OBULI SAI NAREN

18CSR125

Installation of Apache Hadoop

TUTORIAL - V

Tutorial 5

Date: 09.05.2021

Apache Hadoop Installation & Execution

Installation of Apache Hadoop on Windows 10

✓ Prerequisites: -

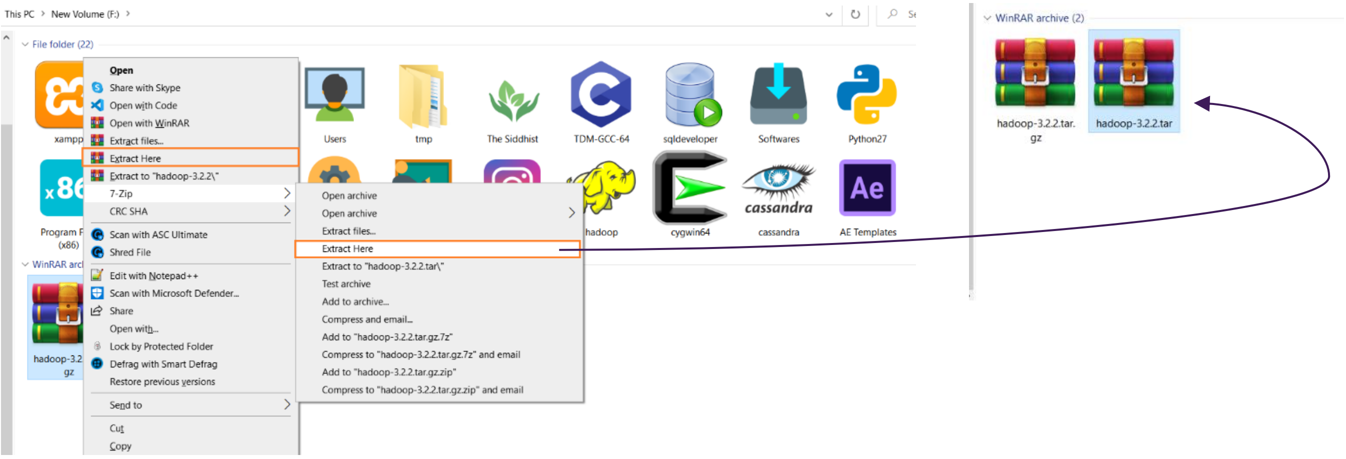
- Java 8 runtime environment (JRE)
- Preferably Java Development Kit (JDK) v1.8_***. (seems to work on all JDK)
- Any Zip Extractors [WinRAR/ 7-Zip]

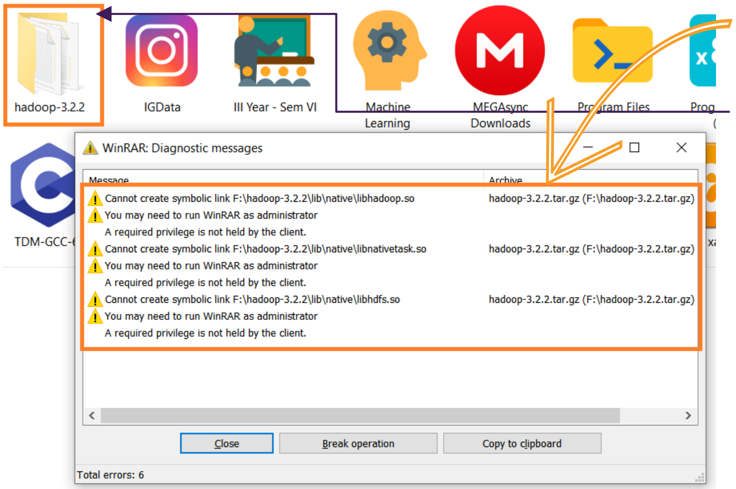
✓ Hadoop Download: -

- Follow the link here 👉 Hadoop Releases .
- Download the one version older than the latest one. (for stability)
- Here I have used hadoop-3.2.2 version.
    - Click here hadoop-3.2.2.tar.gz for direct download.

✓ Hadoop Extraction & Configuring: -

- Extract the downloaded binary [hadoop-3.2.2.tar.gz] file.
    - You will directly get the **Hadoop-3.2.2** folder if you used WinRAR.
    - If you have used 7-Zip you will get " *hadoop-3.2.2.tar* " file which you should extract once again.
- See Image below for Extract Options...!
- Extract Location: “F:\”





