**BDA LAB 2**

**Exploring Hadoop**

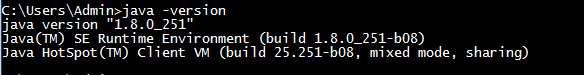
**AIM:** Installation of Apache hadoop 3.2.1 on windows OS.

**Hadoop:**

* Hadoop is a software library that allows distributed processing and storing of data across large clusters of computers.
* It offers high availability of data even on the commodity hardware resources.
* Hadoop is developed by apache foundation and it is an open-source framework.
* Many companies use this java-based solution for their big data-related needs.

**Task 1:** Download and install hadoop on your workstation or seek alternatives. Monitor the same using version specific navigation links.

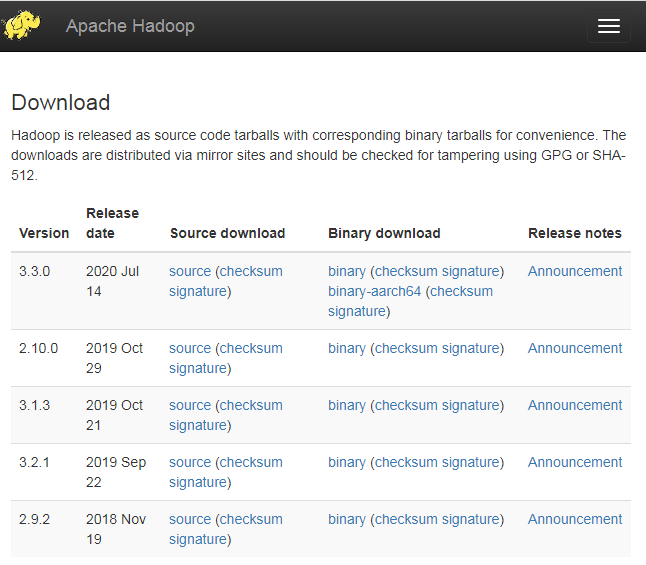
* To install Hadoop first we should have Java version 1.8 in our system.



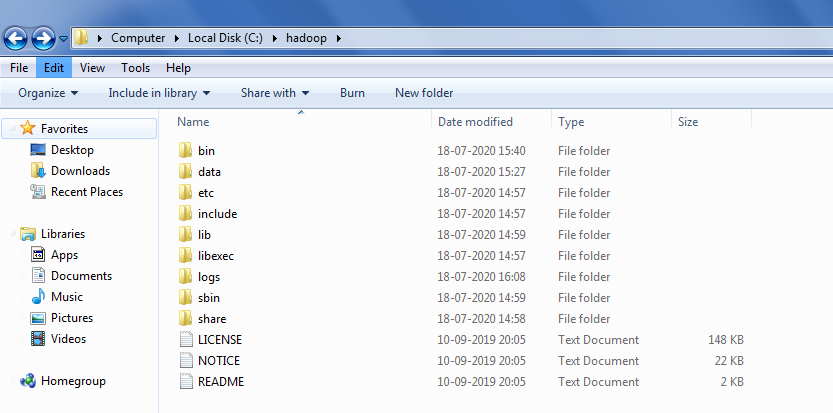
If Java is not installed in our system, we can install it from official website of oracle.

* After downloading java version 1.8 , now we need to download Hadoop version 3.2.1 from the below link:

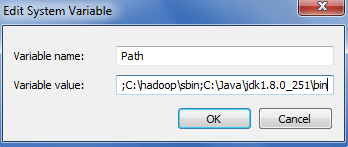
<https://hadoop.apache.org/releases.html>



Extract the file into C drive in hadoop folder.

