Apache Hadoop

History Of Hadoop:

Hadoop originated from Google's research on the Google File System (GFS) and MapReduce, leading Doug Cutting and Mike Cafarella to develop it as part of the Nutch project in 2004. By 2006, Hadoop became a top-level project of the Apache Software Foundation, with its first official release offering basic MapReduce and HDFS functionalities. Over the next few years, it gained traction with significant milestones, such as the introduction of the YARN resource management framework in Hadoop 1.x (2010) and the Hadoop 2.0 release (2012) that included YARN for better scalability and resource management. The ecosystem expanded with additional projects like Apache Hive and Apache HBase, cementing Hadoop's role in big data processing.

In recent years, Hadoop continued to evolve with the release of Hadoop 3.x in 2017, which brought improvements like erasure coding and enhanced YARN capabilities. The platform saw growing adoption of complementary tools such as Apache Spark, and its integration with cloud-based platforms became more prevalent. The Hadoop community remains active, focusing on advancements in performance, security, and ecosystem integration, maintaining Hadoop's relevance and utility in handling large-scale data processing tasks.

Versions of Hadoop:

Hadoop 0.x Series

- **0.1.0 (2006):** The initial release included basic implementations of MapReduce and HDFS.
- **0.17 (2008):** Added significant features and stability improvements, including the introduction of the JobTracker and TaskTracker components.

Hadoop 1.x Series

- **1.0.0 (2012):** Marked the release of Hadoop 1.x, featuring the YARN (Yet Another Resource Negotiator) architecture. This version decoupled MapReduce from resource management, enabling better resource allocation and scalability.

Hadoop 2.x Series

- **2.0.0 (2013):** Official release of Hadoop 2.x with YARN as the default resource manager, improving scalability and flexibility by allowing non-MapReduce applications to run on the Hadoop cluster.
- **2.7.x (2015-2016):** Introduced features such as improved HDFS high availability and enhancements to YARN for better resource management and performance.

Hadoop 3.x Series

- **3.0.0 (2017):** Major release that included support for erasure coding, which improves storage efficiency and fault tolerance. Enhanced YARN for better resource management and the addition of the Hadoop Distributed File System (HDFS) Federation.
- **3.1.x (2018):** Added support for dynamic resource pools and improved support for running multiple versions of YARN.
- **3.2.x (2019):** Included more stability improvements, additional features for performance, and further enhancements to YARN and HDFS.

Hadoop 3.3.x Series

- **3.3.0 (2020):** Focused on continuing improvements, including better scalability, security features, and performance optimizations.

Hadoop 4.x Series

- **4.0.x (2024):** Expected to bring further advancements in scalability, performance, and integration with modern data processing tools and cloud environments. (Note: As of the last update, Hadoop 4.x is

anticipated and in development phases; exact features and release details may evolve.)

System Requirements for Hadoop:

Hadoop System Requirements for Windows:

1. Operating System:
    - Windows 10, Windows Server 2016/2019/2022 (Windows Subsystem for Linux or Docker may be used for a more compatible environment).
2. Hardware:
    - **CPU**: 2 GHz multi-core processor (minimum); more cores are recommended for better performance.
    - **RAM**: Minimum 8 GB of RAM (16 GB or more is recommended for larger datasets).
    - **Disk Space**: At least 100 GB of free disk space (more if handling large datasets).
3. Software:
    - **Java**: JDK 8 or JDK 11 (Hadoop 3.x and later versions support Java 11).
    - **SSH**: For a full Hadoop setup, you'll need SSH; however, on Windows, you can use tools like PuTTY or Cygwin for SSH capabilities.
    - **Hadoop Distribution**: Download and configure a Hadoop distribution compatible with Windows (e.g., Hadoop binaries or using WSL/Docker).

Installation Steps with Commands:

- Install Java SDK and set the path in environment variables.
- Download Hadoop and set its path in environment variables.
  `C:\hadoop\bin`
- Configure Hadoop core-site.xml file.
  `<configuration>`

```
<property>
<name>hadoop.tmp.dir</name>
<value>/Users/<YOUR_COMPUTER_NAME>/hdfs/tmp/</value>
</property>
<property>
<name>fs.default.name</name>
<value>hdfs://127.0.0.1:9000</value>
</property>
</configuration>
```

- Configure Hadoop hdfs-site.xml file.

```
<configuration>
<property>
<name>dfs.data.dir</name>
<value>/Users/<YOUR_COMPUTER_NAME>/hdfs/namenode</value>
</property>
<property>
<name>dfs.data.dir</name>
<value>/Users/<YOUR_COMPUTER_NAME>/hdfs/datanode</value>
</property>
<property><name>dfs.replication</name>
<value>1</value>
</property>
</configuration>
```

- Configure Hadoop mapred-site.xml

```
<configuration>
<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
</property>
</configuration>
```