WinRAR: Diagnostic messages

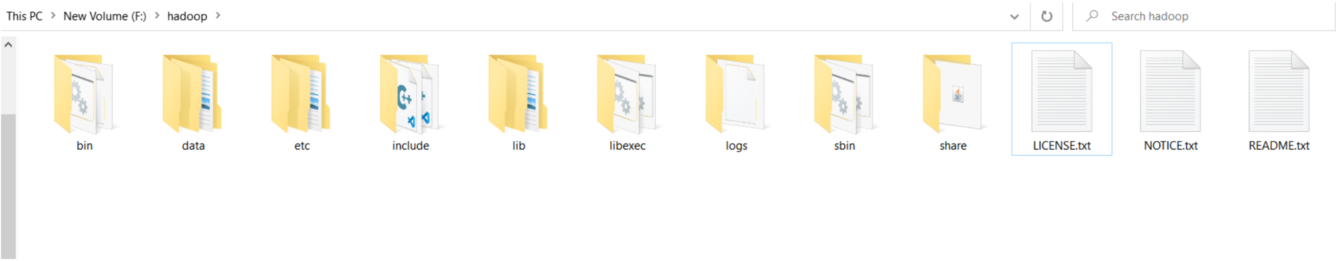
Message	Archive
Cannot create symbolic link F:\hadoop-3.2.2\lib\native\libhadoop.so	hadoop-3.2.2.tar.gz (F:\hadoop-3.2.2.tar.gz)
You may need to run WinRAR as administrator	
A required privilege is not held by the client.	
Cannot create symbolic link F:\hadoop-3.2.2\lib\native\libnativetask.so	hadoop-3.2.2.tar.gz (F:\hadoop-3.2.2.tar.gz)
You may need to run WinRAR as administrator	
A required privilege is not held by the client.	
Cannot create symbolic link F:\hadoop-3.2.2\lib\native\libhdfs.so	hadoop-3.2.2.tar.gz (F:\hadoop-3.2.2.tar.gz)
You may need to run WinRAR as administrator	
A required privilege is not held by the client.	

Total errors: 6

Ignore these warnings that show up at the end.

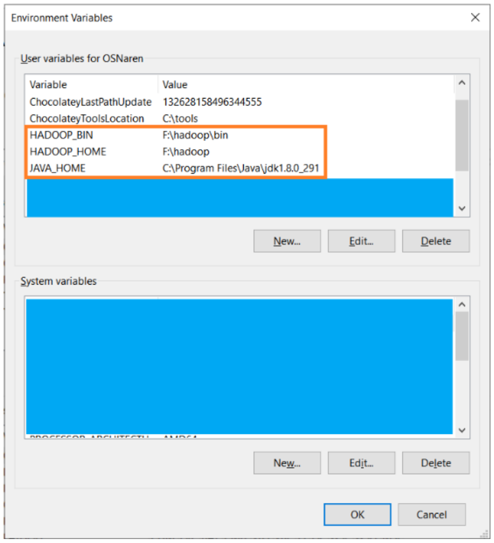
You will get a folder as in the image after extraction. Rename folder “hadoop-3.2.2” ⟶ “hadoop”

Folder format after extraction. *(Ignore ‘data’ folder)*



❖ *Setting up environment variables: -*

- JAVA_HOME: JDK installation folder path : C:\Program Files\Java\jdk1.8.0_291
- HADOOP_HOME: Hadoop installation folder path : F:\hadoop
- HADOOP_BIN: Hadoop Bin path : F:\hadoop\bin



Environment Variables

User variables for OSNaren

Variable	Value
ChocolateyLastPathUpdate	132628158496344555
ChocolateyToolsLocation	C:\tools
HADOOP_BIN	F:\hadoop\bin
HADOOP_HOME	F:\hadoop
JAVA_HOME	C:\Program Files\Java\jdk1.8.0_291

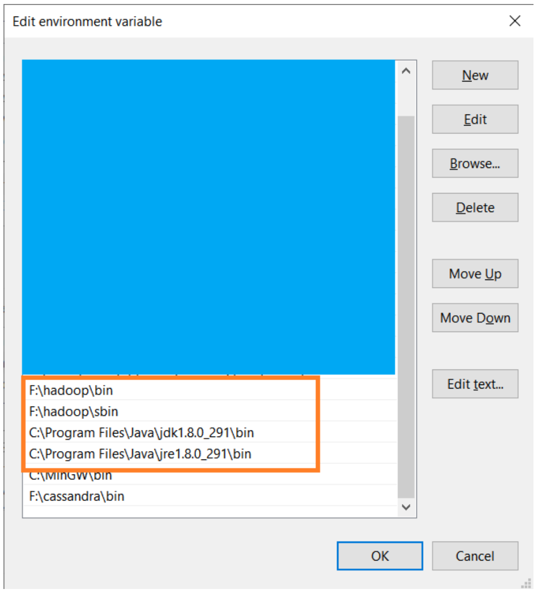
[redacted]

System variables

[redacted]

- Create 3 New variables as shown.

- Add these 4 Path variables as shown.



Edit environment variable

[redacted]

F:\hadoop\bin
F:\hadoop\sbin
C:\Program Files\Java\jdk1.8.0_291\bin
C:\Program Files\Java\jre1.8.0_291\bin
C:\MinGW\bin
F:\cassandra\bin

- ***Configure Hadoop XMLs: -***

  There are five files we should configure in the path below…

- Path: “F:\hadoop\etc\hadoop\”
  - core-site.xml
  - hdfs-site.xml
  - mapred-site.xml
  - yarn-site.xml
  - hadoop-env.cmd

- Just download & extract the files from Config Zip and replace the existing files in the path mentioned.

