Assignment 1 : Hadoop Installation

Aim – Installation and configuration of Hadoop in the windows operating system.

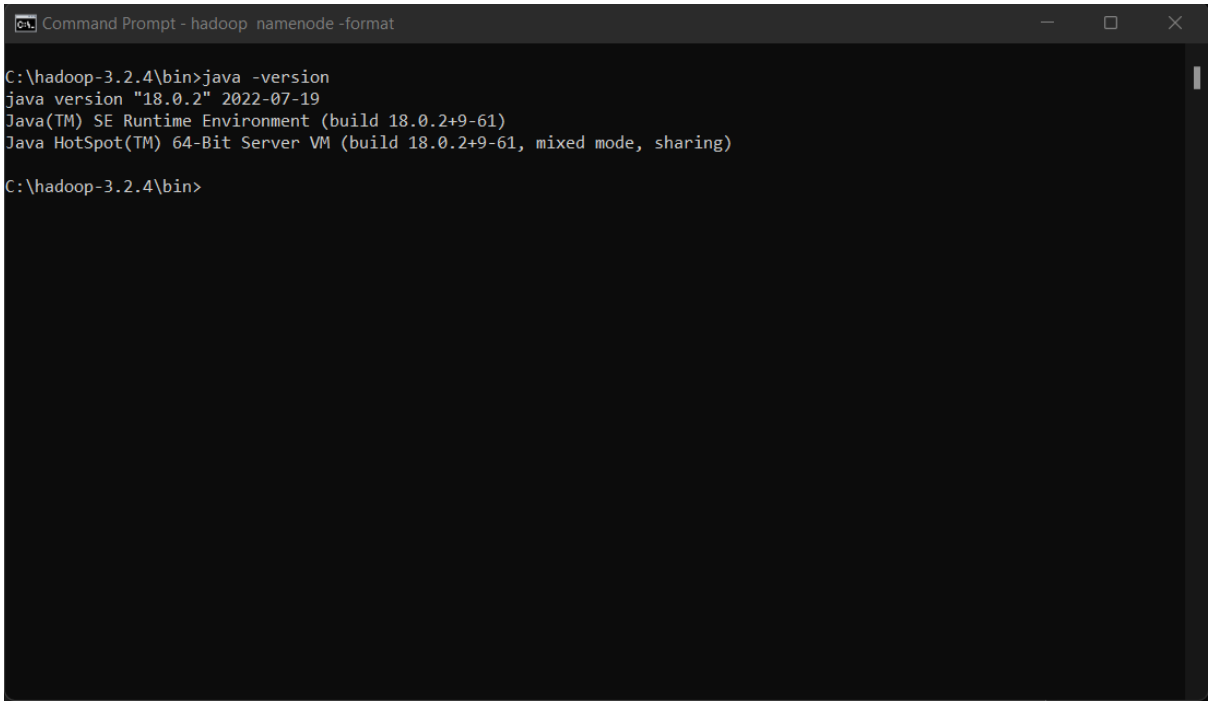
Step 1: Installation of Java

1. Hadoop required java to run it's services hence java latest version is downloaded from official oracle website https://www.oracle.com/java/technologies/downloads/.

Linux macOS Windows

Product/file description	File size	Download
x64 Compressed Archive	172.93 MB	https://download.oracle.com/java/18/latest/jdk-18_windows-x64_bin.zip (sha256)
x64 Installer	153.45 MB	https://download.oracle.com/java/18/latest/jdk-18_windows-x64_bin.exe (sha256)
x64 MSI Installer	152.33 MB	https://download.oracle.com/java/18/latest/jdk-18_windows-x64_bin.msi (sha256)

2. After successful installation of java jdk install the downloaded x64 installer.
3. To run java into the system we have to set the environment variable into local machine so that java will directly shift too the bin location whenever the java or javac command is encountered.