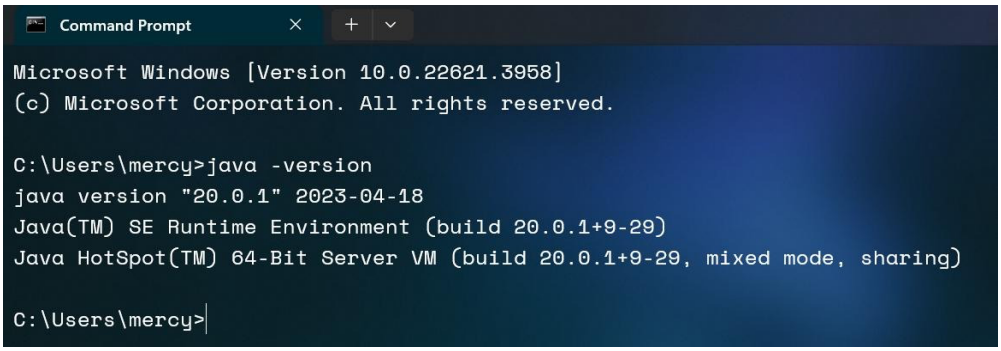
- Configure Hadoop yarn-site.xml

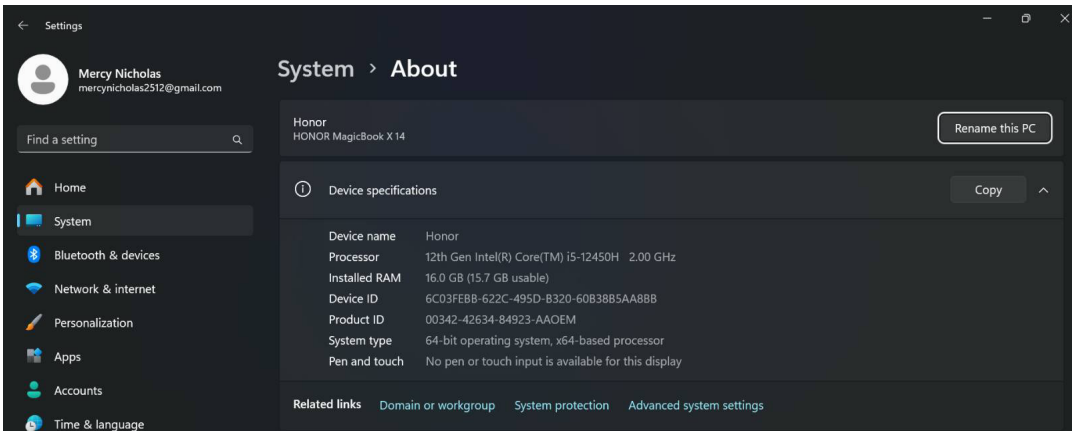
```
<configuration>
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
```

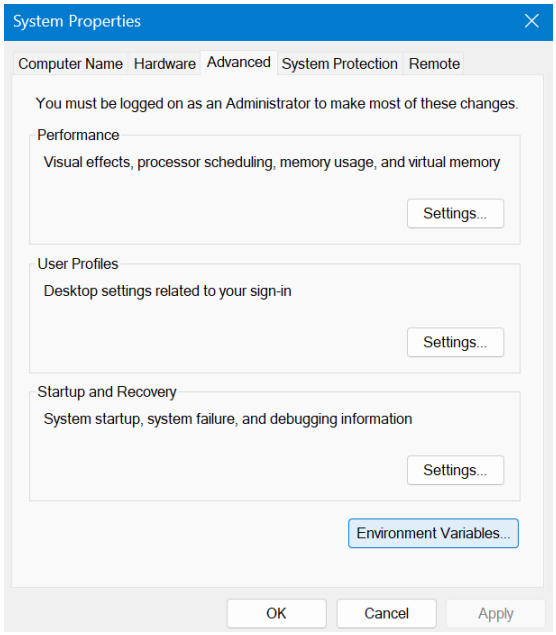
```
</property>
<property>
<name>yarn.nodemanager.aux-
services.mapreduce.shuffle.class</name>
<value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
<property>
<name>yarn.resourcemanager.hostname</name>
<value>127.0.0.1</value>
</property>
<property>
<name>yarn.acl.enable</name>
<value>0</value>
</property><property>
<name>yarn.nodemanager.env-whitelist</name>
<value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOO
P_
CONF_DIR,CLASSPATH_PERPEND_DISTCACHE,HADOOP_YARN_HOME,HADO
OP
_MAPRED_HOME</value>
</property>
</configuration>
```