➢ hdfs-site.xml : Check the path of Namenode & Datanode.

```
<configuration>
    <property>
        <name>dfs.replication</name>
        <value>1</value>
    </property>
    <property>
        <name>dfs.namenode.name.dir</name>
        <value>/F:/hadoop/data/namenode</value>
    </property>
    <property>
        <name>dfs.datanode.data.dir</name>
        <value>/F:/hadoop/data/datanode</value>
    </property>
</configuration>
```

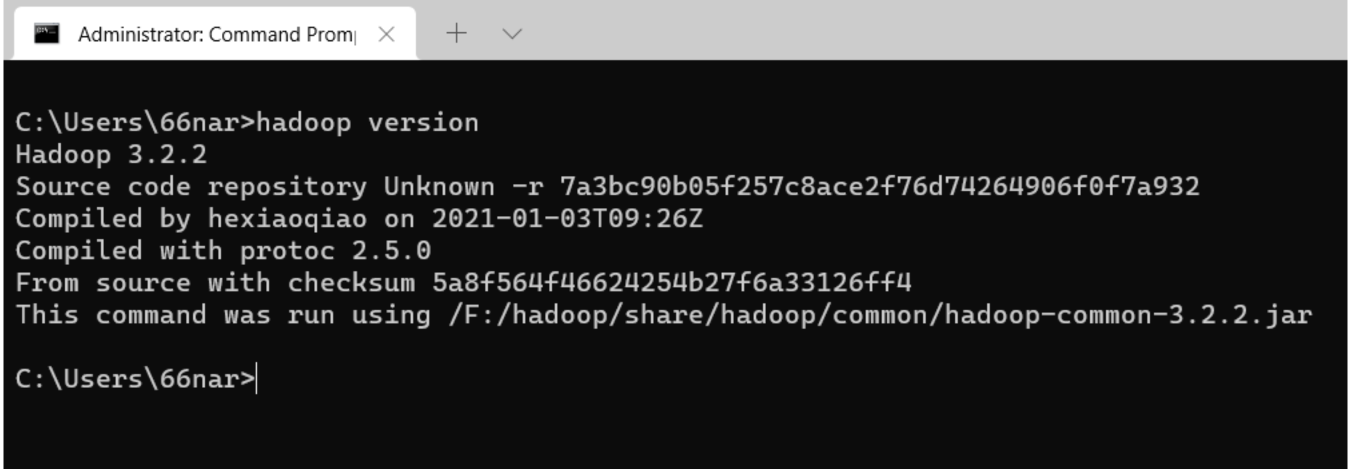
➢ Open "hadoop-env.cmd" & check if the paths are set correct…

```
set JAVA_HOME="C:\Program Files\Java\jdk1.8.0_291"
set HADOOP_PREFIX=%HADOOP_HOME%
set HADOOP_CONF_DIR=%HADOOP_PREFIX%\etc\hadoop
set YARN_CONF_DIR=%HADOOP_CONF_DIR%
set PATH=%PATH%;%HADOOP_PREFIX%\bin
```

- No changes in other 3 files…!
- Now download & extract the files from Bin Zip and replace the existing files in the path mentioned.
    - Path: "F:\hadoop\bin"
- If you are installing some other versions download respective Bin from here☞ Winutils.

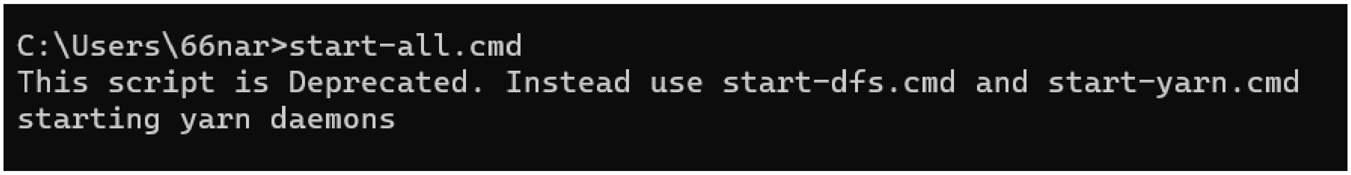
✓ *Running Hadoop: -*

- Open an Administrator Command Prompt.
- Run the command hadoop version and expect an output like below.