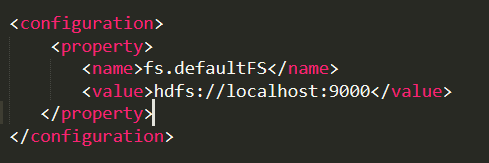


* Now we have to set environment path variables. For this first open system environment variables and the add path.
* Create a new user variable. Put the Variable\_name as HADOOP\_HOME and variable\_value as the path of the bin folder where you extracted hadoop.
* Put Variable value : C:\hadoop\bin
* Similarly, we need to set Hadoop bin directory and Java bin directory path in system variable path.
* Edit path in system variable.
* Click on new and add the bin directory path of Hadoop and Java in it.

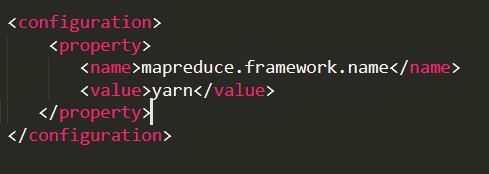


* Now we need to edit some files located in the hadoop directory of the etc folder where we installed hadoop.

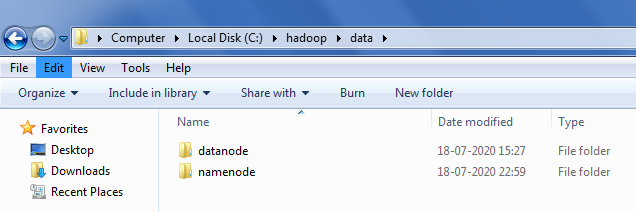
1. Edit the file core-site.xml in the hadoop directory. Copy this xml property in the configuration in the file



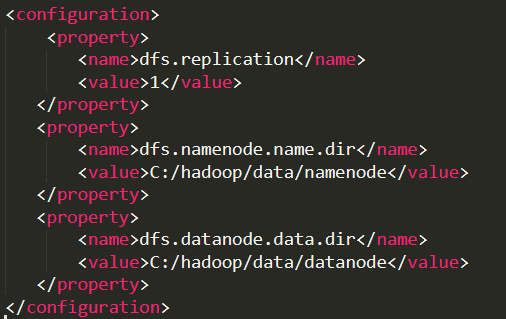
1. Edit mapred-site.xml and copy this property in the configuration.



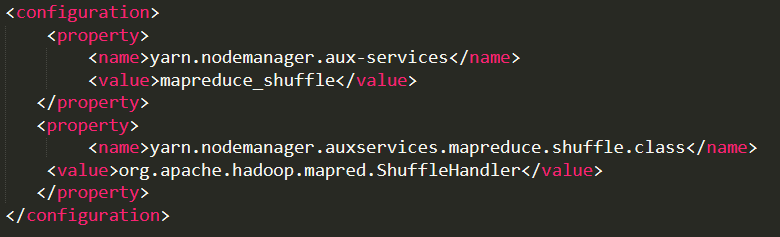
1. Create a folder ‘data’ in the hadoop directory. Then create 2 folders ‘namenode’ , ‘datanode’ in this directory.



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



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



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

C:\Users\Admin\Desktop\10.PNG

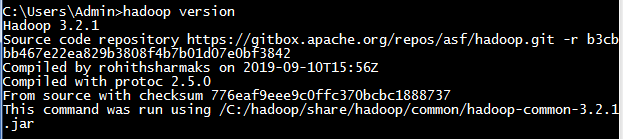
* Hadoop needs windows OS specific files which does not come with default download of hadoop.

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

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

Download it as zip file and extract it and copy the bin folder in it.