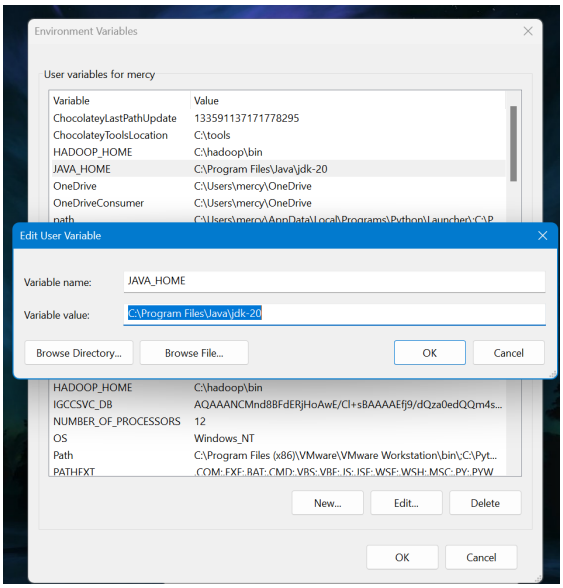
- Start Hadoop
  `start-all.sh`

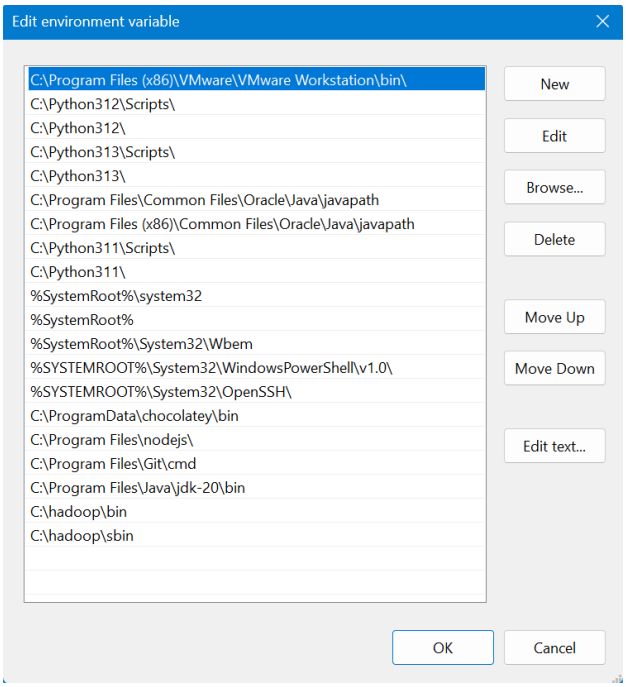
Installation Screenshots

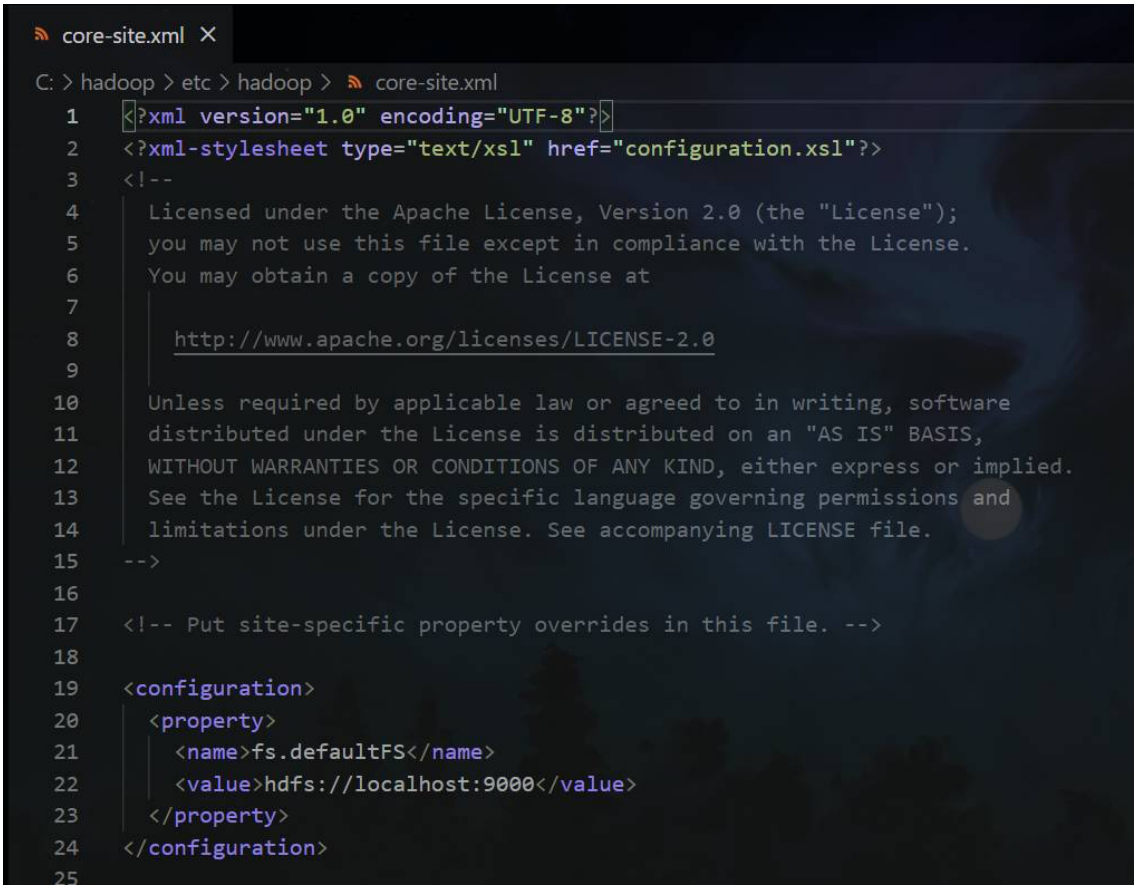



Settings
Mercy Nicholas
mercynicholas2512@gmail.com
Find a setting
Home
System
Bluetooth & devices
Network & internet
Personalization
Apps
Accounts
Time & language
System › About
Honor
HONOR MagicBook X 14
Rename this PC
Device specifications
Copy
Device name Honor
Processor 12th Gen Intel(R) Core(TM) i5-12450H 2.00 GHz
Installed RAM 16.0 GB (15.7 GB usable)
Device ID 6C03FEBB-622C-495D-B320-60B38B5AA8BB
Product ID 00342-42634-84923-AAOEM
System type 64-bit operating system, x64-based processor
Pen and touch No pen or touch input is available for this display
Related links Domain or workgroup System protection Advanced system settings


System Properties
Computer Name Hardware Advanced System Protection Remote
You must be logged on as an Administrator to make most of these changes.
Performance
Visual effects, processor scheduling, memory usage, and virtual memory
Settings...
User Profiles
Desktop settings related to your sign-in
Settings...
Startup and Recovery
System startup, system failure, and debugging information
Settings...
Environment Variables...
OK
Cancel
Apply


Environment Variables
User variables for mercy
Variable Value
ChocolateyLastPathUpdate 133591137171778295
ChocolateyToolsLocation C:\tools
HADOOP_HOME C:\hadoop\bin
JAVA_HOME C:\Program Files\Java\jdk-20
OneDrive C:\Users\mercy\OneDrive
OneDriveConsumer C:\Users\mercy\OneDrive
Edit User Variable
Variable name: JAVA_HOME
Variable value: C:\Program Files\Java\jdk-20
Browse Directory...
Browse File...
OK
Cancel
HADOOP_HOME C:\hadoop\bin
IGCCSVC_DB AQAAANCMnd8BFdERjHoAwE/Cl+sBAAAAEfj9/dQza0edQQm4s...
NUMBER_OF_PROCESSORS 12
OS Windows_NT
Path C:\Program Files (x86)\VMware\VMware Workstation\bin\;C:\Pyt...
PATHEXT .COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC;.PY;.PYW
New...
Edit...
Delete
OK
Cancel