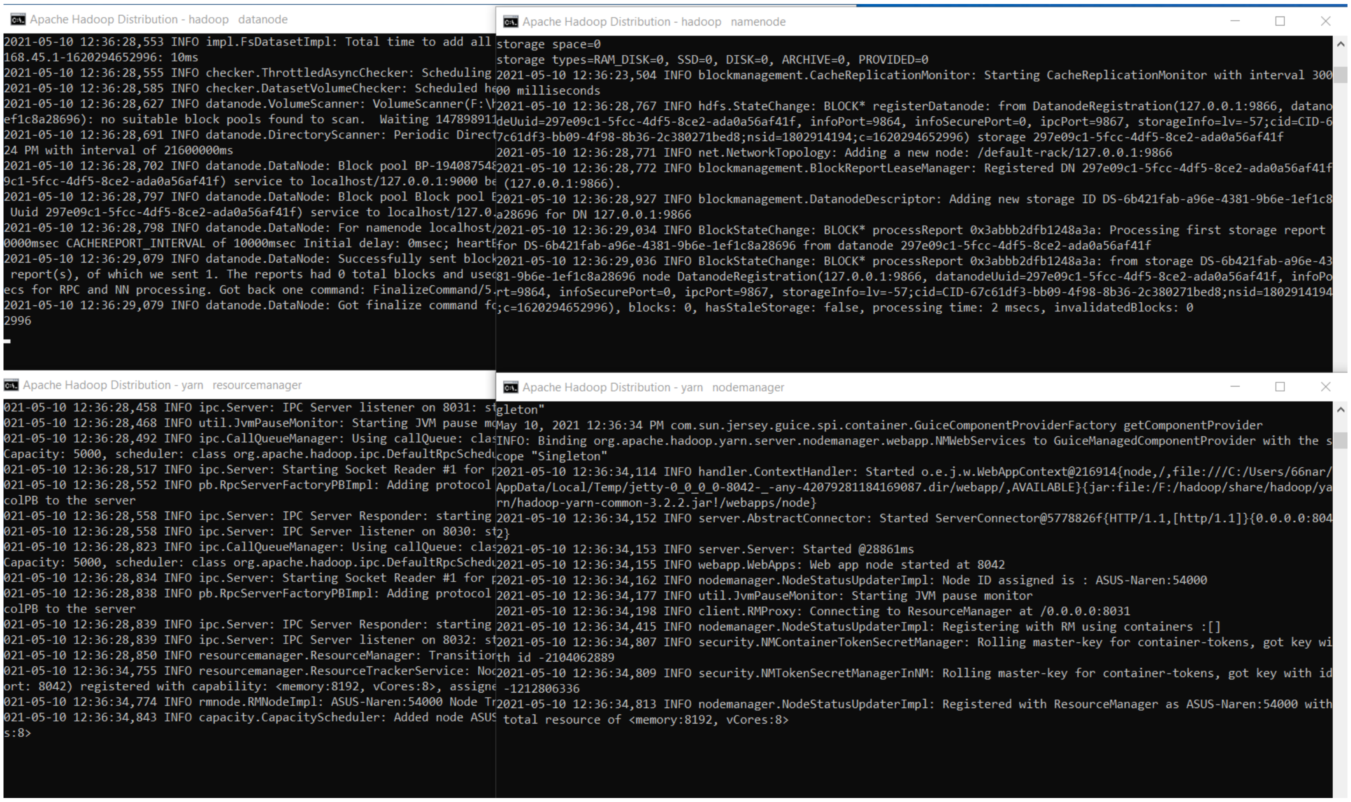
```
C:\Users\66nar>hadoop version
Hadoop 3.2.2
Source code repository Unknown -r 7a3bc90b05f257c8ace2f76d74264906f0f7a932
Compiled by hexiaoqiao on 2021-01-03T09:26Z
Compiled with protoc 2.5.0
From source with checksum 5a8f564f46624254b27f6a33126ff4
This command was run using /F:/hadoop/share/hadoop/common/hadoop-common-3.2.2.jar

C:\Users\66nar>
```

- Now, to format the Namenode → Run the command hdfs namenode -format
  - 🚨 Note – Run this command only once.
  - After long log messages… To the 6th-7th line from the bottom has a message: /F:/hadoop/data/Datanode has been successfully formatted.
- We must start 4 process of Hadoop…
  - Namenode
  - Datanode
  - Resource manager
  - Node manager
  - ➢ Deprecated method → Run the command start-all.cmd



```
C:\Users\66nar>start-all.cmd
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
starting yarn daemons
```

- This opens the 4 processes each in a separate cmd window as below.


Apache Hadoop Distribution - hadoop   datanode
Apache Hadoop Distribution - hadoop   namenode
Apache Hadoop Distribution - yarn   resourcemanager
Apache Hadoop Distribution - yarn   nodemanager

