

Practical No: 6

Date: 17/04/2023

Hadoop Installation:

Aim: Install, configure and run Hadoop and HDFS and explore HDFS on Windows

Steps:

Steps to Install Hadoop

1. Install Java JDK 1.8
2. Download Hadoop and extract and place under C drive
3. Set Path in Environment Variables
4. Config files under Hadoop directory
5. Create folder datanode and namenode under data directory
6. Edit HDFS and YARN files
7. Set Java Home environment in Hadoop environment
8. Setup Complete. Test by executing start-all.cmd

There are two ways to install Hadoop, i.e.

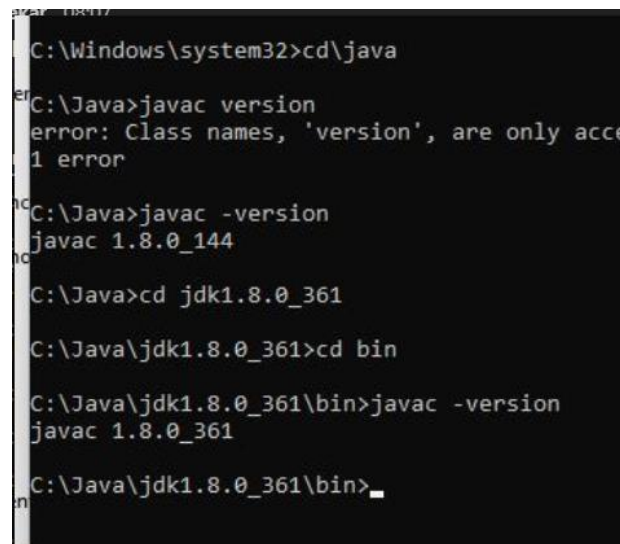
9. Single node

10. Multi node

Here, we use multi node cluster.

1. Install Java

- – Java JDK Link to download
 - <https://www.oracle.com/java/technologies/javase-jdk8-downloads.html>
- – extract and install Java in C:\Java
- – open cmd and type -> javac -version