Edit environment variable
C:\Program Files (x86)\VMware\VMware Workstation\bin\
C:\Python312\Scripts\
C:\Python312\
C:\Python313\Scripts\
C:\Python313\
C:\Program Files\Common Files\Oracle\Java\javapath
C:\Program Files (x86)\Common Files\Oracle\Java\javapath
C:\Python311\Scripts\
C:\Python311\
%SystemRoot%\system32
%SystemRoot%
%SystemRoot%\System32\Wbem
%SYSTEMROOT%\System32\WindowsPowerShell\v1.0\
%SYSTEMROOT%\System32\OpenSSH\
C:\ProgramData\chocolatey\bin
C:\Program Files\nodejs\
C:\Program Files\Git\cmd
C:\Program Files\Java\jdk-20\bin
C:\hadoop\bin
C:\hadoop\sbin
New
Edit
Browse...
Delete
Move Up
Move Down
Edit text...
OK
Cancel


core-site.xml
C: > hadoop > etc > hadoop > core-site.xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at
http://www.apache.org/licenses/LICENSE-2.0
Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->
<!-- Put site-specific property overrides in this file. -->
<configuration>
<property>
<name>fs.defaultFS</name>
<value>hdfs://localhost:9000</value>
</property>
</configuration>

hdfs-site.xml

C: > hadoop > etc > hadoop > hdfs-site.xml

```xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

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

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

mapred-site.xml

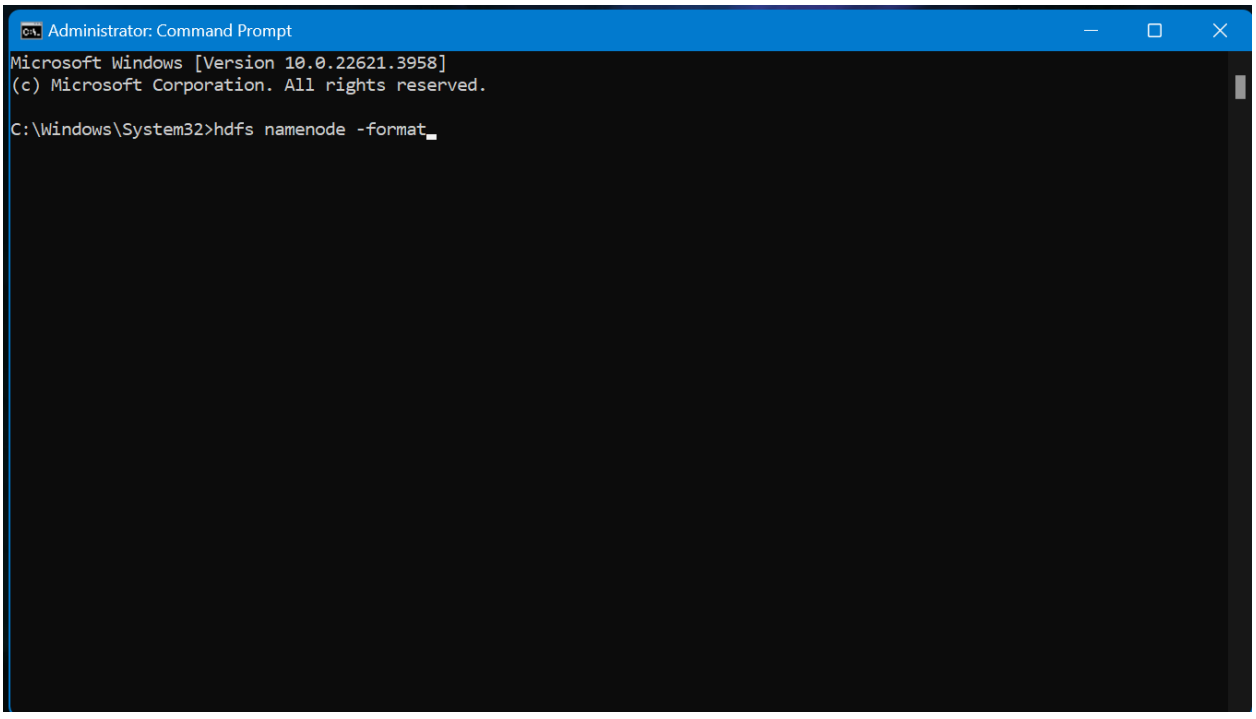
C: > hadoop > etc > hadoop > mapred-site.xml

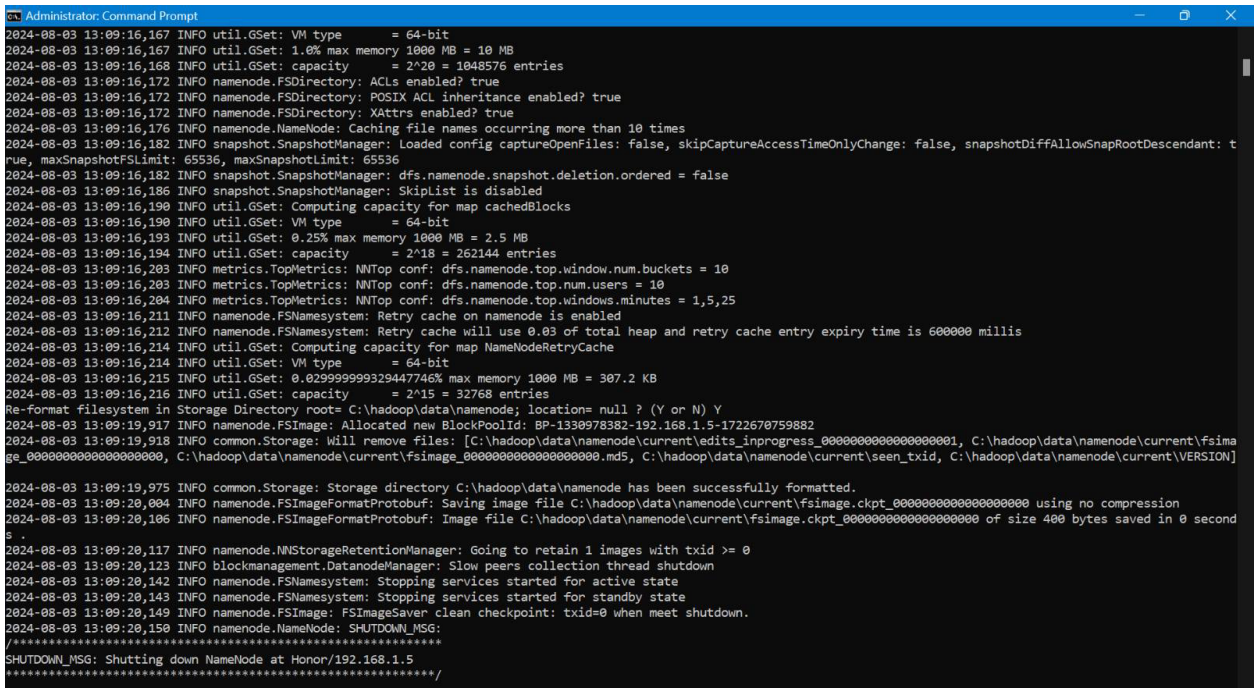
```xml
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