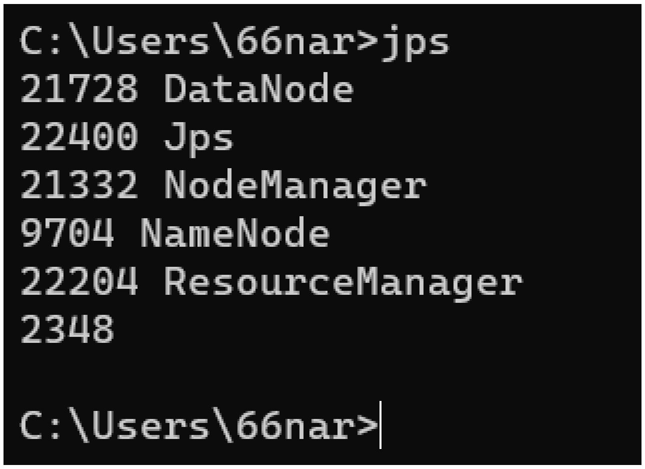
```
2021-05-10 12:36:28,553 INFO impl.FsDatasetImpl: Total time to add all
168.45.1-1620294652996: 10ms
2021-05-10 12:36:28,555 INFO checker.ThrottledAsyncChecker: Scheduling
2021-05-10 12:36:28,585 INFO checker.DatasetVolumeChecker: Scheduled he
2021-05-10 12:36:28,627 INFO datanode.VolumeScanner: VolumeScanner(F:\h
ef1c8a28696): no suitable block pools found to scan.  Waiting 147898911
2021-05-10 12:36:28,691 INFO datanode.DirectoryScanner: Periodic Direct
24 PM with interval of 21600000ms
2021-05-10 12:36:28,702 INFO datanode.DataNode: Block pool BP-19408754
9c1-5fcc-4df5-8ce2-ada0a56af41f) service to localhost/127.0.0.1:9000 be
2021-05-10 12:36:28,797 INFO datanode.DataNode: Block pool Block pool B
 Uuid 297e09c1-5fcc-4df5-8ce2-ada0a56af41f) service to localhost/127.0.
2021-05-10 12:36:28,798 INFO datanode.DataNode: For namenode localhost/
0000msec CACHEREPORT_INTERVAL of 10000msec Initial delay: 0msec; heartB
2021-05-10 12:36:29,079 INFO datanode.DataNode: Successfully sent block
 report(s), of which we sent 1. The reports had 0 total blocks and used
ecs for RPC and NN processing. Got back one command: FinalizeCommand/5.
2021-05-10 12:36:29,079 INFO datanode.DataNode: Got finalize command fo
2996
```

```
storage space=0
storage types=RAM_DISK=0, SSD=0, DISK=0, ARCHIVE=0, PROVIDED=0
2021-05-10 12:36:23,504 INFO blockmanagement.CacheReplicationMonitor: Starting CacheReplicationMonitor with interval 300
00 milliseconds
2021-05-10 12:36:28,767 INFO hdfs.StateChange: BLOCK* registerDatanode: from DatanodeRegistration(127.0.0.1:9866, datano
deUuid=297e09c1-5fcc-4df5-8ce2-ada0a56af41f, infoPort=9864, infoSecurePort=0, ipcPort=9867, storageInfo=lv=-57;cid=CID-6
7c61df3-bb09-4f98-8b36-2c380271bed8;nsid=1802914194;c=1620294652996) storage 297e09c1-5fcc-4df5-8ce2-ada0a56af41f
2021-05-10 12:36:28,771 INFO net.NetworkTopology: Adding a new node: /default-rack/127.0.0.1:9866
2021-05-10 12:36:28,772 INFO blockmanagement.BlockReportLeaseManager: Registered DN 297e09c1-5fcc-4df5-8ce2-ada0a56af41f
 (127.0.0.1:9866).
2021-05-10 12:36:28,927 INFO blockmanagement.DatanodeDescriptor: Adding new storage ID DS-6b421fab-a96e-4381-9b6e-1ef1c8
a28696 for DN 127.0.0.1:9866
2021-05-10 12:36:29,034 INFO BlockStateChange: BLOCK* processReport 0x3abbb2dfb1248a3a: Processing first storage report
for DS-6b421fab-a96e-4381-9b6e-1ef1c8a28696 from datanode 297e09c1-5fcc-4df5-8ce2-ada0a56af41f
2021-05-10 12:36:29,036 INFO BlockStateChange: BLOCK* processReport 0x3abbb2dfb1248a3a: from storage DS-6b421fab-a96e-43
81-9b6e-1ef1c8a28696 node DatanodeRegistration(127.0.0.1:9866, datanodeUuid=297e09c1-5fcc-4df5-8ce2-ada0a56af41f, infoPo
rt=9864, infoSecurePort=0, ipcPort=9867, storageInfo=lv=-57;cid=CID-67c61df3-bb09-4f98-8b36-2c380271bed8;nsid=1802914194
;c=1620294652996), blocks: 0, hasStaleStorage: false, processing time: 2 msecs, invalidatedBlocks: 0
```

```
021-05-10 12:36:28,458 INFO ipc.Server: IPC Server listener on 8031: st
021-05-10 12:36:28,468 INFO util.JvmPauseMonitor: Starting JVM pause mo
021-05-10 12:36:28,492 INFO ipc.CallQueueManager: Using callQueue: clas
Capacity: 5000, scheduler: class org.apache.hadoop.ipc.DefaultRpcSchedu
021-05-10 12:36:28,517 INFO ipc.Server: Starting Socket Reader #1 for p
021-05-10 12:36:28,552 INFO pb.RpcServerFactoryPBImpl: Adding protocol
colPB to the server
021-05-10 12:36:28,558 INFO ipc.Server: IPC Server Responder: starting
021-05-10 12:36:28,558 INFO ipc.Server: IPC Server listener on 8030: st
021-05-10 12:36:28,823 INFO ipc.CallQueueManager: Using callQueue: clas
Capacity: 5000, scheduler: class org.apache.hadoop.ipc.DefaultRpcSchedu
021-05-10 12:36:28,834 INFO ipc.Server: Starting Socket Reader #1 for p
021-05-10 12:36:28,838 INFO pb.RpcServerFactoryPBImpl: Adding protocol
colPB to the server
021-05-10 12:36:28,839 INFO ipc.Server: IPC Server Responder: starting
021-05-10 12:36:28,839 INFO ipc.Server: IPC Server listener on 8032: st
021-05-10 12:36:28,850 INFO resourcemanager.ResourceManager: Transition
021-05-10 12:36:34,755 INFO resourcemanager.ResourceTrackerService: Nod
ort: 8042) registered with capability: <memory:8192, vCores:8>, assigne
021-05-10 12:36:34,774 INFO rmnode.RMNodeImpl: ASUS-Naren:54000 Node Tr
021-05-10 12:36:34,843 INFO capacity.CapacityScheduler: Added node ASUS
s:8>
```

