

Installing Hadoop 3.2.1 Single node cluster on Windows 10

Prerequisites:

https://www.java.com/en/download/windows_offline.jsp

<https://www.oracle.com/java/technologies/downloads/#java8>

<https://www.7-zip.org/download.html>

<https://www.apache.org/dyn/closer.cgi/hadoop/common/hadoop-3.2.1/hadoop-3.2.1.tar.gz>

<https://github.com/cdarlint/winutils>

<https://github.com/cdarlint/winutils/tree/master/hadoop-3.2.1/bin>

<https://github.com/FahaoTang/big-data/blob/master/hadoop-hdfs-3.2.1.jar>

While working on a [project](#) two years ago, I wrote a step-by-step guide to [install Hadoop 3.1.0 on Ubuntu 16.04](#) operating system. Since we are currently working on a new project where we need to install a Hadoop cluster on Windows 10, I decided to write a guide for this process.

This article is a part of a series that we are publishing on TowardsDataScience.com that aims to illustrate how to install Big Data technologies on Windows operating system.

Other published articles in this series:

- [Installing Apache Pig 0.17.0 on Windows 10](#)
- [Installing Apache Hive 3.1.2 on Windows 10](#)

1. Prerequisites

First, we need to make sure that the following prerequisites are installed:

1. Java 8 runtime environment

(JRE): [Hadoop 3 requires a Java 8 installation](#). I prefer using the [offline installer](#).

2. [Java 8 development Kit \(JDK\)](#)

3. To unzip downloaded Hadoop binaries, we should install 7zip.

4. I will create a folder "E:\hadoop-env" on my local machine to store downloaded files.

2. Download Hadoop binaries

The first step is to download Hadoop binaries from the official website. The binary package size is about 342 MB.

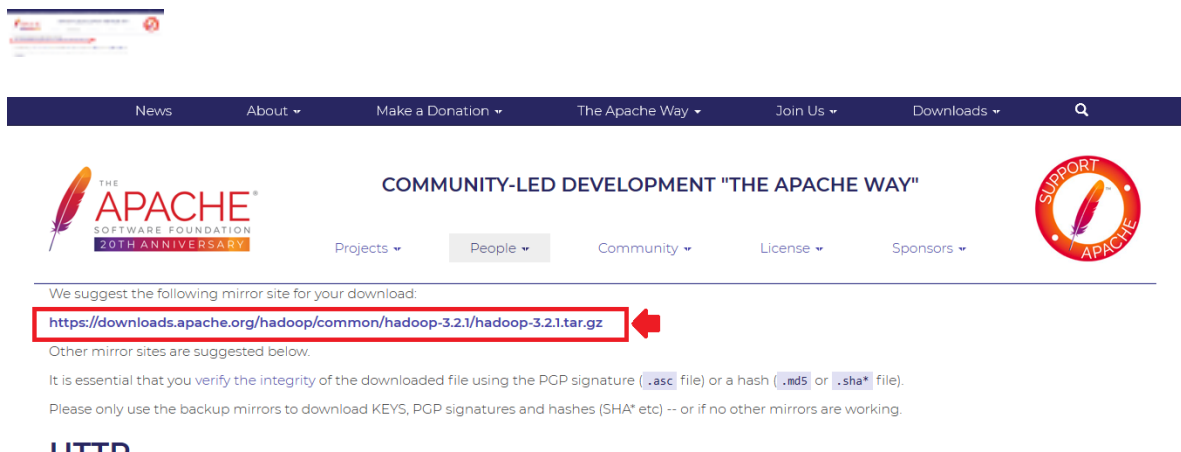


Figure 1 — Hadoop binaries download link
After finishing the file download, we should unpack the package using 7zip in two steps. First, we should extract the hadoop-3.2.1.tar.gz library, and then, we should unpack the extracted tar file:



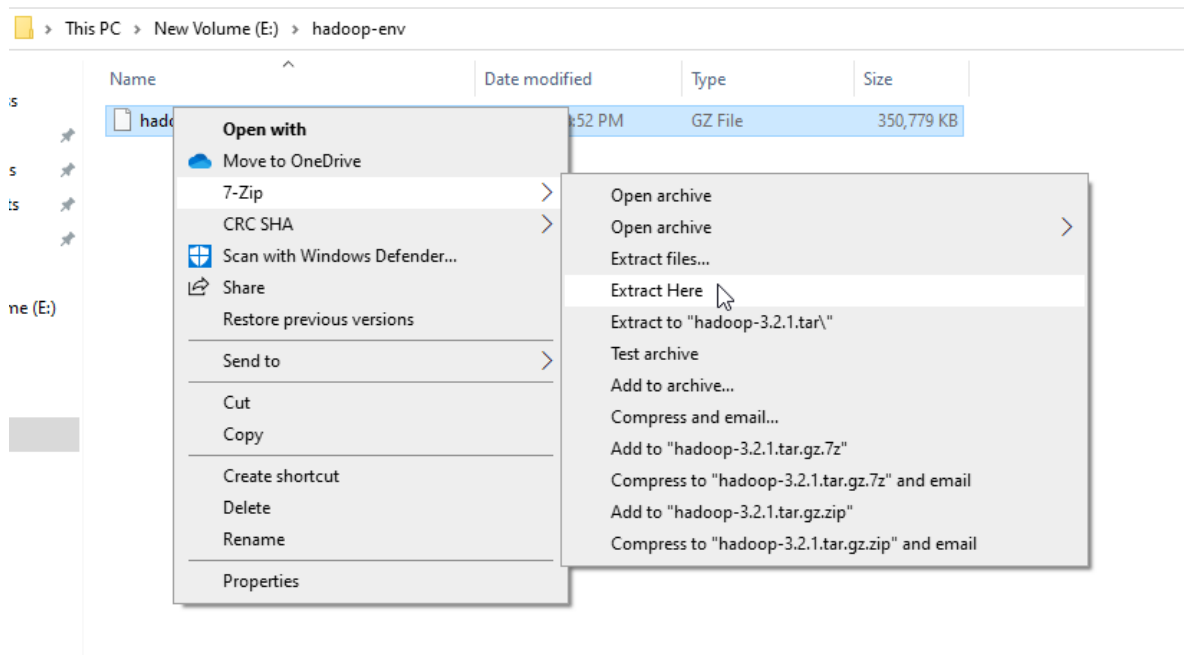


Figure 2 — Extracting hadoop-3.2.1.tar.gz package using 7zip

Name	Date modified	Type	Size
hadoop-3.2.1.tar	9/10/2019 8:11 PM	TAR File	893,250 KB
hadoop-3.2.1.tar.gz	4/15/2020 8:52 PM	GZ File	350,779 KB