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

<configuration>
  <property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
  </property>
</configuration>

```

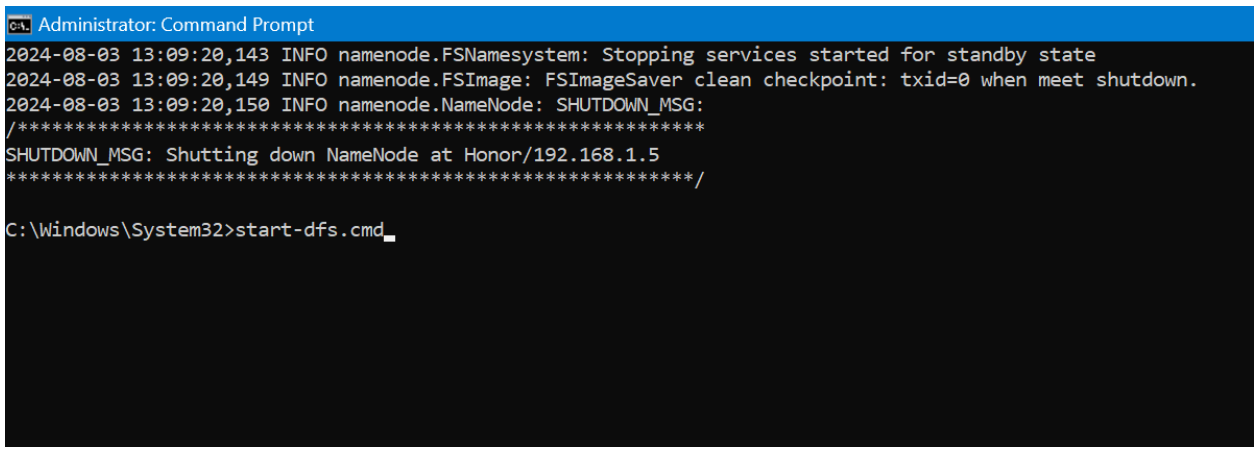
yarn-site.xml

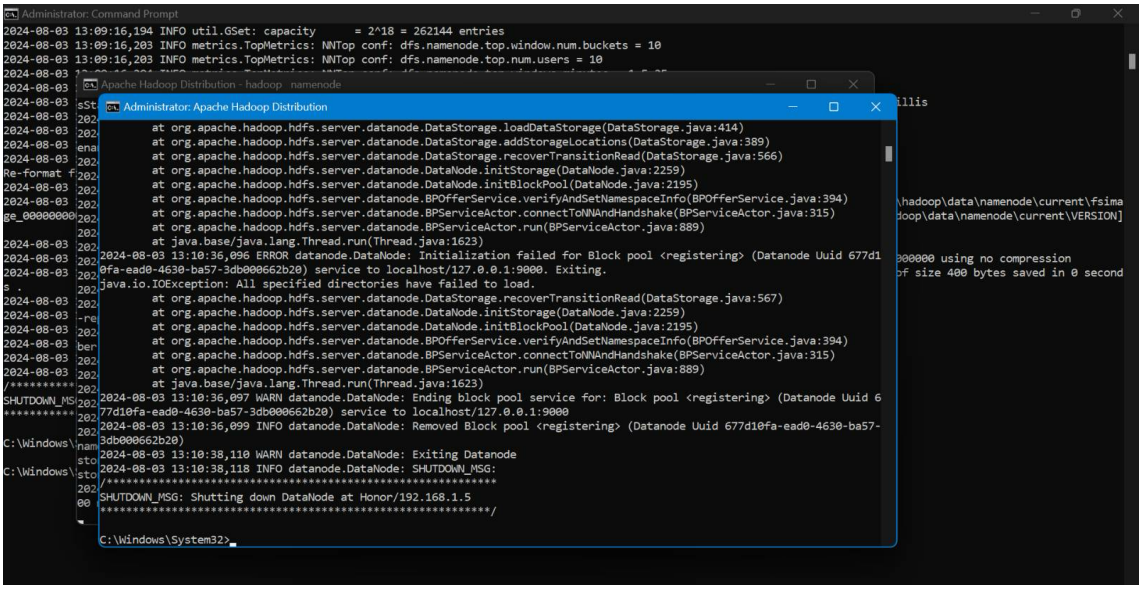
C: > hadoop > etc > hadoop > yarn-site.xml

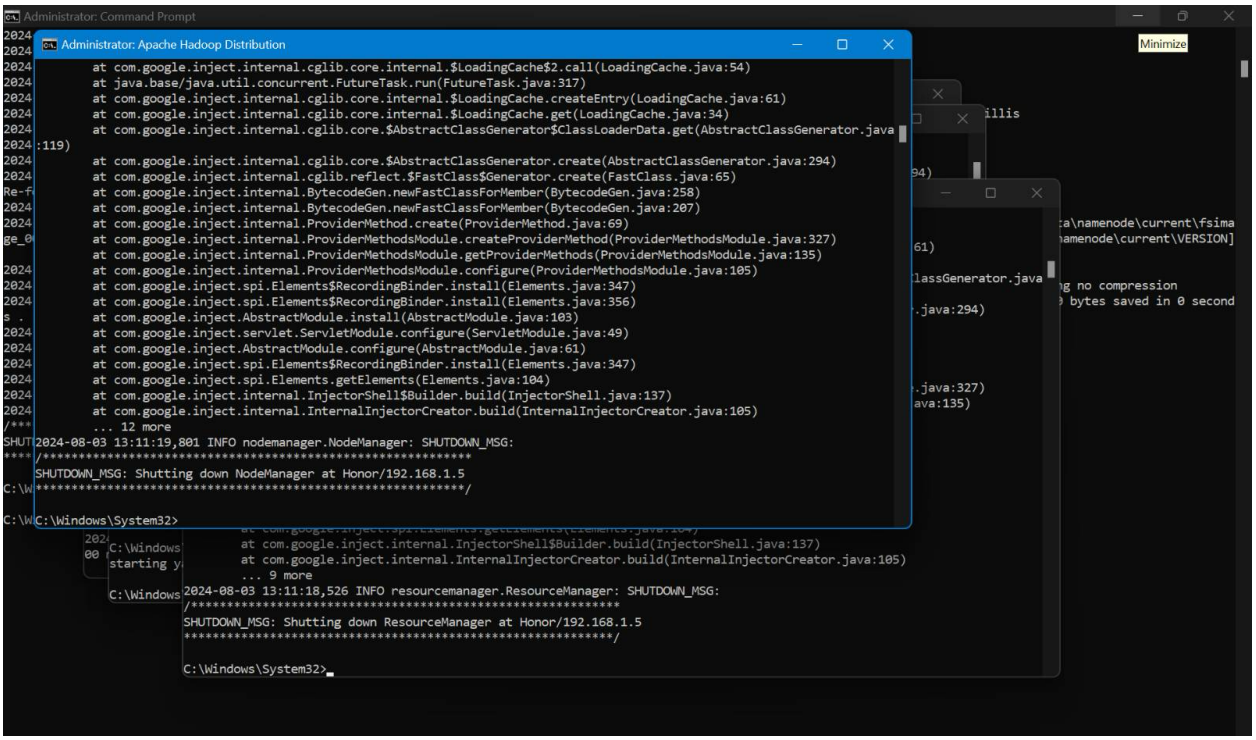
```xml
<?xml version="1.0"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->
<configuration>
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>
</configuration>
```

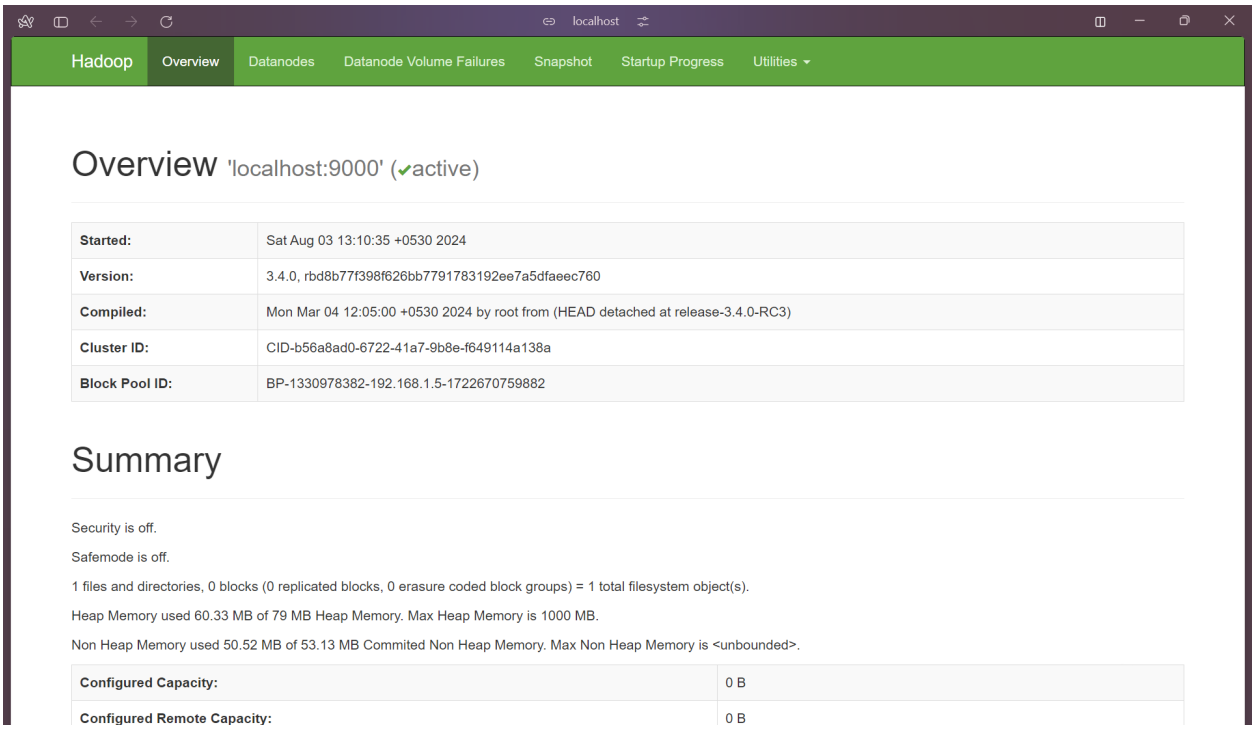



Administrator: Command Prompt
2024-08-03 13:09:16,167 INFO util.GSet: VM type = 64-bit
2024-08-03 13:09:16,167 INFO util.GSet: 1.0% max memory 1000 MB = 10 MB
2024-08-03 13:09:16,168 INFO util.GSet: capacity = 2^20 = 1048576 entries
2024-08-03 13:09:16,172 INFO namenode.FSDirectory: ACLs enabled? true
2024-08-03 13:09:16,172 INFO namenode.FSDirectory: POSIX ACL inheritance enabled? true
2024-08-03 13:09:16,172 INFO namenode.FSDirectory: XAttrs enabled? true
2024-08-03 13:09:16,176 INFO namenode.NameNode: Caching file names occurring more than 10 times
2024-08-03 13:09:16,182 INFO snapshot.SnapshotManager: Loaded config captureOpenFiles: false, skipCaptureAccessTimeOnlyChange: false, snapshotDiffAllowSnapRootDescendant: true, maxSnapshotFSLimit: 65536, maxSnapshotLimit: 65536
2024-08-03 13:09:16,182 INFO snapshot.SnapshotManager: dfs.namenode.snapshot.deletion.ordered = false
2024-08-03 13:09:16,186 INFO snapshot.SnapshotManager: SkipList is disabled
2024-08-03 13:09:16,190 INFO util.GSet: Computing capacity for map cachedBlocks
2024-08-03 13:09:16,190 INFO util.GSet: VM type = 64-bit
2024-08-03 13:09:16,193 INFO util.GSet: 0.25% max memory 1000 MB = 2.5 MB
2024-08-03 13:09:16,194 INFO util.GSet: capacity = 2^18 = 262144 entries
2024-08-03 13:09:16,203 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
2024-08-03 13:09:16,203 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
2024-08-03 13:09:16,204 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
2024-08-03 13:09:16,211 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
2024-08-03 13:09:16,212 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry time is 600000 millis
2024-08-03 13:09:16,214 INFO util.GSet: Computing capacity for map NameNodeRetryCache
2024-08-03 13:09:16,214 INFO util.GSet: VM type = 64-bit
2024-08-03 13:09:16,215 INFO util.GSet: 0.029999999329447746% max memory 1000 MB = 307.2 KB