4. To verify if the path is properly set or not got to command promote and enter command java -version.

Step 2: Installation of Hadoop

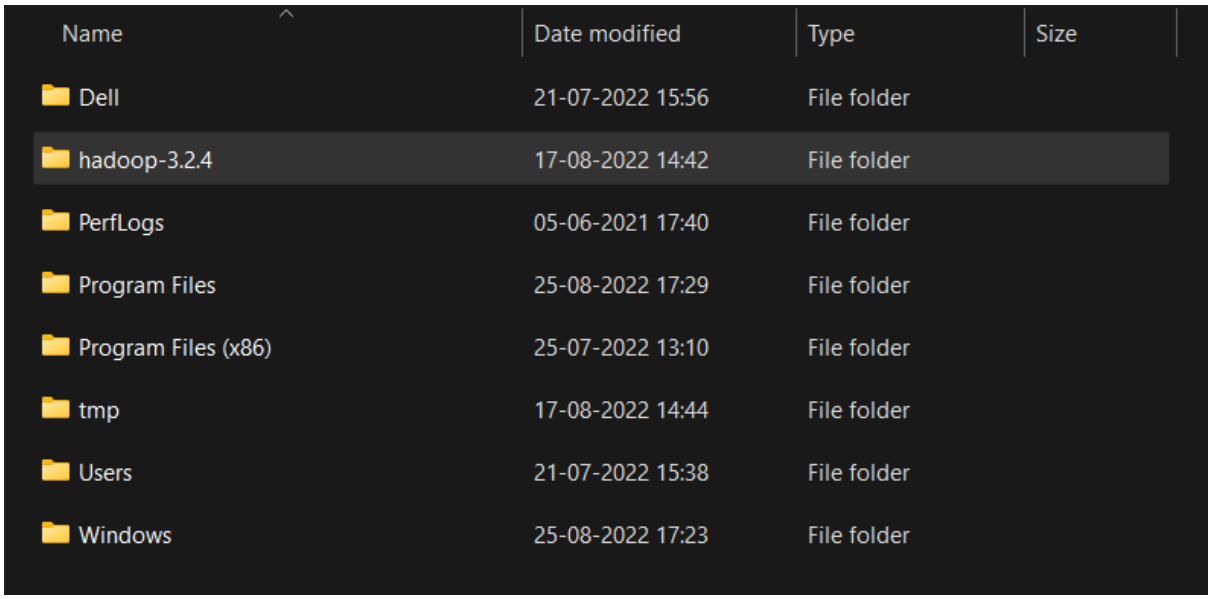
1. To install Hadoop we have to visit the official Hadoop website in order to downloaded the required files. https://hadoop.apache.org/releases.html

Download

Hadoop is released as source code tarballs with corresponding binary tarballs for convenience. The downloads are distributed via mirror sites and should be checked for tampering using GPG or SHA-512.

Version	Release date	Source download	Binary download	Release notes
3.3.4	2022 Aug 8	source (checksum signature)	binary (checksum signature) binary-aarch64 (checksum signature)	Announcement
3.2.4	2022 Jul 22	source (checksum signature)	binary (checksum signature)	Announcement
2.10.2	2022 May 31	source (checksum signature)	binary (checksum signature)	Announcement