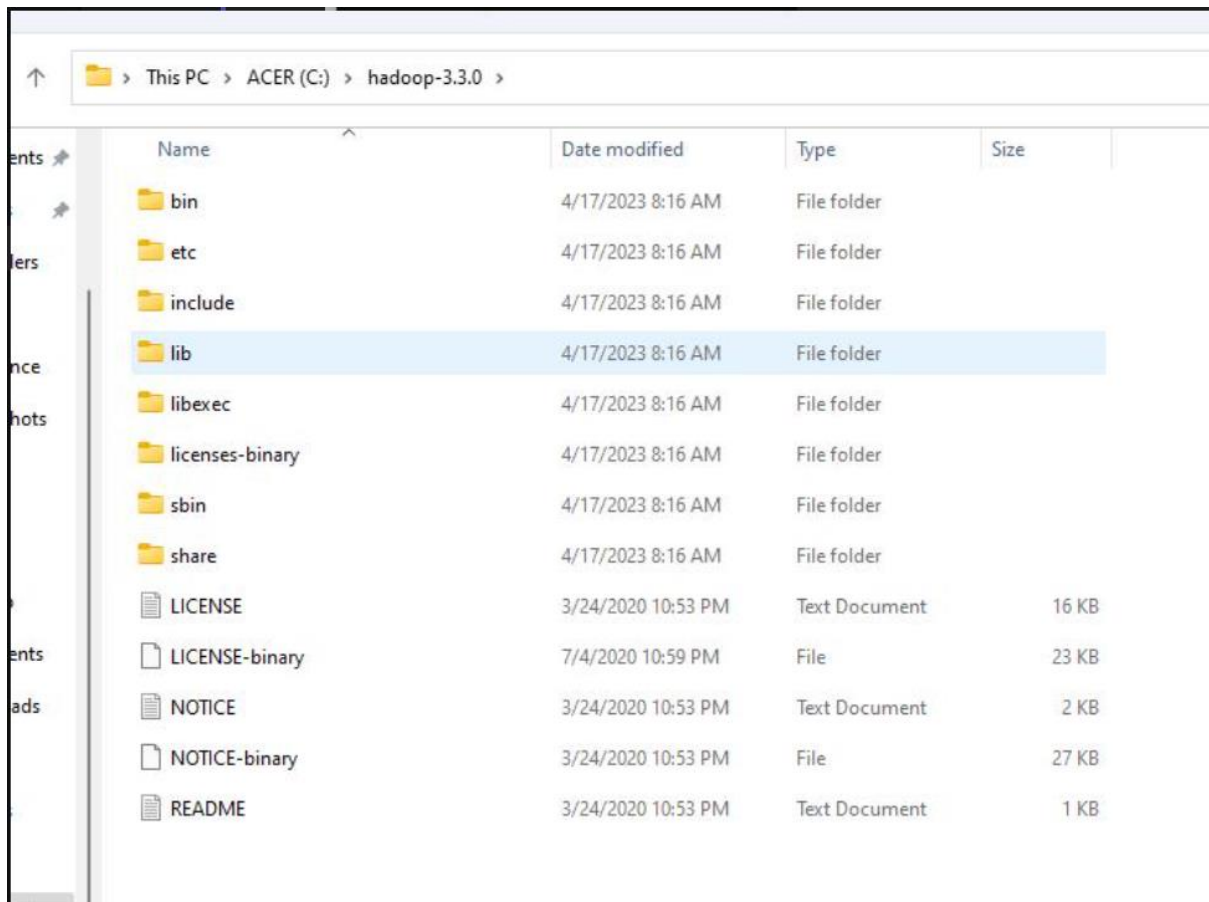


```
C:\Windows\system32>cd\java
C:\Java>javac version
error: Class names, 'version', are only accepted if used by javac
1 error
C:\Java>javac -version
javac 1.8.0_144
C:\Java>cd jdk1.8.0_361
C:\Java\jdk1.8.0_361>cd bin
C:\Java\jdk1.8.0_361\bin>javac -version
javac 1.8.0_361
C:\Java\jdk1.8.0_361\bin>
```

2. Download Hadoop

<https://www.apache.org/dyn/closer.cgi/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz>

- right click .rar.gz file -> show more options -> 7-zip->and extract to C:\Hadoop-3.3.0\

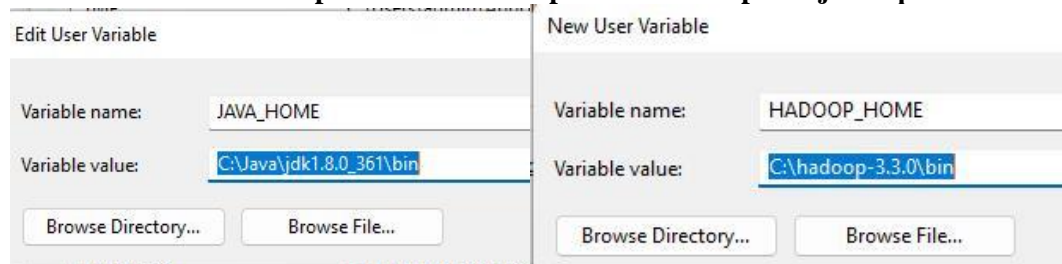


3. Set the path JAVA_HOME Environment variable

4. Set the path HADOOP_HOME Environment variable

Click on New to both user variables and system variables.

Click on user variable -> path -> edit-> add path for Hadoop and java upto 'bin'



Click Ok, Ok, Ok.

5. Configurations

Edit file C:/Hadoop-3.3.0/etc/hadoop/core-site.xml,

paste the xml code in folder and save

```
<configuration>
<property>
  <name>fs.defaultFS</name>
  <value>hdfs://localhost:9000</value>
</property>
</configuration>
```

=====

Rename “mapred-site.xml.template” to “mapred-site.xml” and edit this file C:/Hadoop-3.3.0/etc/hadoop/mapred-site.xml, paste xml code and save this file.