Figure 3 — Extracted hadoop-3.2.1.tar file



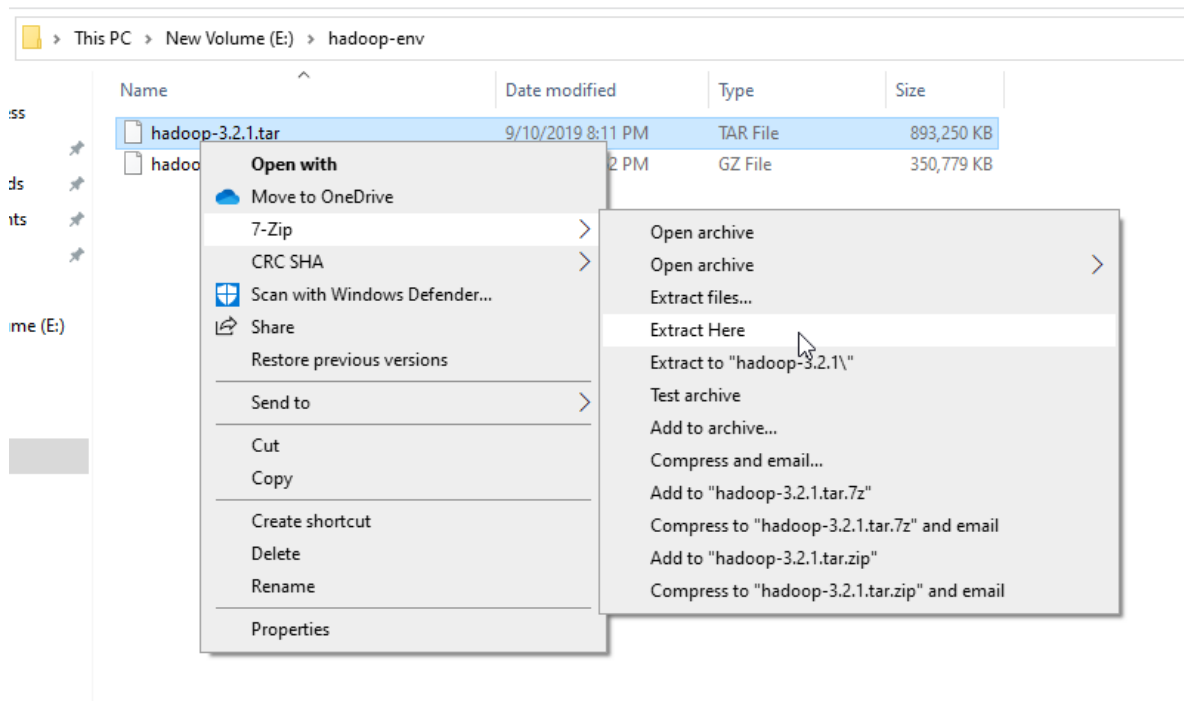


Figure 4 — Extracting the hadoop-3.2.1.tar file

The tar file extraction may take some minutes to finish. In the end, you may see some warnings about symbolic link creation. Just ignore these warnings since they are not related to windows.

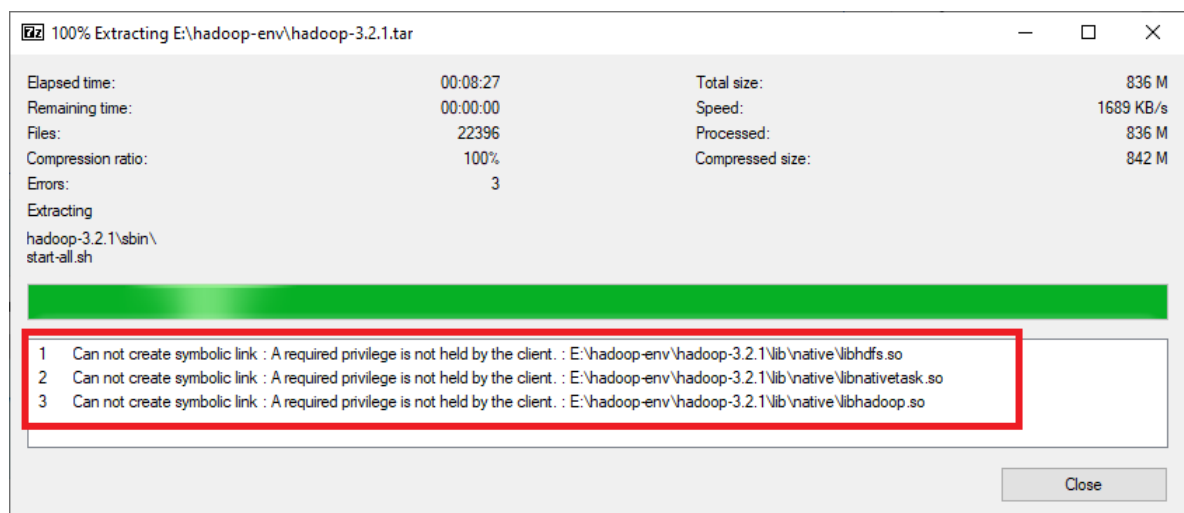
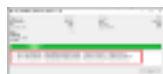


Figure 5 — Symbolic link warnings

After unpacking the package, we should add the Hadoop native IO libraries, which can be found in the following GitHub repository: <https://github.com/cdarlint/winutils>.

Since we are installing Hadoop 3.2.1, we should download the files located in <https://github.com/cdarlint/winutils/tree/master/hadoop-3.2.1/bin> and copy them into the "hadoop-3.2.1\bin" directory.

3. 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.



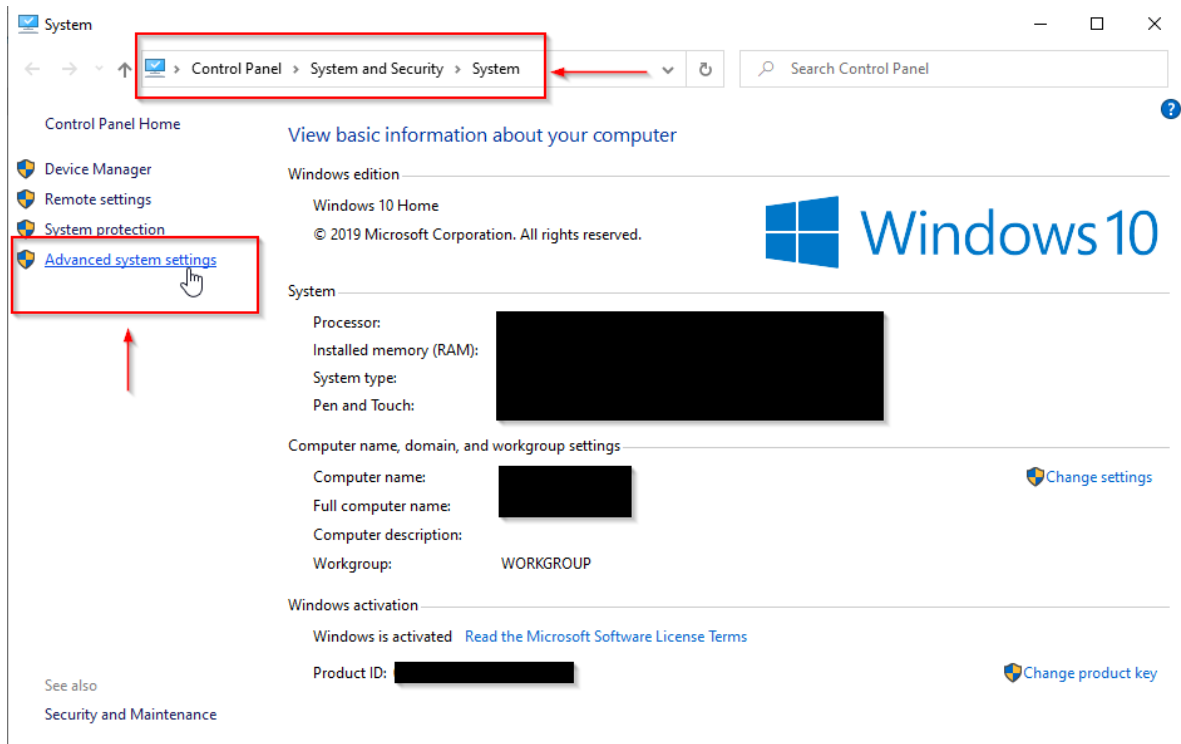


Figure 6 — Opening advanced system settings

When the "Advanced system settings" dialog appears, go to the "Advanced" tab and click on the "Environment variables" button located on the bottom of the dialog.