2024-08-03 13:09:16,216 INFO util.GSet: capacity = 2^15 = 32768 entries
Re-format filesystem in Storage Directory root= C:\hadoop\data\namenode; location= null ? (Y or N) Y
2024-08-03 13:09:19,917 INFO namenode.FSImage: Allocated new BlockPoolId: BP-1330978382-192.168.1.5-1722670759882
2024-08-03 13:09:19,918 INFO common.Storage: Will remove files: [C:\hadoop\data\namenode\current\edits_inprogress_0000000000000000001, C:\hadoop\data\namenode\current\fsimage_0000000000000000000, C:\hadoop\data\namenode\current\fsimage_0000000000000000000.md5, C:\hadoop\data\namenode\current\seen_txid, C:\hadoop\data\namenode\current\VERSION]
2024-08-03 13:09:19,975 INFO common.Storage: Storage directory C:\hadoop\data\namenode has been successfully formatted.
2024-08-03 13:09:20,004 INFO namenode.FSImageFormatProtobuf: Saving image file C:\hadoop\data\namenode\current\fsimage.ckpt_0000000000000000000 using no compression
2024-08-03 13:09:20,106 INFO namenode.FSImageFormatProtobuf: Image file C:\hadoop\data\namenode\current\fsimage.ckpt_0000000000000000000 of size 400 bytes saved in 0 seconds .
2024-08-03 13:09:20,117 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2024-08-03 13:09:20,123 INFO blockmanagement.DatanodeManager: Slow peers collection thread shutdown
2024-08-03 13:09:20,142 INFO namenode.FSNamesystem: Stopping services started for active state
2024-08-03 13:09:20,143 INFO namenode.FSNamesystem: Stopping services started for standby state
2024-08-03 13:09:20,149 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2024-08-03 13:09:20,150 INFO namenode.NameNode: SHUTDOWN_MSG:
/**
SHUTDOWN_MSG: Shutting down NameNode at Honor/192.168.1.5
**/


Administrator: Command Prompt
2024-08-03 13:09:20,143 INFO namenode.FSNamesystem: Stopping services started for standby state
2024-08-03 13:09:20,149 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2024-08-03 13:09:20,150 INFO namenode.NameNode: SHUTDOWN_MSG:
/**
SHUTDOWN_MSG: Shutting down NameNode at Honor/192.168.1.5
**/
C:\Windows\System32>start-dfs.cmd


Administrator: Command Prompt
Apache Hadoop Distribution - hadoop namenode
Administrator: Apache Hadoop Distribution
at org.apache.hadoop.hdfs.server.datanode.DataStorage.loadDataStorage(DataStorage.java:414)
at org.apache.hadoop.hdfs.server.datanode.DataStorage.addStorageLocations(DataStorage.java:389)
at org.apache.hadoop.hdfs.server.datanode.DataStorage.recoverTransitionRead(DataStorage.java:566)