=====

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

=====

Create folder “data” under “C:\Hadoop-3.3.0”
Create folder “datanode” under “C:\Hadoop-3.3.0\data”
Create folder “namenode” under “C:\Hadoop-3.3.0\data”

=====

Edit file C:/Hadoop-3.3.0/etc/hadoop/hdfs-site.xml, paste xml code and save this file.

```
<configuration>
<property>
  <name>dfs.replication</name>
  <value>1</value>
</property>

<property>
  <name>dfs.namenode.name.dir</name>
  <value>/hadoop-3.3.0/data/namenode</value>
</property>
<property>
  <name>dfs.datanode.data.dir</name>
  <value>/hadoop-3.3.0/data/datanode</value>
</property>
</configuration>
```

=====

Edit file C:/Hadoop-3.3.0/etc/hadoop/yarn-site.xml, paste xml code and save this file.

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

  <property>
    <name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
    <value>org.apache.hadoop.mapred.ShuffleHandler</value>
  </property>
  <property>
    <name>yarn.resourcemanager.address</name>
    <value>127.0.0.1:8032</value>
  </property>
  <property>
    <name>yarn.resourcemanager.scheduler.address</name>
    <value>127.0.0.1:8030</value>
```

```

</property>
<property>
    <name>yarn.resourcemanager.resource-tracker.address</name>
    <value>127.0.0.1:8031</value>
</property>
</configuration>

```

6. Edit file C:/Hadoop-3.3.0/etc/hadoop/hadoop-env.cmd

Find "JAVA_HOME=%JAVA_HOME%" and replace it as
 set JAVA_HOME="C:\Java\jdk1.8.0_361"

7. Download "redistributable" package

Download and run VC_redist.x64.exe

8. Hadoop Configurations

Download bin folder from <https://github.com/s911415/apache-hadoop-3.1.0-winutils>

– Copy the bin folder to c:\hadoop-3.3.0. Replace the existing bin folder.

9. copy "hadoop-yarn-server-timelineservice-3.0.3.jar" from ~\hadoop-3.0.3\share\hadoop\yarn\timelineservice to ~\hadoop-3.0.3\share\hadoop\yarn folder.

10. Format the NameNode

– Open cmd 'Run as Administrator' and type command "hdfs namenode –format"

```

2023-04-17 09:14:34,324 INFO util.GSet: VM type = 64-bit
2023-04-17 09:14:34,324 INFO util.GSet: 0.029999999329447746% max memory 889 MB = 273.1 KB
2023-04-17 09:14:34,325 INFO util.GSet: capacity = 2^15 = 32768 entries
Re-format filesystem in Storage Directory root= C:\hadoop-3.3.0\data\namenode; location= null ? (Y or N) y
2023-04-17 09:14:55,710 INFO namenode.FSImage: Allocated new BlockPoolId: BP-813810208-169.254.162.181-16817
2023-04-17 09:14:55,711 INFO common.Storage: Will remove files: [C:\hadoop-3.3.0\data\namenode\current\fsimage_00000000000000000000000000000000.md5, C:\hadoop-3.3.0\data\namenode\current\fsimage_00000000000000000000000000000000]
2023-04-17 09:14:55,752 INFO common.Storage: Storage directory C:\hadoop-3.3.0\data\namenode has been successfully formatted.
2023-04-17 09:14:55,775 INFO namenode.FSImageFormatProtobuf: Saving image file C:\hadoop-3.3.0\data\namenode\current\image.ckpt_00000000000000000000000000000000 using no compression
2023-04-17 09:14:55,840 INFO namenode.FSImageFormatProtobuf: Image file C:\hadoop-3.3.0\data\namenode\current\image.ckpt_00000000000000000000000000000000 of size 397 bytes saved in 0 seconds .
2023-04-17 09:14:55,856 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2023-04-17 09:14:55,862 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2023-04-17 09:14:55,863 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at VSIT-X121-15/169.254.162.181
*****/
C:\hadoop-3.3.0\bin>