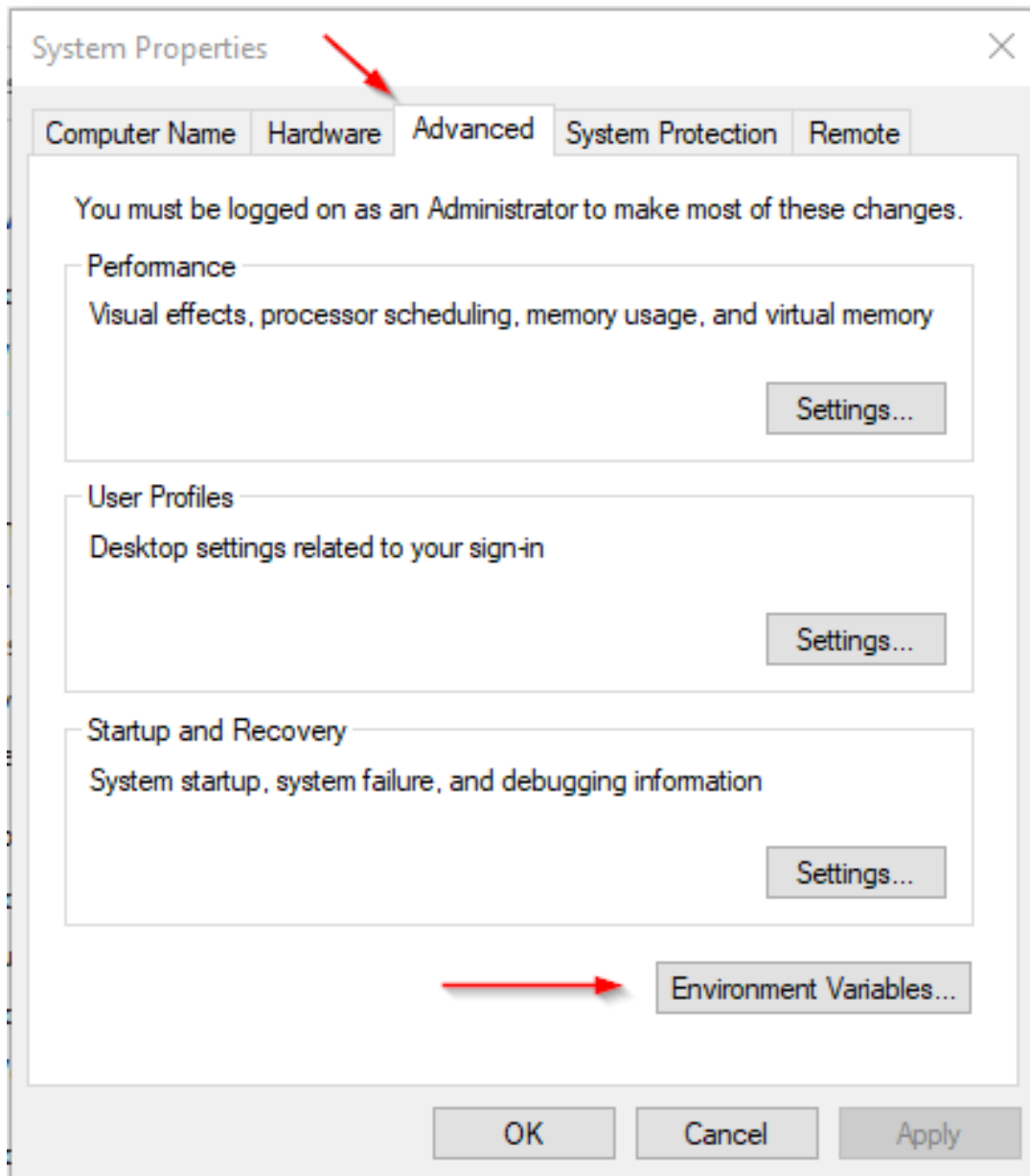


Figure 7 — Advanced system settings dialog

In the "Environment Variables" dialog, press the "New" button to add a new variable.

Note: In this guide, we will add user variables since we are configuring Hadoop for a single user. If you are looking to configure Hadoop for multiple users, you can define System variables instead.

There are two variables to define:

1. JAVA_HOME: JDK installation folder path
2. HADOOP_HOME: Hadoop installation folder path

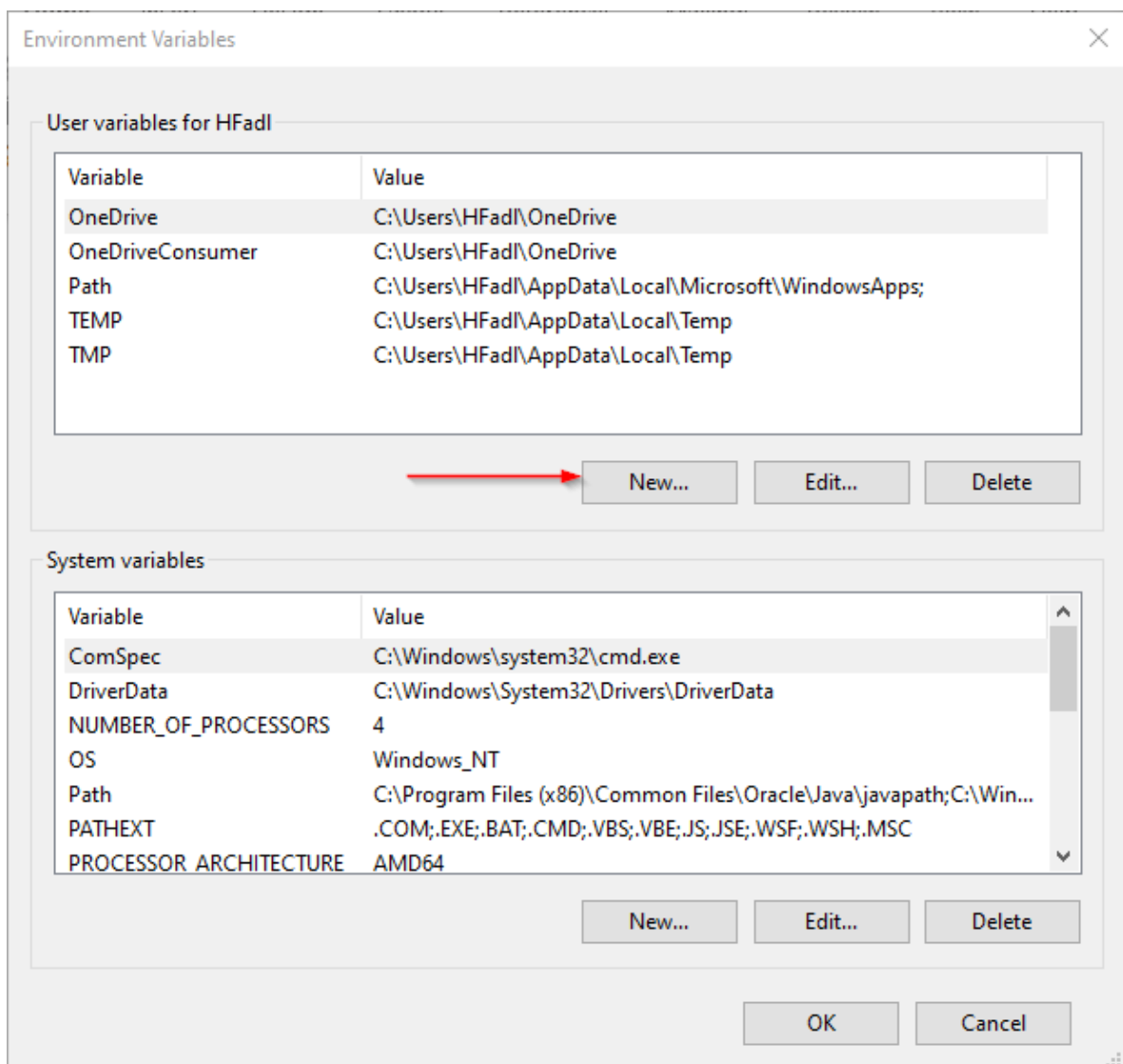


Figure 8 — Adding JAVA_HOME variable



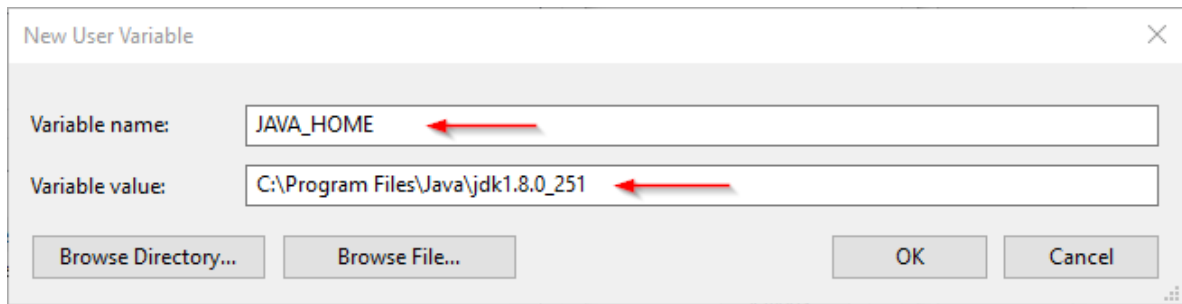


Figure 9 — Adding HADOOP_HOME variable

Now, we should edit the PATH variable to add the Java and Hadoop binaries paths as shown in the following screenshots.

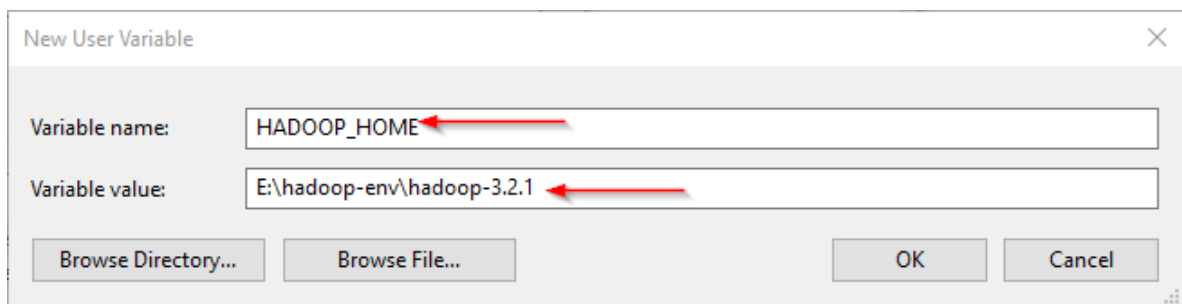
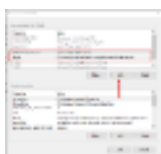


Figure 10 — Editing the PATH variable



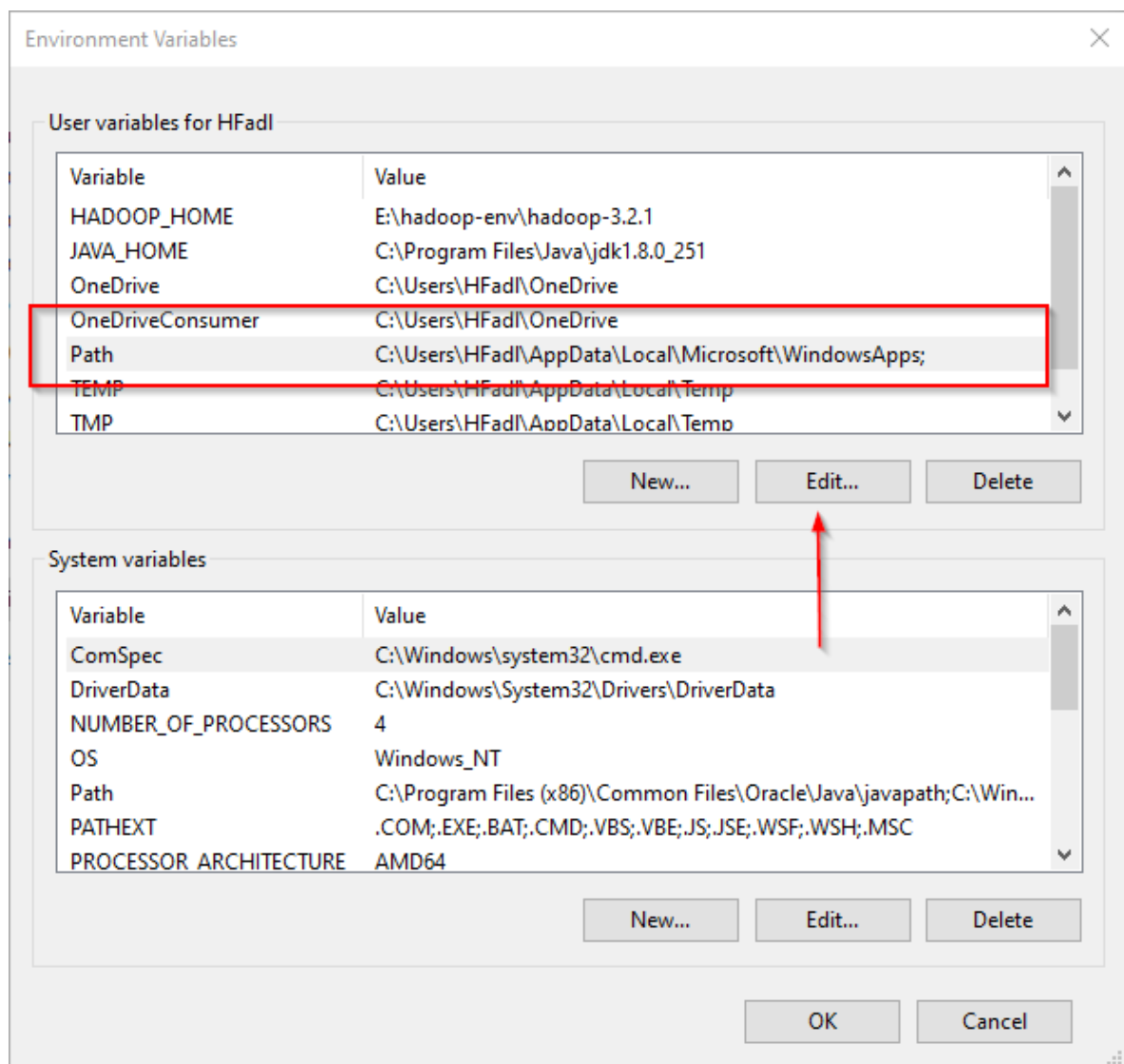
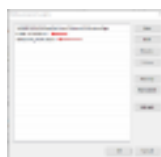