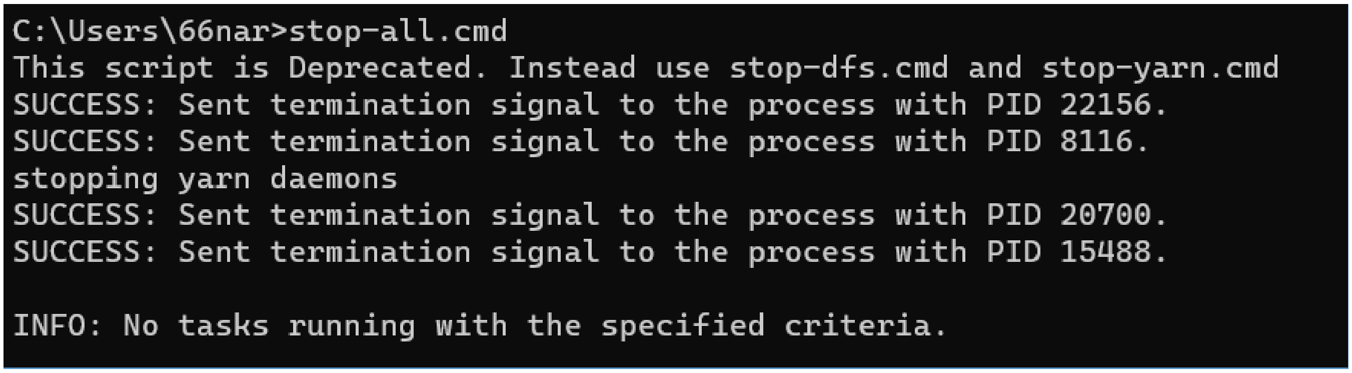
```
gleton"
May 10, 2021 12:36:34 PM com.sun.jersey.guice.spi.container.GuiceComponentProviderFactory getComponentProvider
INFO: Binding org.apache.hadoop.yarn.server.nodemanager.webapp.NMWebServices to GuiceManagedComponentProvider with the s
cope "Singleton"
2021-05-10 12:36:34,114 INFO handler.ContextHandler: Started o.e.j.w.WebAppContext@216914{node,/,file:///C:/Users/66nar/
AppData/Local/Temp/jetty-0_0_0_0-8042-_-any-42079281184169087.dir/webapp/,AVAILABLE}{jar:file:/F:/hadoop/share/hadoop/ya
rn/hadoop-yarn-common-3.2.2.jar!/webapps/node}
2021-05-10 12:36:34,152 INFO server.AbstractConnector: Started ServerConnector@5778826f{HTTP/1.1,[http/1.1]}{0.0.0.0:804
2}
2021-05-10 12:36:34,153 INFO server.Server: Started @28861ms
2021-05-10 12:36:34,155 INFO webapp.WebApps: Web app node started at 8042
2021-05-10 12:36:34,162 INFO nodemanager.NodeStatusUpdaterImpl: Node ID assigned is : ASUS-Naren:54000
2021-05-10 12:36:34,177 INFO util.JvmPauseMonitor: Starting JVM pause monitor
2021-05-10 12:36:34,198 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8031
2021-05-10 12:36:34,415 INFO nodemanager.NodeStatusUpdaterImpl: Registering with RM using containers :[]
2021-05-10 12:36:34,807 INFO security.NMContainerTokenSecretManager: Rolling master-key for container-tokens, got key wi
th id -2104062889
2021-05-10 12:36:34,809 INFO security.NMTokenSecretManagerInNM: Rolling master-key for container-tokens, got key with id
 -1212806336
2021-05-10 12:36:34,813 INFO nodemanager.NodeStatusUpdaterImpl: Registered with ResourceManager as ASUS-Naren:54000 with
 total resource of <memory:8192, vCores:8>
```