at org.apache.hadoop.hdfs.server.datanode.DataNode.initStorage(DataNode.java:2259)
at org.apache.hadoop.hdfs.server.datanode.DataNode.initBlockPool(DataNode.java:2195)
at org.apache.hadoop.hdfs.server.datanode.BPOfferService.verifyAndSetNamespaceInfo(BPOfferService.java:394)
at org.apache.hadoop.hdfs.server.datanode.BPServiceActor.connectToNNAndHandshake(BPServiceActor.java:315)
at org.apache.hadoop.hdfs.server.datanode.BPServiceActor.run(BPServiceActor.java:889)
at java.base/java.lang.Thread.run(Thread.java:1623)
2024-08-03 13:10:36,096 ERROR datanode.DataNode: Initialization failed for Block pool <registering> (Datanode Uuid 677d10fa-ead0-4630-ba57-3db000662b20) service to localhost/127.0.0.1:9000. Exiting.
java.io.IOException: All specified directories have failed to load.
at org.apache.hadoop.hdfs.server.datanode.DataStorage.recoverTransitionRead(DataStorage.java:567)
at org.apache.hadoop.hdfs.server.datanode.DataNode.initStorage(DataNode.java:2259)
at org.apache.hadoop.hdfs.server.datanode.DataNode.initBlockPool(DataNode.java:2195)
at org.apache.hadoop.hdfs.server.datanode.BPOfferService.verifyAndSetNamespaceInfo(BPOfferService.java:394)
at org.apache.hadoop.hdfs.server.datanode.BPServiceActor.connectToNNAndHandshake(BPServiceActor.java:315)
at org.apache.hadoop.hdfs.server.datanode.BPServiceActor.run(BPServiceActor.java:889)
at java.base/java.lang.Thread.run(Thread.java:1623)
2024-08-03 13:10:36,097 WARN datanode.DataNode: Ending block pool service for: Block pool <registering> (Datanode Uuid 677d10fa-ead0-4630-ba57-3db000662b20) service to localhost/127.0.0.1:9000
2024-08-03 13:10:36,099 INFO datanode.DataNode: Removed Block pool <registering> (Datanode Uuid 677d10fa-ead0-4630-ba57-3db000662b20)
2024-08-03 13:10:38,110 WARN datanode.DataNode: Exiting Datanode
2024-08-03 13:10:38,118 INFO datanode.DataNode: SHUTDOWN_MSG:
/**
SHUTDOWN_MSG: Shutting down DataNode at Honor/192.168.1.5
**/
C:\Windows\System32>


Administrator: Command Prompt
Administrator: Apache Hadoop Distribution
at com.google.inject.internal.cglib.core.internal.$LoadingCache$2.call(LoadingCache.java:54)
at java.base/java.util.concurrent.FutureTask.run(FutureTask.java:317)
at com.google.inject.internal.cglib.core.internal.$LoadingCache.createEntry(LoadingCache.java:61)
at com.google.inject.internal.cglib.core.internal.$LoadingCache.get(LoadingCache.java:34)
at com.google.inject.internal.cglib.core.$AbstractClassGenerator$ClassLoaderData.get(AbstractClassGenerator.java:119)