Figure 11 — Editing PATH variable



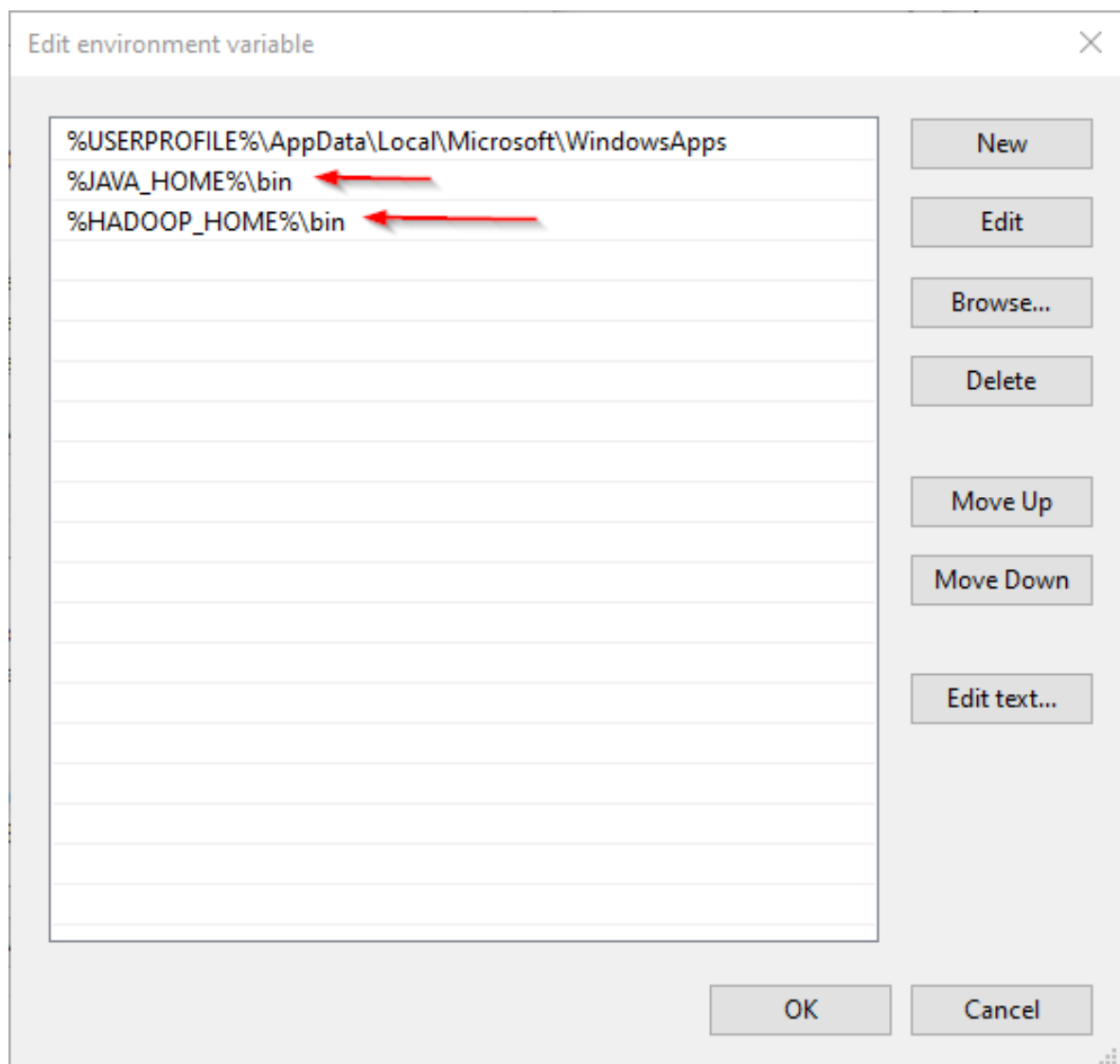


Figure 12— Adding new paths to the PATH variable

3.1. JAVA_HOME is incorrectly set error

Now, let's open PowerShell and try to run the following command:

```
hadoop -version
```

In this example, since the JAVA_HOME path contains spaces, I received the following error:

JAVA_HOME is incorrectly set



```
Windows PowerShell
PS C:\Users\HFad1> hadoop -version
The system cannot find the path specified.
Error: JAVA_HOME is incorrectly set.
Please update E:\hadoop-env\hadoop-3.2.1\etc\hadoop\hadoop-env.cmd
'-Xmx512m' is not recognized as an internal or external command,
operable program or batch file.
PS C:\Users\HFad1> 
```

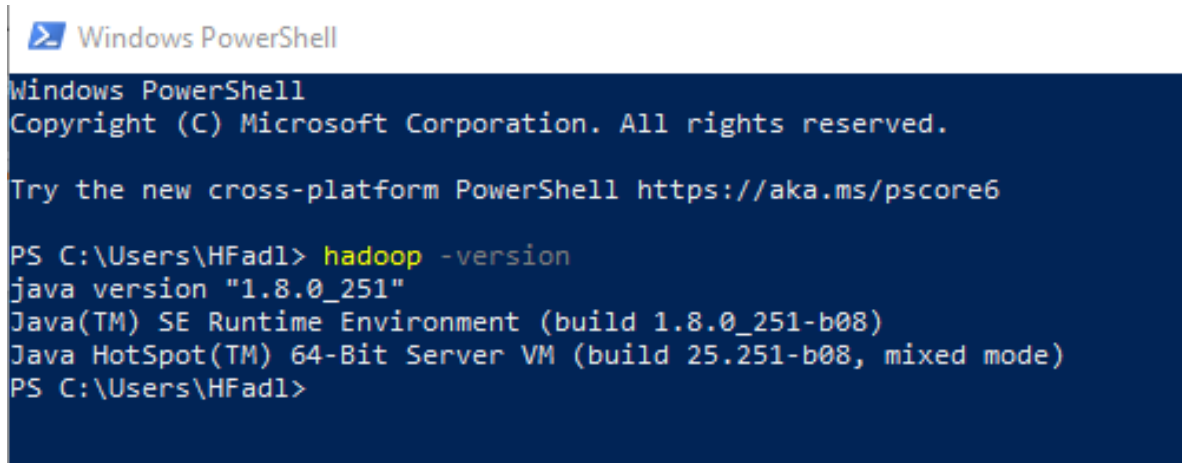
Figure 13 — JAVA_HOME error

To solve this issue, we should use the windows 8.3 path instead. As an example:

- Use "Progra~1" instead of "Program Files"
- Use "Progra~2" instead of "Program Files(x86)"

After replacing "Program Files" with "Progra~1", we closed and reopened PowerShell and tried the same command. As shown in the screenshot below, it runs without errors.