- Or you can → Run the commands:
  - start-dfs.cmd - Starts namenode and datanode.
  - start-yarn.cmd - Starts node manager and resource manager.
- To check whether all the process has started and running properly:
  - Run the command → *jps*



```
C:\Users\66nar>jps
21728 DataNode
22400 Jps
21332 NodeManager
9704 NameNode
22204 ResourceManager
2348

C:\Users\66nar>
```

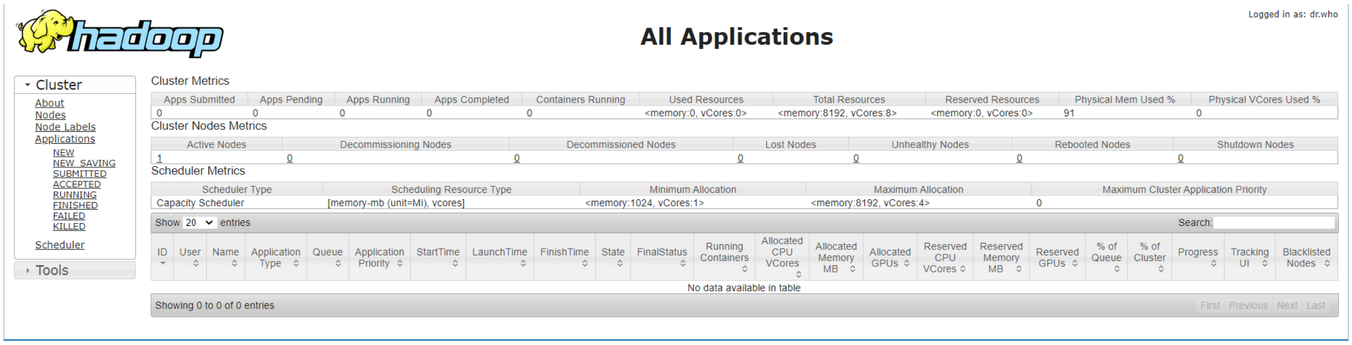
- To stop the processes: Use respective commands.
  - stop-all.cmd
  - stop-dfs.cmd
  - stop-yarn.cmd



```
C:\Users\66nar>stop-all.cmd
This script is Deprecated. Instead use stop-dfs.cmd and stop-yarn.cmd
SUCCESS: Sent termination signal to the process with PID 22156.
SUCCESS: Sent termination signal to the process with PID 8116.
stopping yarn daemons
SUCCESS: Sent termination signal to the process with PID 20700.
SUCCESS: Sent termination signal to the process with PID 15488.

INFO: No tasks running with the specified criteria.
```

Hadoop Web UI: -

- These 3-localhost address works if hadoop is running without and errors.
  - http://localhost:8088 - *Yarn page*



All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	Total Resources	Reserved Resources	Physical Mem Used %	Physical VCores Used %
0	0	0	0	0	<memory:0, vCores:0>	<memory:8192, vCores:8>	<memory:0, vCores:0>	91	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
1	0	0	0	0	0	0

Scheduler Metrics

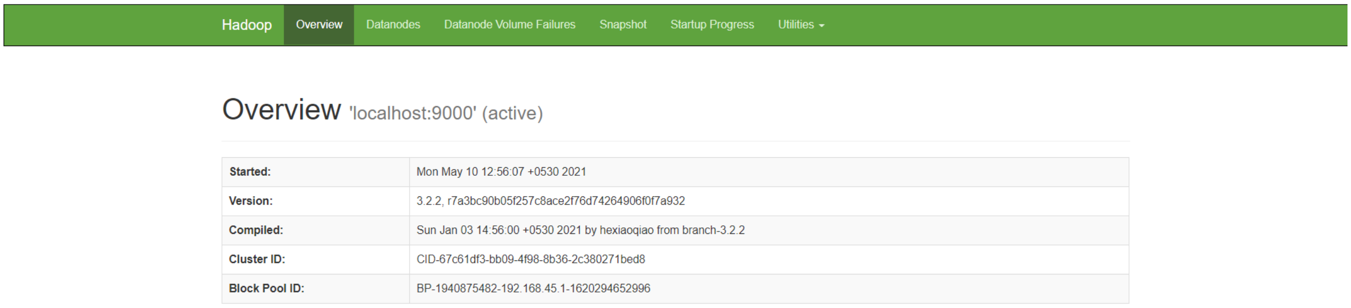
Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0

Show 20 entries Search:

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Allocated GPUs	Reserved CPU VCores	Reserved Memory MB	Reserved GPUs	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes
No data available in table																						