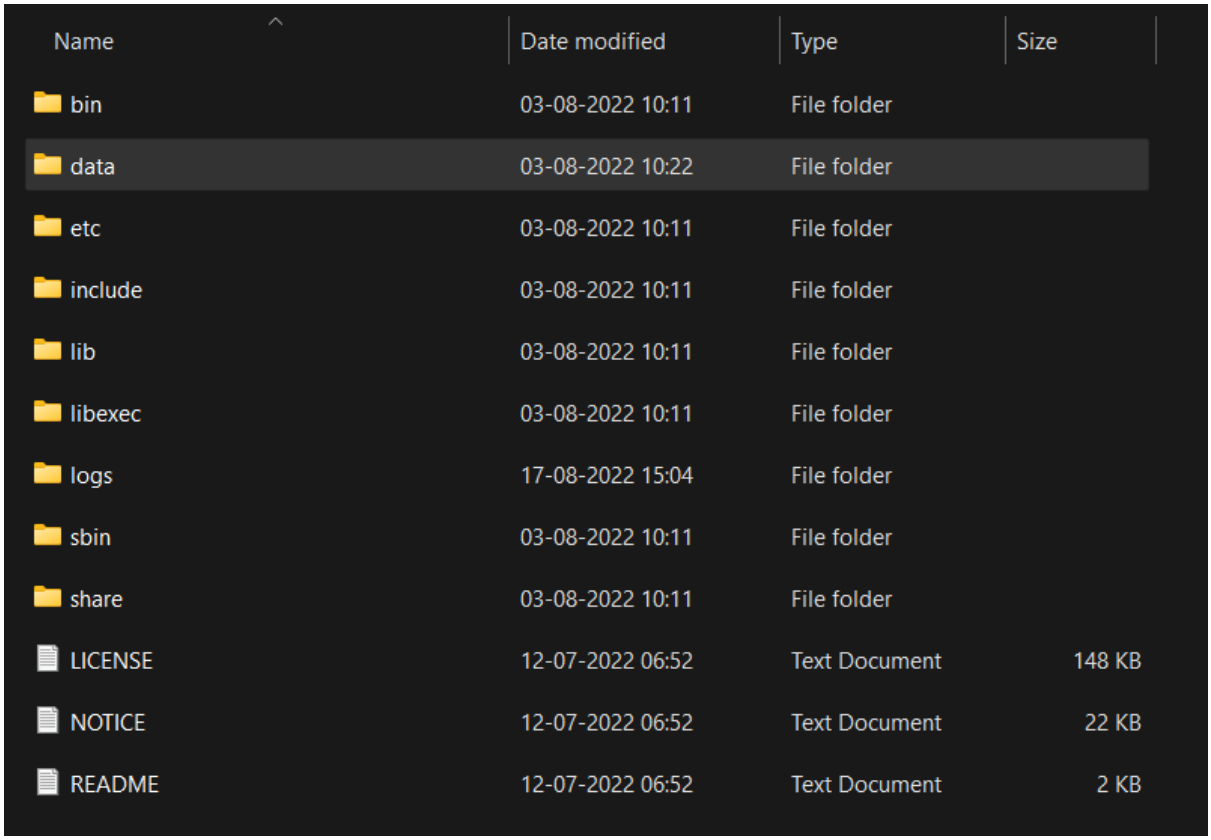
2. Download the latest version of Hadoop with binary download which allows us to modify the Hadoop files and it consist of all the required files which is in the compress format.
3. After downloading the compress .tar file extract the file into C drive (it is not exe file which will automatically extracted)
4. After successful extraction go the the Hadoop 3.3.4 folder



5. Now in order to create datanode and namenode which are important node of Hadoop architecture we have to create a new folder named as data.
6. After creation of data folder we have to create 2 new folders named as datanode and namenode which are empty for now.