A screenshot of a Windows PowerShell terminal window. The title bar at the top says 'Windows PowerShell'. The terminal text includes the copyright notice for Microsoft Corporation, a link to the new cross-platform PowerShell, and the output of the 'hadoop -version' command. The command is entered at the prompt 'PS C:\Users\HFadl>'. The output shows the Java version '1.8.0_251', the Java(TM) SE Runtime Environment (build 1.8.0_251-b08), and the Java HotSpot(TM) 64-Bit Server VM (build 25.251-b08, mixed mode).

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

Try the new cross-platform PowerShell https://aka.ms/pscore6

PS C:\Users\HFadl> hadoop -version
java version "1.8.0_251"
Java(TM) SE Runtime Environment (build 1.8.0_251-b08)
Java HotSpot(TM) 64-Bit Server VM (build 25.251-b08, mixed mode)
PS C:\Users\HFadl>
```

Figure 14 — hadoop -version command executed successfully

4. Configuring Hadoop cluster

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

- %HADOOP_HOME%\etc\hadoop\hdfs-site.xml
- %HADOOP_HOME%\etc\hadoop\core-site.xml
- %HADOOP_HOME%\etc\hadoop\mapred-site.xml
- %HADOOP_HOME%\etc\hadoop\yarn-site.xml

4.1. HDFS site configuration

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:

- E:\hadoop-env\hadoop-3.2.1\data\dfs\namenode
- E:\hadoop-env\hadoop-3.2.1\data\dfs\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:///E:/hadoop-env/
```

```
hadoop-3.2.1/data/dfs/namenode</value>
```

```
</property>
```

```
<property>
```

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

```
<value>file:///E:/hadoop-env/
```

```
hadoop-3.2.1/data/dfs/datanode</value>
```

```
</property>
```

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

4.2. Core site configuration

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>
```

4.3. Map Reduce site configuration

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>
```

4.4. Yarn site configuration

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>
```

5. Formatting Name node

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

```
hdfs namenode -format
```

Due to a [bug in the Hadoop 3.2.1 release](#), you will receive the following error:

```
2020-04-17 22:04:01,503 ERROR
```

```
namenode.NameNode: Failed to start namenode.
```

```
java.lang.UnsupportedOperationException
at
```

```
java.nio.file.Files.setPosixFilePermissions(Files.java:2044)
```

at

org.apache.hadoop.hdfs.server.common.Storage\$StorageDirectory.clearDirectory(Storage.java:452)

at

org.apache.hadoop.hdfs.server.namenode.NNStorage.format(NNStorage.java:591)

at

org.apache.hadoop.hdfs.server.namenode.NNStorage.format(NNStorage.java:613)

at

org.apache.hadoop.hdfs.server.namenode.FSImage.format(FSImage.java:188)

at

org.apache.hadoop.hdfs.server.namenode.NameNode.format(NameNode.java:1206)

at

org.apache.hadoop.hdfs.server.namenode.NameNode.createNameNode(NameNode.java:1649)

at

org.apache.hadoop.hdfs.server.namenode.NameNode.main(NameNode.java:1759)

2020-04-17 22:04:01,511 INFO

util.ExitUtil: Exiting with status 1:

java.lang.UnsupportedOperationException
2020-04-17 22:04:01,518 INFO

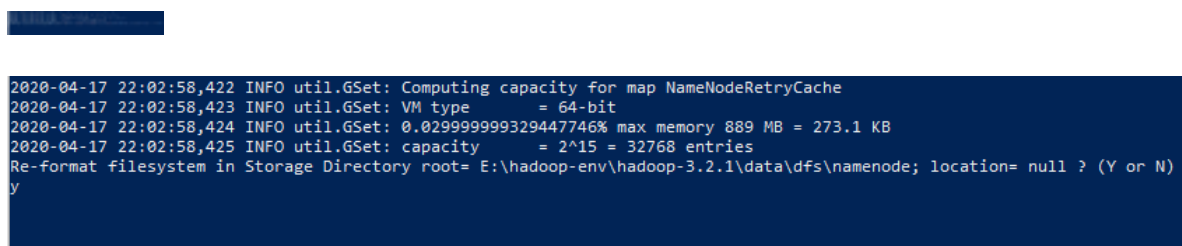
namenode.NameNode:

SHUTDOWN_MSG:

This issue will be solved within the next release. For now, you can fix it temporarily using the following steps ([reference](#)):

- Download hadoop-hdfs-3.2.1.jar file from the [following 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

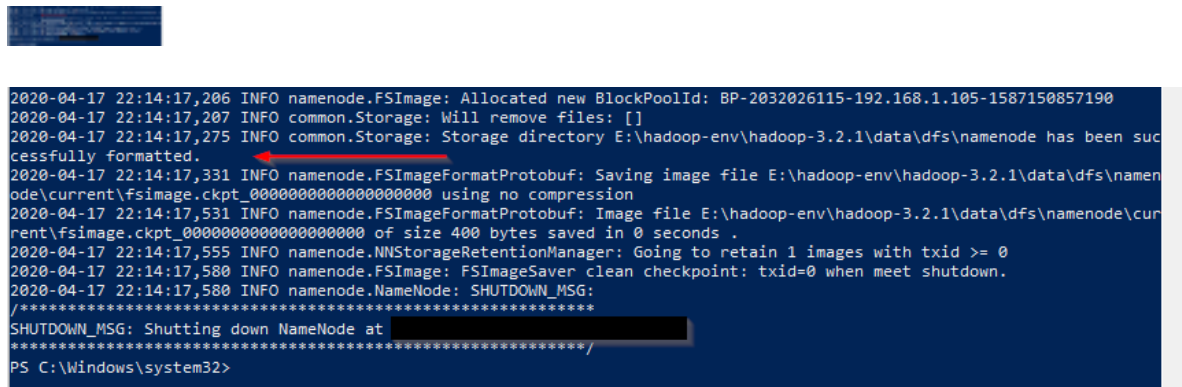
Now, if we try to re-execute the format command (Run the command prompt or PowerShell as administrator), you need to approve file system format.



