**Hadoop Installation on Windows 10**

As a beginner, you might feel reluctant in performing cloud computing which requires subscriptions. While you can install a virtual machine as well in your system, it requires allocation of a large amount of RAM for it to function smoothly else it would hang constantly.

You can install Hadoop in your system as well which would be a feasible way to learn Hadoop.

We will be installing single node pseudo-distributed hadoop cluster on windows 10.

**Prerequisite**: To install Hadoop, you should have Java version 1.8 in your system.

Check your java version through this command on command prompt

|  |  |
| --- | --- |
| 1 | java –version |

Text

Description automatically generated

If java is not installed in your system, then –

Go this link –

[https://www.oracle.com/technetwork/java/javase/downloads/jdk8-downl...](https://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html)

Accept the license,

Table

Description automatically generated

Download the file according to your operating system. Keep the java folder directly under the local disk directory (C:\Java\jdk1.8.0\_152) rather than in Program Files (C:\Program Files\Java\jdk1.8.0\_152) as it can create errors afterwards.

Graphical user interface, application

Description automatically generated

After downloading java version 1.8, download hadoop version 3.1 from this link –

[https://archive.apache.org/dist/hadoop/common/hadoop-3.1.0/hadoop-3...](https://archive.apache.org/dist/hadoop/common/hadoop-3.1.0/hadoop-3.1.0.tar.gz)

Extract it to a folder.

Graphical user interface, application, table

Description automatically generated

**Setup System Environment Variables**

Open control panel to edit the system environment variable

Graphical user interface, text, application, email

Description automatically generated

Go to environment variable in system properties

Graphical user interface, text, application, email

Description automatically generated

Create a new user variable. Put the Variable\_name as HADOOP\_HOME and Variable\_value as the path of the bin folder where you extracted hadoop.

Graphical user interface, text, application, email

Description automatically generated

Likewise, create a new user variable with variable name as JAVA\_HOME and variable value as the path of the bin folder in the Java directory.

Graphical user interface, text, application, email

Description automatically generated

Now we need to set Hadoop bin directory and Java bin directory path in system variable path.

Edit Path in system variable

Graphical user interface, text, application

Description automatically generated

Click on New and add the bin directory path of Hadoop and Java in it.

Graphical user interface, text, application

Description automatically generated

**Configurations**

Now we need to edit some files located in the hadoop directory of the etc folder where we installed hadoop. The files that need to be edited have been highlighted.

Graphical user interface, table

Description automatically generated with medium confidence

1. Edit the file core-site.xml in the hadoop directory. Copy this xml property in the configuration in the file

|  |  |
| --- | --- |
| 1  2  3  4  5  6 | /span>configuration>     /span>property>         /span>name>fs.defaultFS/span>/name>         /span>value>hdfs://localhost:9000</value>     /span>/property>  /span>/configuration> |

2. Edit mapred-site.xml and copy this property in the cofiguration

|  |  |
| --- | --- |
| 1  2  3  4  5  6 | /span>configuration>     /span>property>         /span>name>mapreduce.framework.name/span>/name>         /span>value>yarn/span>/value>     /span>/property>  /span>/configuration> |

3. Create a folder ‘data’ in the hadoop directory

Graphical user interface, table

Description automatically generated

Create a folder with the name ‘datanode’ and a folder ‘namenode’ in this data directory

Graphical user interface, text, application

Description automatically generated

4. Edit the file hdfs-site.xml and add below property in the configuration

Note: The path of namenode and datanode across value would be the path of the datanode and namenode folders you just created.

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14 | /span>configuration>     /span>property>         /span>name>dfs.replication/span>/name>         /span>value>1/span>/value>     /span>/property>     /span>property>         /span>name>dfs.namenode.name.dir/span>/name>         /span>value>C:\Users\hp\Downloads\hadoop-3.1.0\hadoop-3.1.0\data\namenode/span>/value>     /span>/property>     /span>property>         /span>name>dfs.datanode.data.dir/span>/name>         /span>value> C:\Users\hp\Downloads\hadoop-3.1.0\hadoop-3.1.0\data\datanode/span>/value>     /span>/property>  /span>/configuration> |

5. Edit the file yarn-site.xml and add below property in the configuration

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10 | /span>configuration>     /span>property>       /span>name>yarn.nodemanager.aux-services/span>/name>       /span>value>mapreduce\_shuffle/span>/value>     /span>/property>     /span>property>         /span>name>yarn.nodemanager.auxservices.mapreduce.shuffle.class/span>/name>  /span>value>org.apache.hadoop.mapred.ShuffleHandler/span>/value>     /span>/property>  /span>/configuration> |

6. Edit hadoop-env.cmd and replace %JAVA\_HOME% with the path of the java folder where your jdk 1.8 is installed

Graphical user interface, text, application, email

Description automatically generated

Hadoop needs windows OS specific files which does not come with default download of hadoop.