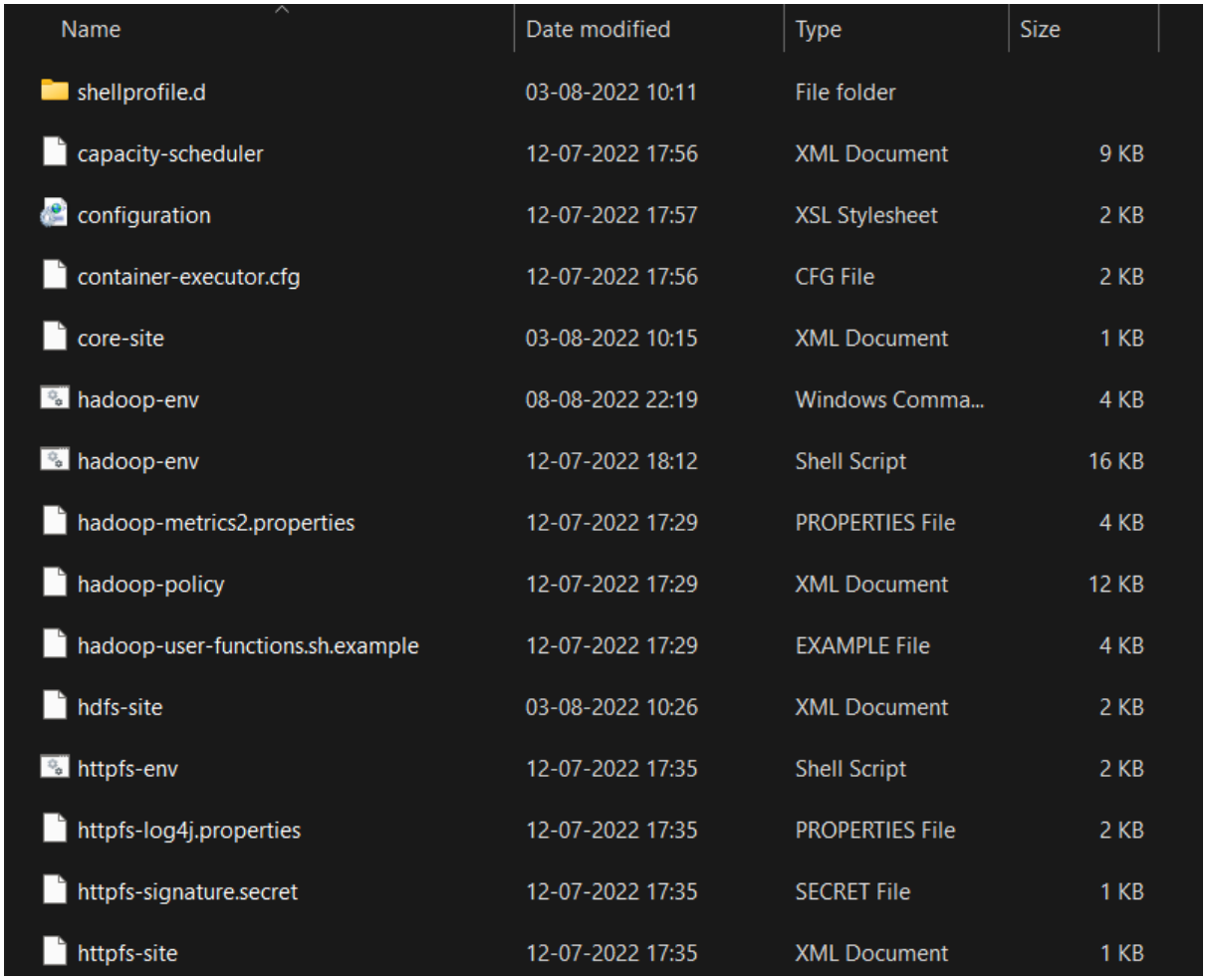
Step 3: Setting up the Hadoop files

1. To make your Hadoop run all the services we have to modify the core Hadoop xml files such as hdfs.xml, yarn.xml, and so on.
2. To modify the file go to etc folder of Hadoop installation



Name	Date modified	Type	Size
bin	03-08-2022 10:11	File folder	
data	03-08-2022 10:22	File folder	
etc	03-08-2022 10:11	File folder	
include	03-08-2022 10:11	File folder	
lib	03-08-2022 10:11	File folder	
libexec	03-08-2022 10:11	File folder	
logs	17-08-2022 15:04	File folder	
sbin	03-08-2022 10:11	File folder	
share	03-08-2022 10:11	File folder	
LICENSE	12-07-2022 06:52	Text Document	148 KB
NOTICE	12-07-2022 06:52	Text Document	22 KB
README	12-07-2022 06:52	Text Document	2 KB

3. Go to etc folder and you see all the required file that need to modify.



Name	Date modified	Type	Size
shellprofile.d	03-08-2022 10:11	File folder	
capacity-scheduler	12-07-2022 17:56	XML Document	9 KB
configuration	12-07-2022 17:57	XSL Stylesheet	2 KB
container-executor.cfg	12-07-2022 17:56	CFG File	2 KB
core-site	03-08-2022 10:15	XML Document	1 KB
hadoop-env	08-08-2022 22:19	Windows Comma...	4 KB
hadoop-env	12-07-2022 18:12	Shell Script	16 KB
hadoop-metrics2.properties	12-07-2022 17:29	PROPERTIES File	4 KB
hadoop-policy	12-07-2022 17:29	XML Document	12 KB
hadoop-user-functions.sh.example	12-07-2022 17:29	EXAMPLE File	4 KB
hdfs-site	03-08-2022 10:26	XML Document	2 KB
httpfs-env	12-07-2022 17:35	Shell Script	2 KB
httpfs-log4j.properties	12-07-2022 17:35	PROPERTIES File	2 KB
httpfs-signature.secret	12-07-2022 17:35	SECRET File	1 KB
httpfs-site	12-07-2022 17:35	XML Document	1 KB