Showing 0 to 0 of 0 entries First Previous Next Last

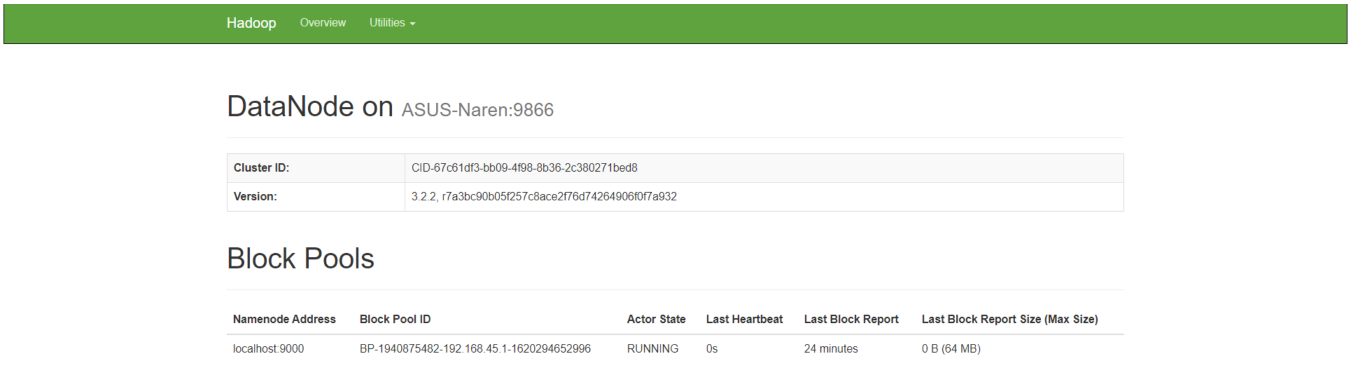
  - http://localhost:9870 - *Name node page*



Overview 'localhost:9000' (active)

Started:	Mon May 10 12:56:07 +0530 2021
Version:	3.2.2, r7a3bc90b05f257c8ace2f76d74264906f0f7a932
Compiled:	Sun Jan 03 14:56:00 +0530 2021 by hexiaoqiao from branch-3.2.2
Cluster ID:	CID-67c61df3-bb09-4f98-8b36-2c380271bed8
Block Pool ID:	BP-1940875482-192.168.45.1-1620294652996

- http://localhost:9864 - *Data node page*



DataNode on ASUS-Naren:9866

Cluster ID:	CID-67c61df3-bb09-4f98-8b36-2c380271bed8
Version:	3.2.2, r7a3bc90b05f257c8ace2f76d74264906f0f7a932

Block Pools

Namenode Address	Block Pool ID	Actor State	Last Heartbeat	Last Block Report	Last Block Report Size (Max Size)
localhost:9000	BP-1940875482-192.168.45.1-1620294652996	RUNNING	0s	24 minutes	0 B (64 MB)

✓ Handling Errors in Installation Process: -

1. *JAVA_HOME is incorrectly set :–*

a. Use “Progra~1” instead of “Program Files”

b. Use “Progra~2” instead of “Program Files(x86)”

- Replace this wherever you have given JAVA path.
    - hadoop-env.cmd – Path variables – JAVA_HOME

2. *Datanode shutting down / Not starting / Exception in datanode cmd :–*

a. Copy the Cluster ID from the namenode VERSION file in the directory → *“F:\hadoop\data\namenode\current”.*

b. Paste the Cluster ID that you copied to the datanode VERSION file in the directory → *“F:\hadoop\data\datanode\current”.*

- Do not give *hdfs namenode -format* more than once, this resets the Cluster ID in the namenode. So, you have to copy & paste each and every time.

3. ***Exception while formatting Namenode / formatting failure :–***
    a. Run the ***hdfs namenode -format*** command in administrator command prompt.
    b. Do not use hadoop-3.2.1
    c. If you are using hadoop-3.2.1…
        - Download hadoop-hdfs-3.2.1.jar file from the link.
        - Rename the file name *hadoop-hdfs-3.2.1.jar* to *hadoop-hdfs-3.2.1.bak* in folder %HADOOP_HOME%\share\hadoop\hdfs
        - Copy the downloaded hadoop-hdfs-3.2.1.jar to folder %HADOOP_HOME%\share\hadoop\hdfs

4. ***Exception in Resource manager / Not running :–***
    a. Check whether the file *"hadoop-yarn-server-resourcemanager-3.2.2.jar"* is present in the locations…
        - F:\hadoop\share\hadoop\yarn
        - F:\hadoop\share\hadoop\yarn\sources
    b. If not copy & paste from the directory that has the file to the directory that does not have the file.

Hadoop MapReduce

✓ *MapReduce Program Source Code*

Hadoop.java.html Choices.java.html MapMR.java.html

✓ *MapReduce Sample Data*

Sample Data.html

✓ Hadoop using IntelliJ – Follow this link to run your Hadoop MapReduce program in IntelliJ.

✓ Hadoop using Eclipse – Follow this link to run your Hadoop MapReduce program in Eclipse.