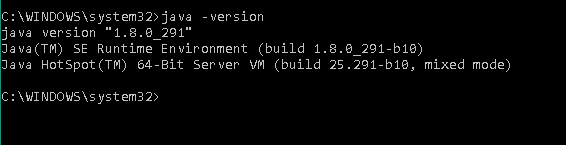
Installation of Hadoop

Prerequisite:

Should have Java version 1.8 in your system.

Check your java version through this command on command prompt

java –version



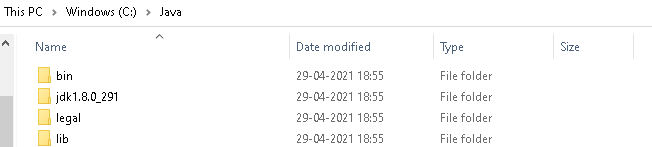
If the java version is not installed please visit the oracle website:

<https://www.oracle.com/java/technologies/javase/javase-jdk8-downloads.html>

Select and download the jdk according to your operating system. You have to create or login to your oracle account which is a free account and accept the license agreement to download the jdk file.

While installation of Java, create a new folder directly in your C drive called “Java”, therefore it will not create any problem while running hadoop in your system.

The path should be C:\Java\jdk1.8.0\_291



Copy the jdk file after java installation from C:\Program Files\Java\ jdk1.8.0\_291 to the specified path above. Delete the Java file inside the Program Files directory to avoid duplicates.

Download the hadoop files from:

<https://hadoop.apache.org/releases.html> by clicking on binary for the version you want to install.

Extract it to folder and paste it directly into the C drive. C:\hadoop

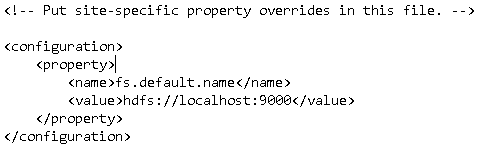
Create a new folder called data and create the respective subfolder datanode and namenode.

C:\hadoop\data\datanode & C:\hadoop\data\namenode

Go to C:\hadoop\etc\hadoop to configure 5 files:

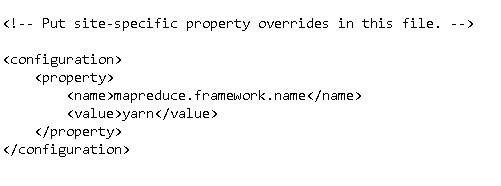
* core-site.xml

Edit the file by writing the properties:



* mapred-site.xml

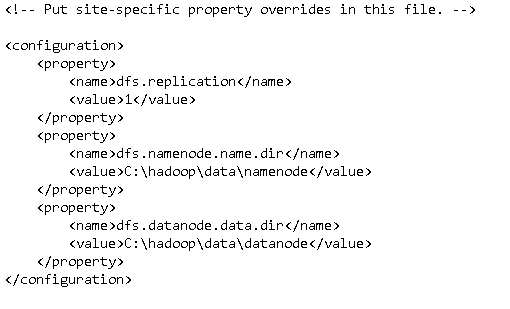
Edit the file by writing the properties:



* hdfs-site.xml

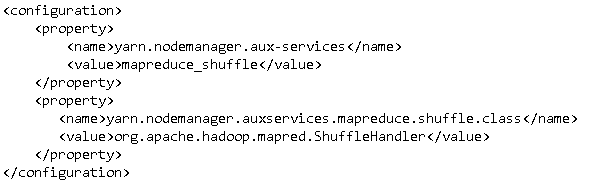
Edit the file by writing the properties:

Note:- Please make sure you have created the respective datanode and namenode folders inside the data directory.



