVM or host machine (eg GC): pause of approximately
 220230ms
No GCs detected
2023-11-14 18:19:22,362 INFO localizer.ResourceLocalizationService: Cache Size Before Clean: 0, Total Deleted: 0, Public
 Deleted: 0, Private Deleted: 0
```

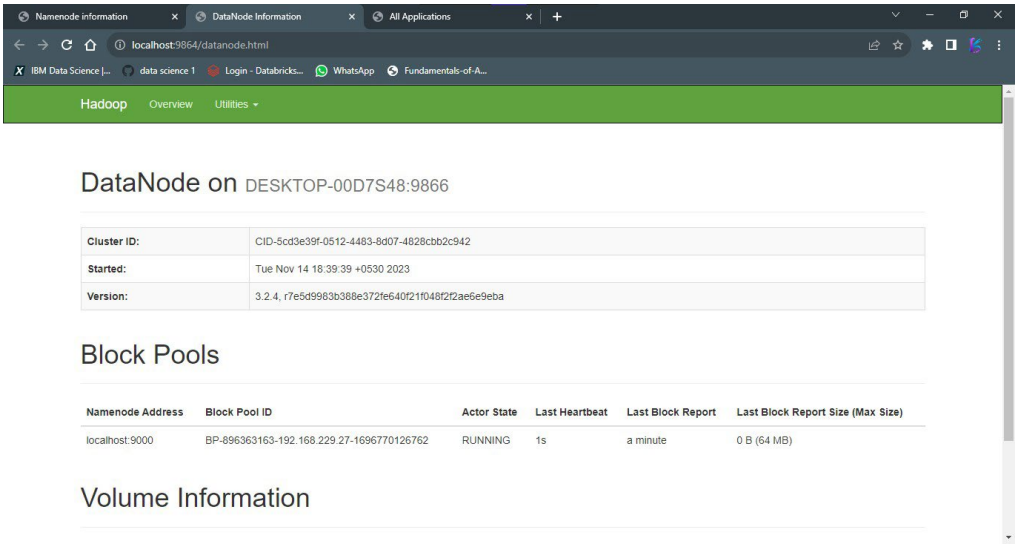
❖ Hadoop Web UI

There are three web user interfaces to be used:

Name node web page: http://localhost:9870/dfshealth.html



Data node web page: http://localhost:9864/datanode.html



Yarn web page: http://localhost:8088/cluster