To include those files, replace the bin folder in hadoop directory with the bin folder provided in this github link.

<https://github.com/s911415/apache-hadoop-3.1.0-winutils>

Download it as zip file. Extract it and copy the bin folder in it. If you want to save the old bin folder, rename it like bin\_old and paste the copied bin folder in that directory.

Graphical user interface, application, table

Description automatically generated

Check whether hadoop is successfully installed by running this command on cmd-

|  |  |
| --- | --- |
| 1 | hadoop version |

Text

Description automatically generated

Since it doesn’t throw error and successfully shows the hadoop version, that means hadoop is successfully installed in the system.

**Format the NameNode**

Formatting the NameNode is done once when hadoop is installed and not for running hadoop filesystem, else it will delete all the data inside HDFS. Run this command-

|  |  |
| --- | --- |
| 1 | hdfs namenode –format |

It would appear something like this –

Graphical user interface, text

Description automatically generated

Graphical user interface, text

Description automatically generated

Now change the directory in cmd to sbin folder of hadoop directory with this command,

(Note: Make sure you are writing the path as per your system)

|  |  |
| --- | --- |
| 1 | cd C:\Users\hp\Downloads\hadoop-3.1.0\hadoop-3.1.0\sbin |

Start namenode and datanode with this command –

|  |  |
| --- | --- |
| 1 | start-dfs.cmd |

Text

Description automatically generated

Two more cmd windows will open for NameNode and DataNode

Now start yarn through this command-

|  |  |
| --- | --- |
| 1 | start-yarn.cmd |

Text

Description automatically generated

Two more windows will open, one for yarn resource manager and one for yarn node manager.

Text

Description automatically generated

Note: Make sure all the 4 Apache Hadoop Distribution windows are up n running. If they are not running, you will see an error or a shutdown message. In that case, you need to debug the error.

To access information about resource manager current jobs, successful and failed jobs, go to this link in browser-

<http://localhost:8088/cluster>

Graphical user interface, application, Word

Description automatically generated

To check the details about the hdfs (namenode and datanode),

Open this link on browser-

<http://localhost:9870/>

Note: If you are using Hadoop version prior to 3.0.0 – Alpha 1, then use port <http://localhost:50070/>

Graphical user interface, text, application, email

Description automatically generated

**Working with HDFS**

I will be using a small text file in my local file system. To put it in hdfs using hdfs command line tool.

I will create a directory named ‘sample’ in my hadoop directory using the following command-

|  |  |
| --- | --- |
| 1 | hdfs dfs –mkdir /sample |



To verify if the directory is created in hdfs, we will use ‘ls’ command which will list the files present in hdfs –

|  |  |
| --- | --- |
| 1 | hdfs dfs –ls / |

Text

Description automatically generated

Then I will copy a text file named ‘potatoes’ from my local file system to this folder that I just created in hdfs using copyFromLocal command-

|  |  |
| --- | --- |
| 1 | hdfs dfs -copyFromLocal C:\Users\hp\Downloads\potatoes.txt /sample |



To verify if the file is copied to the folder, I will use ‘ls’ command by specifying the folder name which will read the list of files in that folder –

|  |  |
| --- | --- |
| 1 | hdfs dfs –ls /sample |

Text

Description automatically generated

To view the contents of the file we copied, I will use cat command-

|  |  |
| --- | --- |
| 1 | hdfs dfs –cat /sample/potatoes.txt |

A picture containing text

Description automatically generated

To Copy file from hdfs to local directory, I will use get command –

|  |  |
| --- | --- |
| 1 | hdfs dfs -get /sample/potatoes.txt C:\Users\hp\Desktop\priyanka |

Text

Description automatically generated

These were some basic hadoop commands. You can refer to this HDFS commands guide to learn more.

[https://hadoop.apache.org/docs/r3.1.0/hadoop-project-dist/hadoop-hd...](https://hadoop.apache.org/docs/r3.1.0/hadoop-project-dist/hadoop-hdfs/HDFSCommands.html)

Hadoop MapReduce can be used to perform data processing activity. However, it possessed limitations due to which frameworks like Spark and Pig emerged and have gained popularity. A 200 lines of MapReduce code can be written with less than 10 lines of Pig code. Hadoop has various other components in its ecosystem like Hive, Sqoop, Oozie, and HBase. You can download these software as well in your windows system to perform data processing operations using cmd.