4. Now to modify or configure the core xml to make Hadoop execute command first open the core-site.xml file and add the below code
    a. <configuration>
    b. <property>
    c. <name>fs.defaulterFS</name>
    d. <value>hdfs://localhost:900</value>
    e. </property>
    f. </configuration>

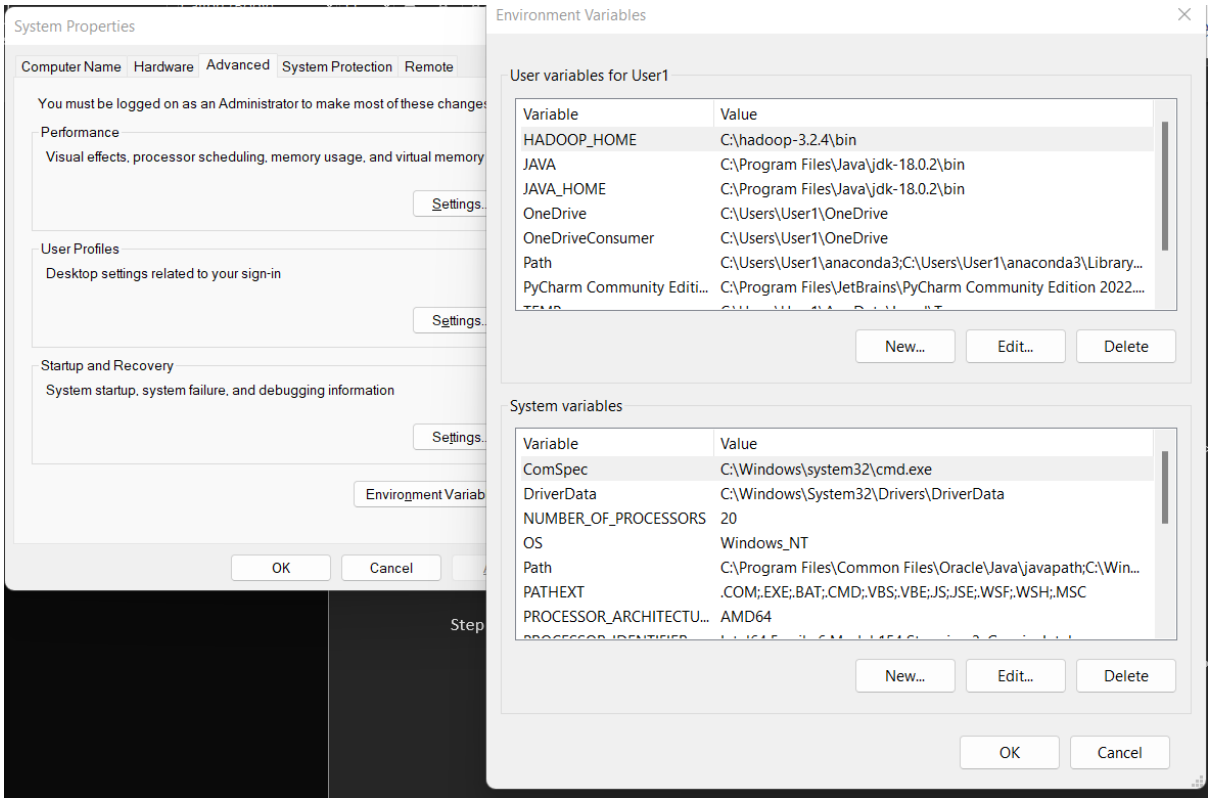
```xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    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.defaulterFS</name>
        <value>hdfs://localhost:900</value>
    </property>
</configuration>

```

5. The second file that need to modify is hdfs.xml file add the below code into that file
    a. <configuration>
    b.
    c. <property>
    d. <name>dfs.replication</name>
    e. <value>1</value>
    f. </property>
    g.
    h. <property>
    i. <name>dfs.namenode.name.dir</name>
    j. <value>C:\hadoop-3.2.4\data\namenode</value>
    k. </property>
    l.
    m. <property>
    n. <name>dfs.datanode.name.dir</name>
    o. <value>C:\hadoop-3.2.4\data\datanode</value>
    p. </property>
6. The third file need to configure is mapred.xml which is useful for performing the map reduce operation which is the main functionality of Hadoop to configure this use below code.
    a. <configuration>
    b. <property>
    c. <name>mapreduce.framework.name</name>
    d. <value> yarn</value>
    e. </property>
    f.
    g. </configuration>
7. The last 4th file need to configure is yarn-site.xml which is a framework which keep track of other data clusters which helps in location of datanode and namenode clusters for that use below code.