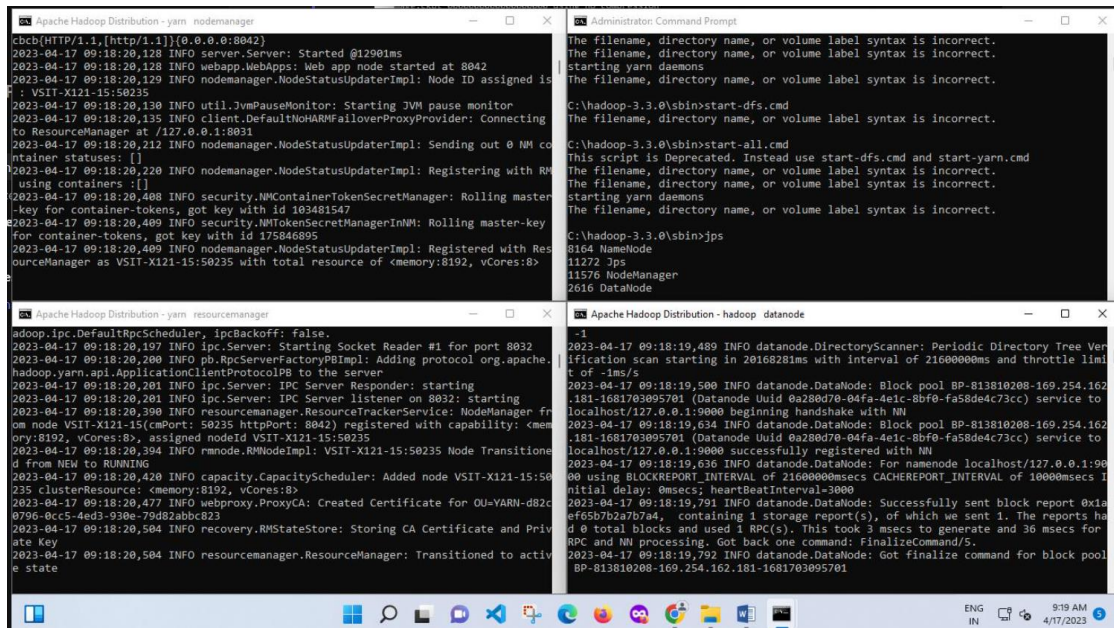
```

11. Testing

- Open cmd 'Run as Administrator' and change directory to C:\Hadoop-3.3.0\sbin
- type start-all.cmd
- OR
- type start-dfs.cmd
- type start-yarn.cmd

```
C:\hadoop-3.3.0\sbin>start-all.cmd
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd
The filename, directory name, or volume label syntax is incorrect.
The filename, directory name, or volume label syntax is incorrect.
starting yarn daemons
The filename, directory name, or volume label syntax is incorrect.
```

– You will get 4 more running threads for Datanode, namenode, resource manager and node manager



Output:

12. Type JPS command to start-all.cmd command prompt, you will get following output.

```
C:\hadoop-3.3.0\sbin>jps
8164 NameNode
11272 Jps
11576 NodeManager
2616 DataNode
2952 ResourceManager
C:\hadoop-3.3.0\sbin>
```

13. Run <http://localhost:9870/> from any browser

Overview 'localhost:9000' (✓active)

Started:	Mon Apr 17 09:18:18 +0530 2023
Version:	3.3.0, raa96f1871bfd858f9bac59cf2a81ec470da649af
Compiled:	Tue Jul 07 00:14:00 +0530 2020 by brahma from branch-3.3.0
Cluster ID:	CID-7483febe-8254-46e7-bac1-9e1d4d8cbee0
Block Pool ID:	BP-813810208-169.254.162.181-1681703095701

Summary

Security is off.

Browse Directory

/ Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
No data available in table							

Showing 0 to 0 of 0 entries

Previous Next

Hadoop, 2020.