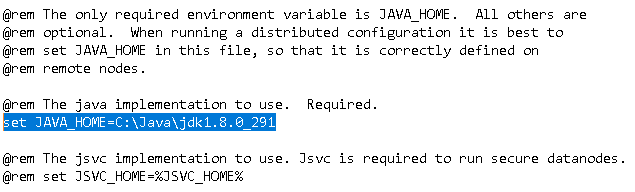
* yarn-site.xml

Edit the file by writing the properties:



* hadoop-env.cmd

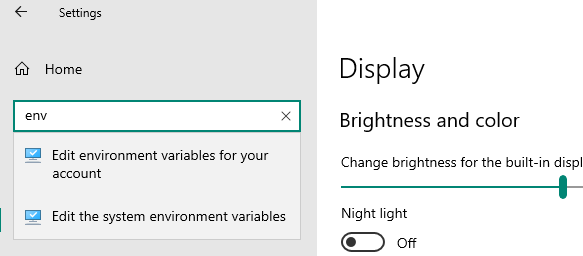
Edit the file by replacing %JAVA\_HOME% with the path of the java folder where your jdk is installed:



Set the environment variables:

Go to settings, select System

Search for env in the search bar in the left corner:

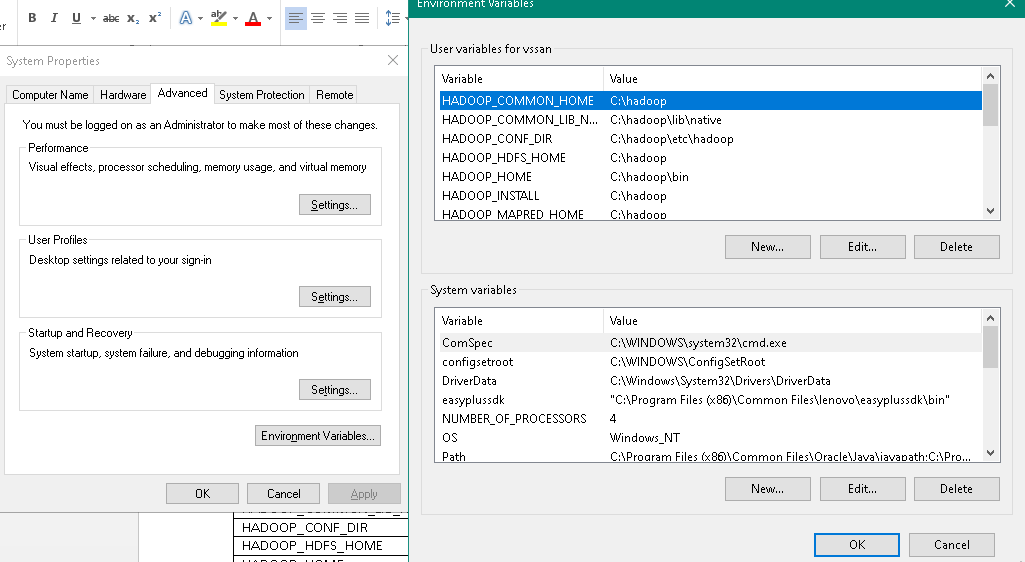


Select Edit the system environment variables, from the System properties pop window on Advanced section click on Environment variables:

|  |  |
| --- | --- |
| **Variable** | **Value** |
| HADOOP\_COMMON\_HOME | C:\hadoop |
| HADOOP\_COMMON\_LIB\_NATIVE\_DIR | C:\hadoop\lib\native |
| HADOOP\_CONF\_DIR | C:\hadoop\etc\hadoop |
| HADOOP\_HDFS\_HOME | C:\hadoop |
| HADOOP\_HOME | C:\hadoop\bin |
| HADOOP\_INSTALL | C:\hadoop |
| HADOOP\_MAPRED\_HOME | C:\hadoop |
| JAVA\_HOME | C:\Java\jdk1.8.0\_291 |
| YARN\_HOME | C:\hadoop |

Enter the above variables and value respectively under user variables for vssan.

Make sure the respective paths specified in value are present inside the System variables Path as well.