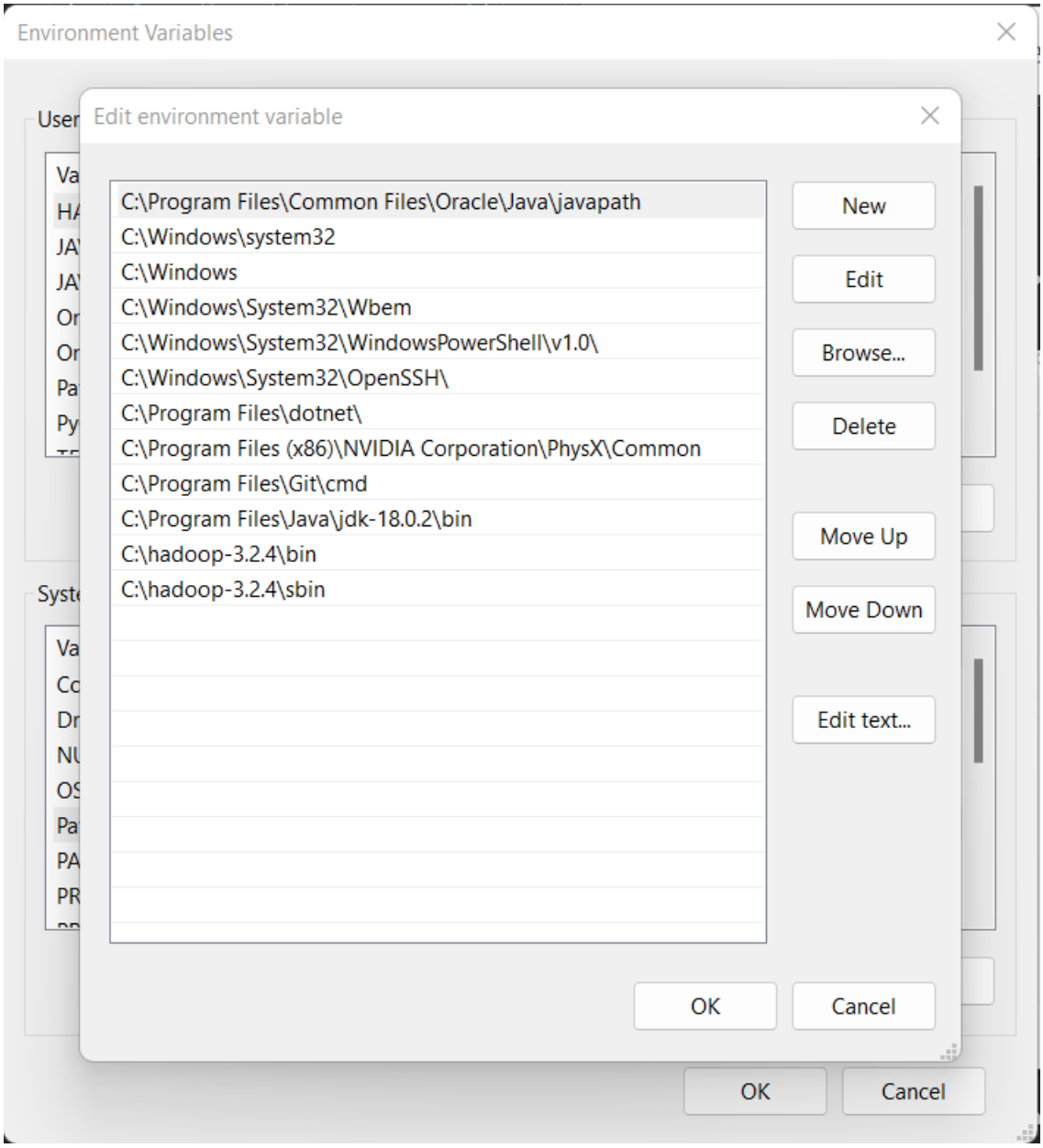
```
a.  <configuration>
b.
c.  <!-- Site specific YARN configuration properties -->
d.
e.      <property>
f.          <name>yarn.nodemanager.aux-services</name>
g.          <value>mapreduce_shuffle</value>
h.      </property>
i.
j.      <property>
k.          <name>yarn.nodemanager.auxservices.mapreduce.shuffle</name>
l.          <value>org.apache.hadoop.mapred.shufflehandler</value>
m.      </property>
n.
o.  </configuration>
```

Step 4: Setting up the environment path

1. In order to execute the Hadoop code we have to set the path of bin and sbin folder into the environmental variable for that search for environmental variable.