at com.google.inject.internal.cglib.core.$AbstractClassGenerator.create(AbstractClassGenerator.java:294)
at com.google.inject.internal.cglib.reflect.$FastClass$Generator.create(FastClass.java:65)
at com.google.inject.internal.BytecodeGen.newFastClassForMember(BytecodeGen.java:258)
at com.google.inject.internal.BytecodeGen.newFastClassForMember(BytecodeGen.java:207)
at com.google.inject.internal.ProviderMethod.create(ProviderMethod.java:69)
at com.google.inject.internal.ProviderMethodsModule.createProviderMethod(ProviderMethodsModule.java:327)
at com.google.inject.internal.ProviderMethodsModule.getProviderMethods(ProviderMethodsModule.java:135)
at com.google.inject.internal.ProviderMethodsModule.configure(ProviderMethodsModule.java:105)
at com.google.inject.spi.Elements$RecordingBinder.install(Elements.java:347)
at com.google.inject.spi.Elements$RecordingBinder.install(Elements.java:356)
at com.google.inject.AbstractModule.install(AbstractModule.java:103)
at com.google.inject.servlet.ServletModule.configure(ServletModule.java:49)
at com.google.inject.AbstractModule.configure(AbstractModule.java:61)
at com.google.inject.spi.Elements$RecordingBinder.install(Elements.java:347)
at com.google.inject.spi.Elements.getElements(Elements.java:104)
at com.google.inject.internal.InjectorShell$Builder.build(InjectorShell.java:137)
at com.google.inject.internal.InternalInjectorCreator.build(InternalInjectorCreator.java:105)
... 12 more
2024-08-03 13:11:19,801 INFO nodemanager.NodeManager: SHUTDOWN_MSG:
SHUTDOWN_MSG: Shutting down NodeManager at Honor/192.168.1.5
C:\Windows\System32>
... 9 more
2024-08-03 13:11:18,526 INFO resourcemanager.ResourceManager: SHUTDOWN_MSG:
SHUTDOWN_MSG: Shutting down ResourceManager at Honor/192.168.1.5
C:\Windows\System32>
Minimize


localhost
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Overview 'localhost:9000' (active)
Started: Sat Aug 03 13:10:35 +0530 2024
Version: 3.4.0, rbd8b77f398f626bb7791783192ee7a5dfaeec760
Compiled: Mon Mar 04 12:05:00 +0530 2024 by root from (HEAD detached at release-3.4.0-RC3)
Cluster ID: CID-b56a8ad0-6722-41a7-9b8e-f649114a138a
Block Pool ID: BP-1330978382-192.168.1.5-1722670759882
Summary
Security is off.
Safemode is off.
1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).
Heap Memory used 60.33 MB of 79 MB Heap Memory. Max Heap Memory is 1000 MB.
Non Heap Memory used 50.52 MB of 53.13 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.
Configured Capacity: 0 B
Configured Remote Capacity: 0 B