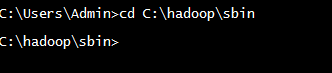
* Check whether hadoop is successfully installed by running this command on cmd-



* 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-



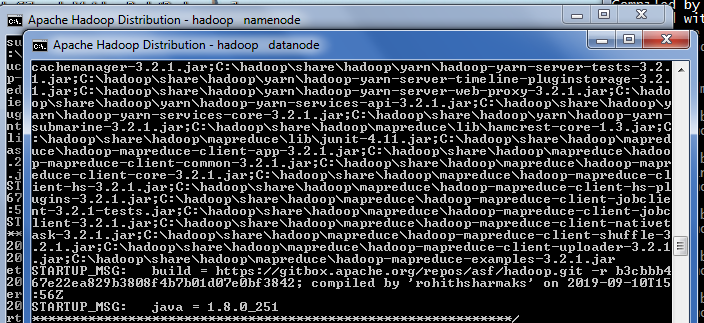
* Now change the directory to sbin folder



* Start namenode and datanode by the command given below

C:\Users\Admin\Desktop\13.PNG

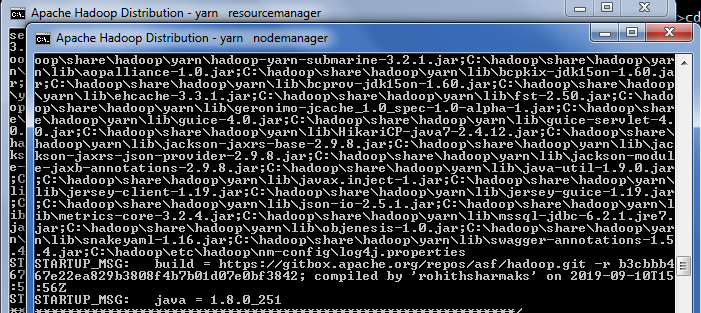
Two more cmd windows will open for NameNode and DataNode.



* Now, start yarn by the given command

C:\Users\Admin\Desktop\14.PNG

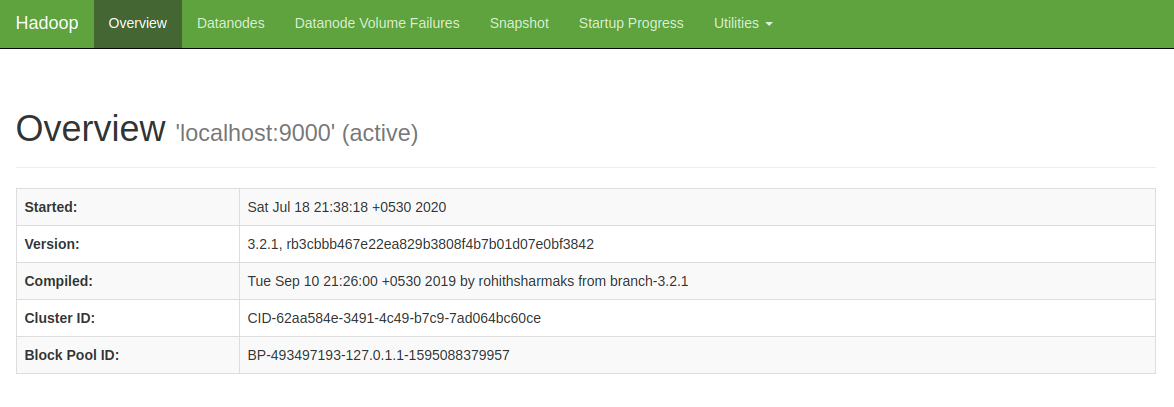
Two more windows will open, one for yarn resource manager and one for yarn node manager.



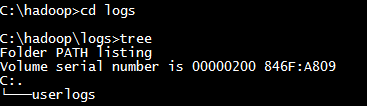
* After executing these 2 commands, make sure all the command windows are open.
* To access information about resource manager current jobs, successful and failed jobs, go this link in browser-