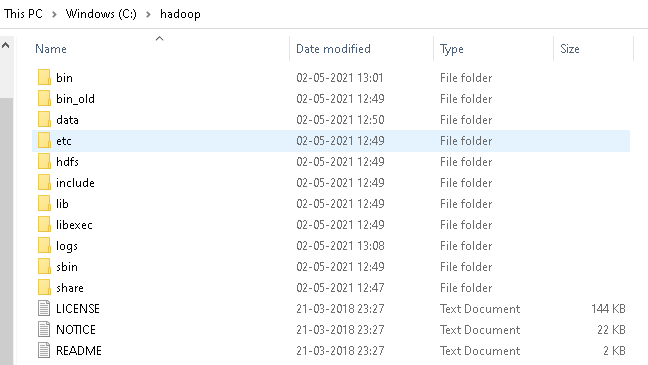


Hadoop needs windows OS specific files which does not come with default download of hadoop.

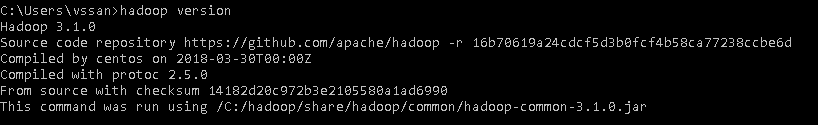
To include those files, replace the bin folder in hadoop directory with the bin folder provided in this github link.

<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.



To check whether hadoop is successfully installed, run the command in cmd prompt:



Format the Namenode by running the command:

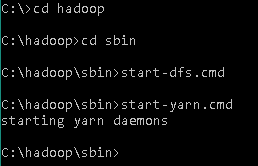
* hdfs namenode –format

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

* cd hadoop/sbin

Start the dfs and yarn deamons by running the command one after the other:

* start-dfs.cmd
* start-yarn.cmd



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>

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

Open this link on browser-

<http://localhost:9870/>

hdfs command practice:

To create a new directory:

* hdfs dfs -mkdir -p /usr/hadoop/data

To list the number of files in the directory:

* hdfs dfs -ls /usr/hadoop/data

To view a file in the directory:

* hdfs dfs -cat /usr/hadoop/data/data.txt

Map Reducer:

* hadoop jar C:\hadoop\share\hadoop\tools\lib\hadoop-streaming-3.1.0.jar -file mapper.py -mapper mapper.py -file reducer.py -reducer reducer.py -input /usr/hadoop/data/data.txt -output /usr/hadoop/data/output3

Error Surfaced During Installation of Hadoop:

Error1:

YARN localizer.ResourceLocalizationService: Failed to setup local dir /tmp/hadoop-vssan/nm-local-dir, which was marked as good. org.apache.hadoop.yarn.exceptions.YarnRuntimeException: Permissions incorrectly set for dir /tmp/hadoop-vssan/nm-local-dir/filecache, should be rwxr-xr-x, actual value = rwxrwxr-x

Resolved by running the command prompt in Admin mode.

Error2:

FATAL resourcemanager.ResourceManager: Error starting ResourceManager java.lang.NoClassDefFoundError: org/apache/hadoop/yarn/server/timelineservice/collector/TimelineCollectorManager

Resolved by following this method:

<https://stackoverflow.com/questions/51118358/noclassdeffounderror-org-apache-hadoop-yarn-server-timelineservice-collector-tim>

Error3:

Task Id : attempt\_1620181620749\_0001\_m\_000001\_0, Status : FAILEDError: java.lang.RuntimeException: Error in configuring object

Unable to ressolve

Reference:

<https://www.datasciencecentral.com/profiles/blogs/how-to-install-and-run-hadoop-on-windows-for-beginners>

<https://www.youtube.com/watch?v=26w6lL8_7Q8>

<https://data-flair.training/blogs/top-hadoop-hdfs-commands-tutorial/>