```
2020-04-17 22:02:58,422 INFO util.GSet: Computing capacity for map NameNodeRetryCache
2020-04-17 22:02:58,423 INFO util.GSet: VM type = 64-bit
2020-04-17 22:02:58,424 INFO util.GSet: 0.029999999329447746% max memory 889 MB = 273.1 KB
2020-04-17 22:02:58,425 INFO util.GSet: capacity = 2^15 = 32768 entries
Re-format filesystem in Storage Directory root= E:\hadoop-env\hadoop-3.2.1\data\dfs\namenode; location= null ? (Y or N)
y
```

Figure 15 — File system format approval
And the command is executed

successfully:



```
2020-04-17 22:14:17,206 INFO namenode.FSImage: Allocated new BlockPoolId: BP-2032026115-192.168.1.105-1587150857190
2020-04-17 22:14:17,207 INFO common.Storage: Will remove files: []
2020-04-17 22:14:17,275 INFO common.Storage: Storage directory E:\hadoop-env\hadoop-3.2.1\data\dfs\namenode has been successfully formatted.
2020-04-17 22:14:17,331 INFO namenode.FSImageFormatProtobuf: Saving image file E:\hadoop-env\hadoop-3.2.1\data\dfs\namenode\current\fsimage.ckpt_00000000000000000000 using no compression
2020-04-17 22:14:17,531 INFO namenode.FSImageFormatProtobuf: Image file E:\hadoop-env\hadoop-3.2.1\data\dfs\namenode\current\fsimage.ckpt_00000000000000000000 of size 400 bytes saved in 0 seconds .
2020-04-17 22:14:17,555 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2020-04-17 22:14:17,580 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2020-04-17 22:14:17,580 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at [REDACTED]
*****/
PS C:\Windows\system32>
```

Figure 16 — Command executed successfully

6. Starting Hadoop services

Now, we will open PowerShell, and navigate to "%HADOOP_HOME%\sbin" directory.

Then we will run the following command to start the Hadoop nodes:

`.\start-dfs.cmd`

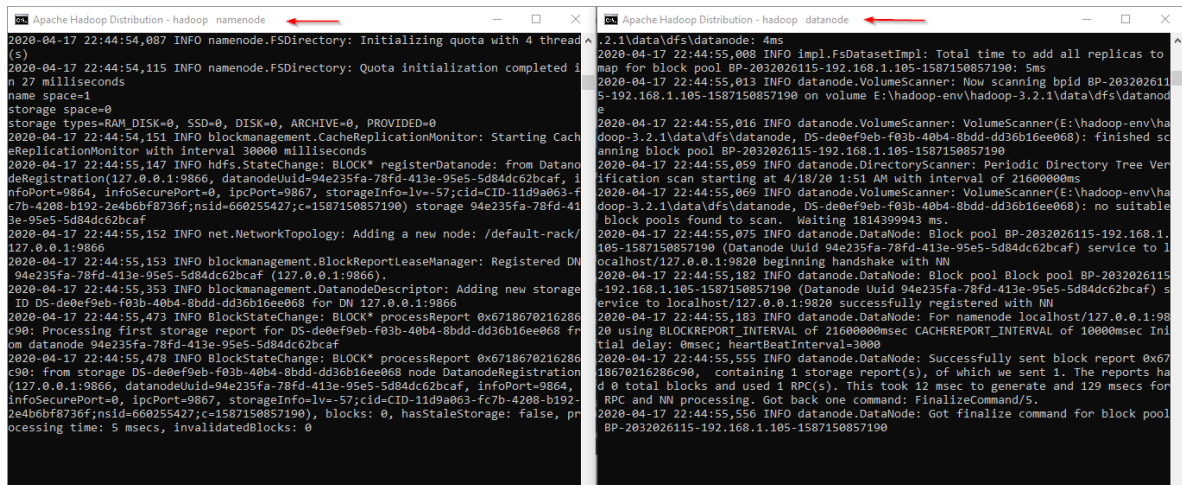


```
Administrator: Windows PowerShell
PS E:\hadoop-env\hadoop-3.2.1\sbin> .\start-dfs.cmd
PS E:\hadoop-env\hadoop-3.2.1\sbin>
```

Figure 17 — Starting Hadoop nodes

Two command prompt windows will open (one for the name node and one for the data node) as follows:

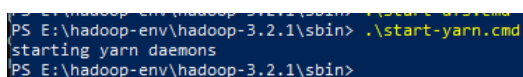




The image shows two terminal windows side-by-side. The left window is titled 'Apache Hadoop Distribution - hadoop namenode' and displays logs for the namenode startup, including quota initialization and block management. The right window is titled 'Apache Hadoop Distribution - hadoop datanode' and displays logs for the datanode startup, including volume scanning and block pool registration.

Figure 18 — Hadoop nodes command prompt windows

Next, we must start the Hadoop Yarn service using the following command:
`./start-yarn.cmd`



A terminal window showing the command prompt. The user enters `PS E:\hadoop-env\hadoop-3.2.1\sbin> .\start-yarn.cmd`, and the output is `starting yarn daemons`. The prompt then changes to `PS E:\hadoop-env\hadoop-3.2.1\sbin>`.

Figure 19 — Starting Hadoop Yarn services
Two command prompt windows will open (one for the resource manager and one for the node manager) as follows:



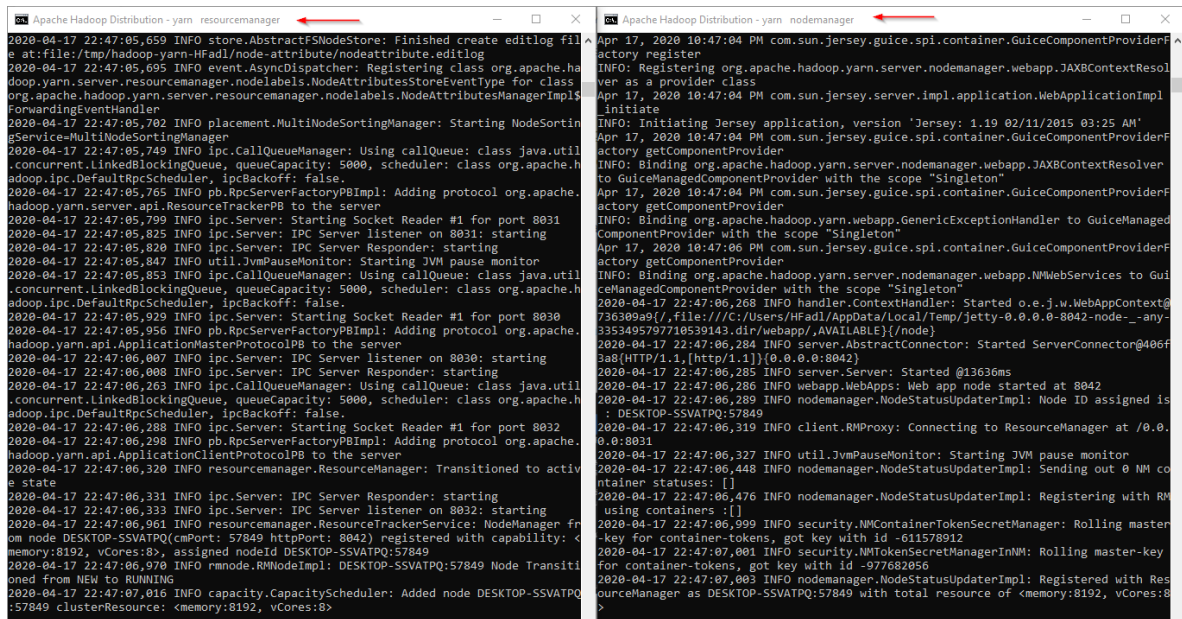


Figure 20— Node manager and Resource manager command prompt windows

To make sure that all services started successfully, we can run the following command:

```
jps
```

It should display the following services:

14560 DataNode

4960 ResourceManager

5936 NameNode

768 NodeManager

14636 Jps

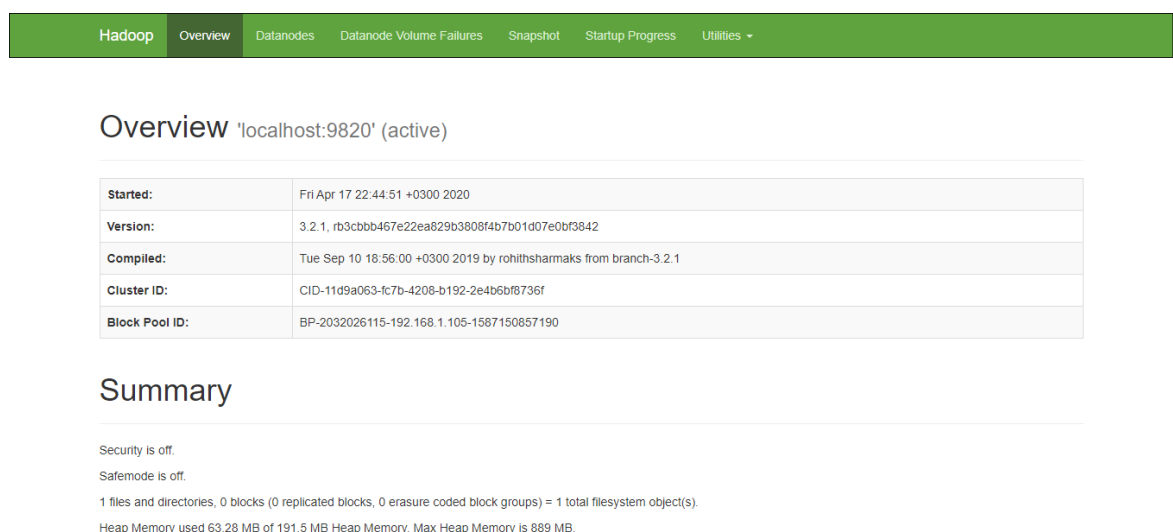
```
PS E:\hadoop-env\hadoop-3.2.1\sbin> jps
14560 DataNode
4960 ResourceManager
5936 NameNode
768 NodeManager
14636 Jps
PS E:\hadoop-env\hadoop-3.2.1\sbin>
```

Figure 21 — Executing jps command

7. Hadoop Web UI

There are three web user interfaces to be used:

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



The screenshot shows the Hadoop web interface with the 'Overview' tab selected. The title is 'Overview 'localhost:9820' (active)'. Below the title is a table with the following information:

Started:	Fri Apr 17 22:44:51 +0300 2020
Version:	3.2.1, rb3cbbb467e22ea829b3808f4b7b01d07e0bf3842
Compiled:	Tue Sep 10 18:56:00 +0300 2019 by rohithsharmaks from branch-3.2.1
Cluster ID:	CID-11d9a063-fc7b-4208-b192-2e4b6bfb736f
Block Pool ID:	BP-2032026115-192.168.1.105-1587150857190

Below the table is a 'Summary' section with the following text:

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 63.28 MB of 191.5 MB Heap Memory. Max Heap Memory is 889 MB.

Figure 22 — Name node web page

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



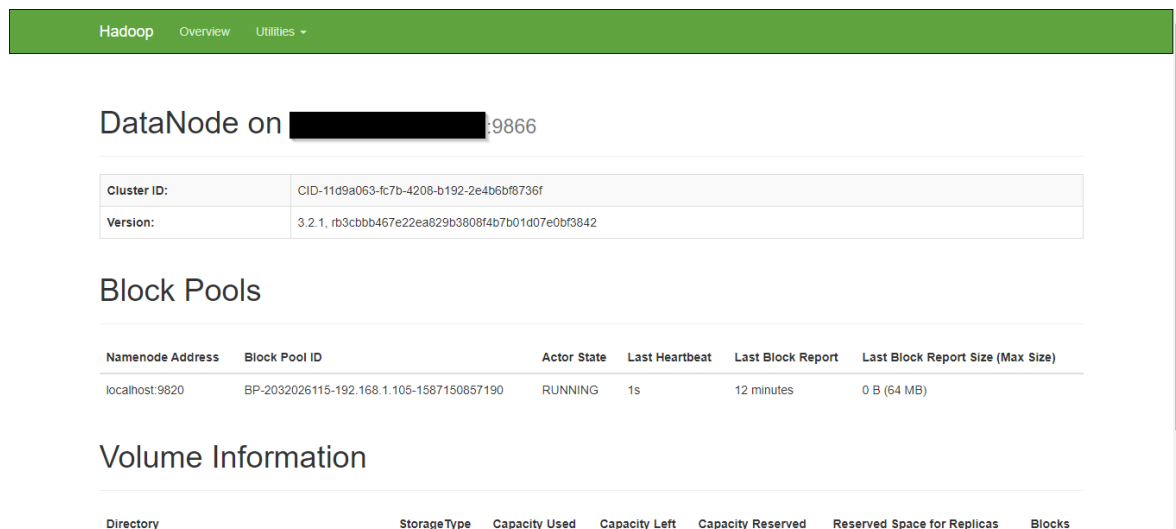


Figure 23 — Data node web page

- Yarn web page: <http://localhost:8088/cluster>

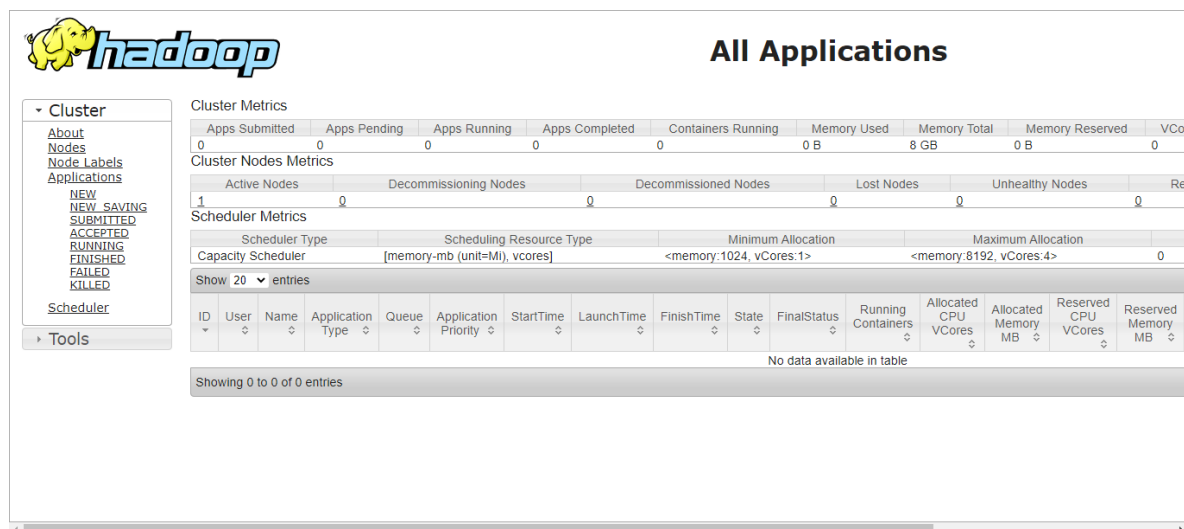


Figure 24 — Yarn web page