<http://localhost:8088/cluster>

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

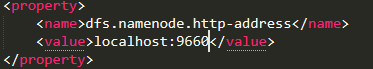
<http://localhost:9870/>

**Task 2:** Locate the log directory and refer for troubleshooting

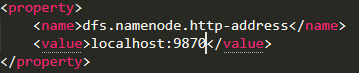


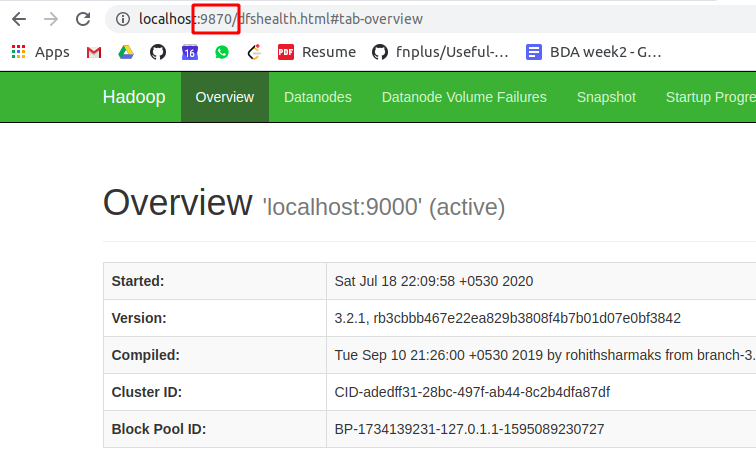
**Task 3:** Change the URL of HDFS Web UI to localhost:51234 instead of localhost:50070 and the access the web UI using new URL.

* For hadoop 3.2.1, web Ui is by default localhost:9870.
* So, for this task, we can change it to localhost:9660.

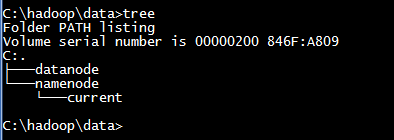


**Task 4:** Undo changes done in task1.



****

**Task 5:** Find out what is current location of namenode and datanode storage directories.



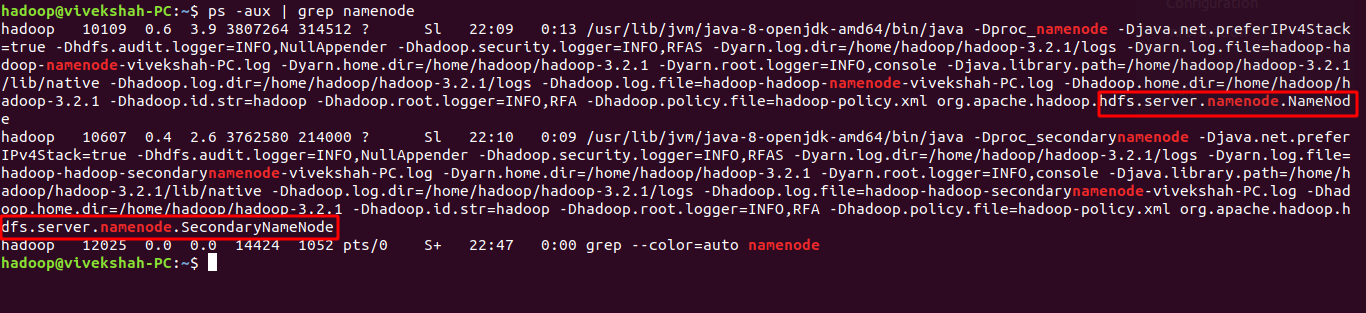
**Task 6:** Analyse the output of following command:

[hadoop@hadoop-clone hadoop]$ ps -ef |grep NameNode

[hadoop@hadoop-clone hadoop]$ ps -ef |grep DataNode

* We know that ps –ef command is used to find PID of the process. Each process will have the unique number which is called as PID of the process.
* And grep is an acronym that stands for Global Regular Expression Print is used to search for a string of character in a specified file. when it finds a match, it prints the line with the result.
* Hence both commands show the result about which process related to Namenode and Datanode respectively.

1. [hadoop@hadoop-clone hadoop]$ ps -ef |grep NameNode

****

1. [hadoop@hadoop-clone hadoop]$ ps -ef |grep DataNode