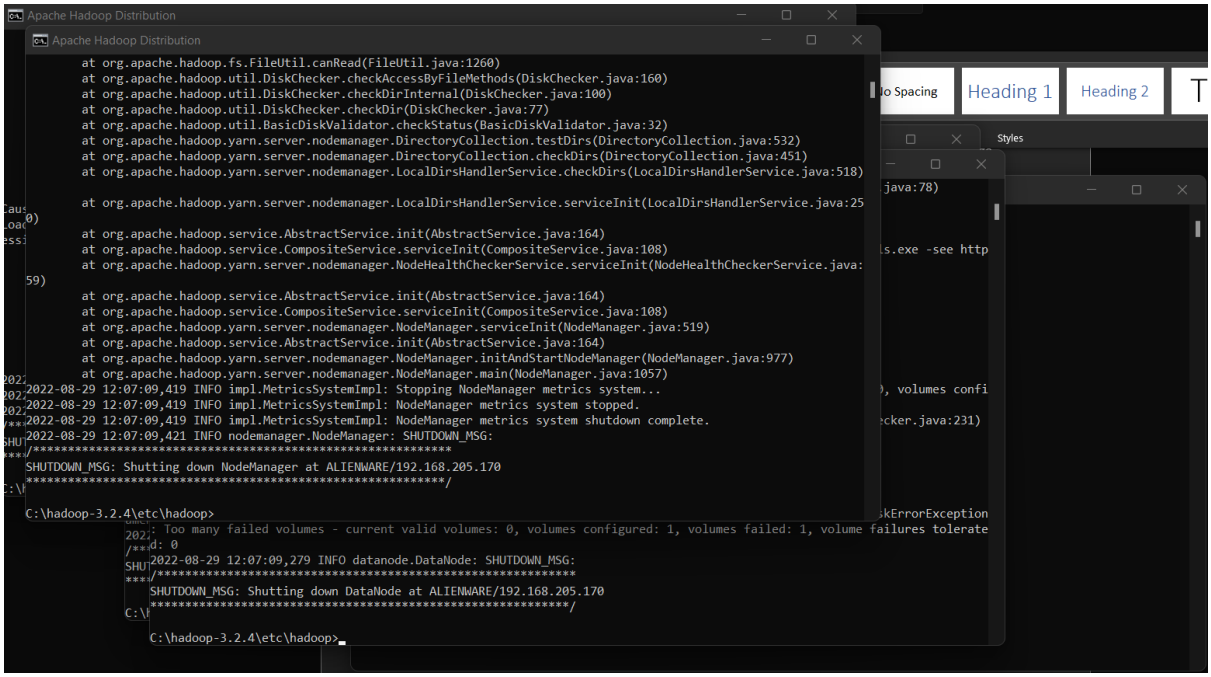
2. On the first section of user variable create new variable and HADOOP_HOME and set the path till bin directory.
3. Similarly for system variable click on path and click on edit you will se the new window.



4. Once you see this window add new path of sbin and bin of Hadoop so that system can identify the base path of Hadoop installation and click on ok and save the data.
5. In order to see if the installation is successful or not go to command prompt and change your current path from xyz location to Hadoop bin directory and use command
    a. Jps

```
C:\hadoop-3.2.4\bin>jps
47428 Jps

C:\hadoop-3.2.4\bin>
```

6. If the installation is successful you see some node number here for all the services of Hadoop
7. To run the Hadoop go to the etc folder and use command start-all.cmd to start the Hadoop server with hall available services.



8. After a proper installation you can see the services are running accurately of some port no and now you are good to go with Hadoop distributed system.

Conclusion – Thus we have studied about the installation of Hadoop and required configuration of Hadoop.