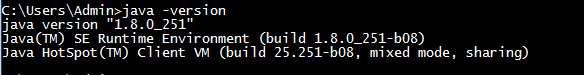
## BDA LAB-2 | Explore hadoop on Single node cluster

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

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

Check your java version through this command on command prompt



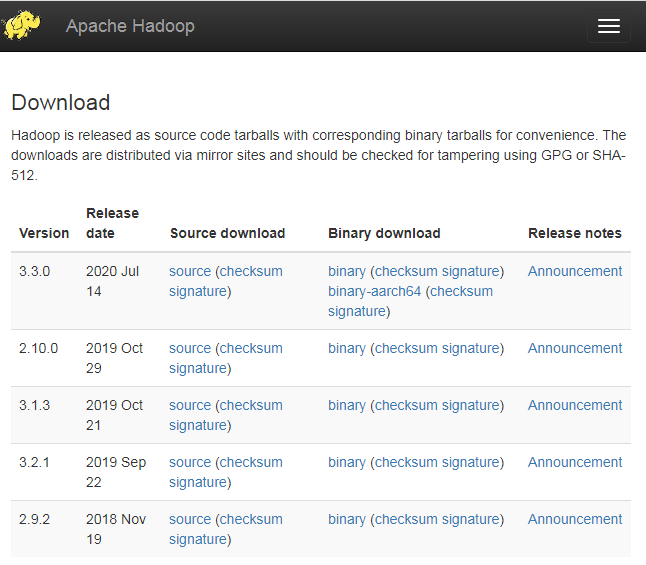
If java is not installed in your system,the –

Download from this link:

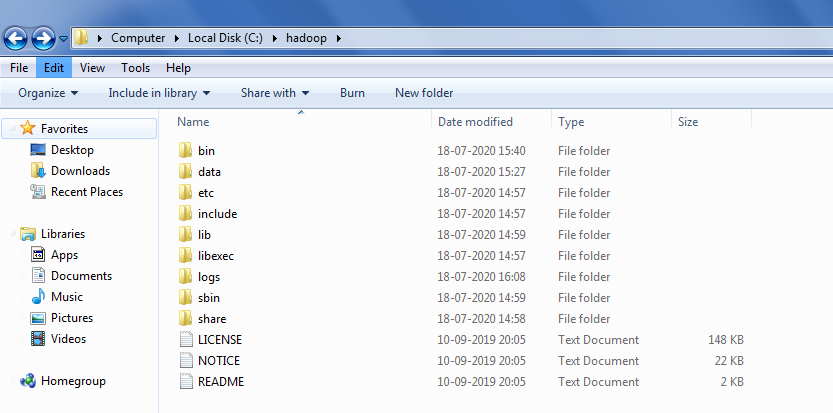
<https://www.oracle.com/java/technologies/javase/javase-jdk8-downloads.html>

After downloading java version 1.8, download hadoop version 3.2.1 from this link –

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



Extract into C: drive in hadoop folder.



Go to environment variable in system properties.

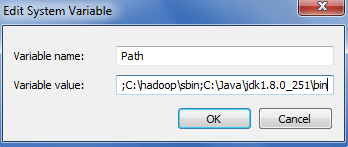
Create a new user variable. Put the Variable\_name as HADOOP\_HOME and variable\_value as the path of the bin folder where you extracted hadoop.

Variable value : C:\hadoop\bin

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

Edit path in system variable.

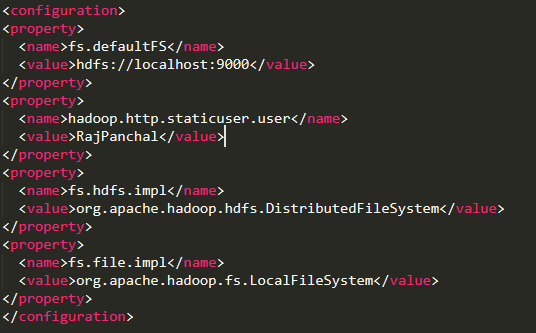
Click on new and add the bin directory path of Hadoop and Java in it.



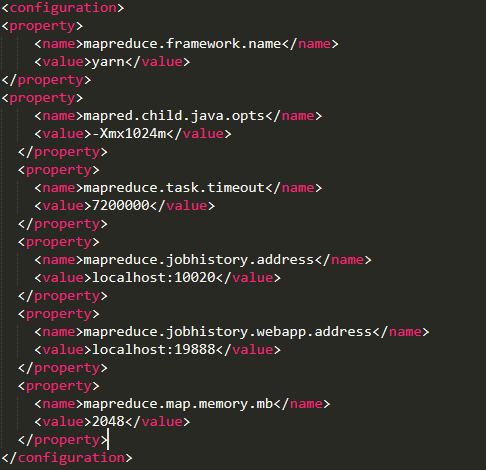
Now we need to edit some files located in the hadoop directory of the etc folder where we installed hadoop. The files that need to be edited have been defined below.

* Core-site.xml
* Hdfs-site.xml
* Mapred-site.xml
* Yarn-site.xml
* Hadoop-env.cmd

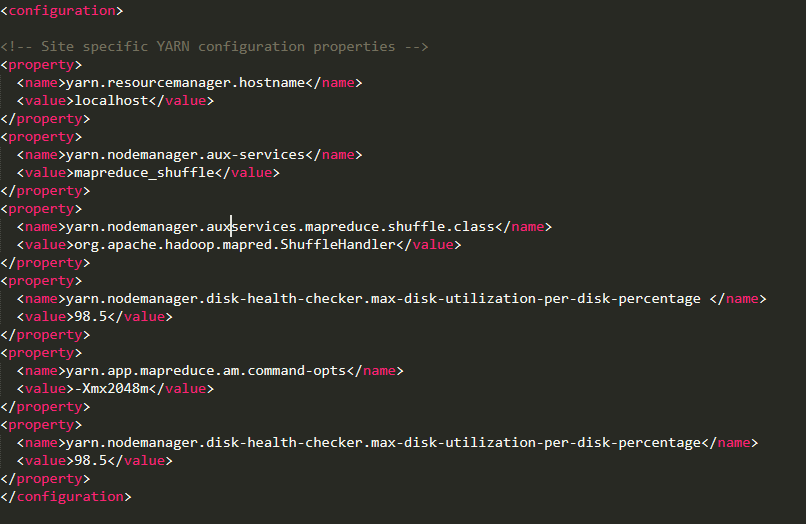
1). Core-site.xml



2). Mapred-site.xml



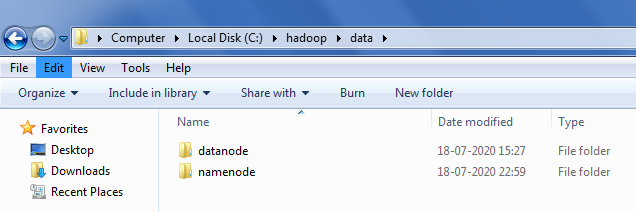
3). Yarn-site.xml



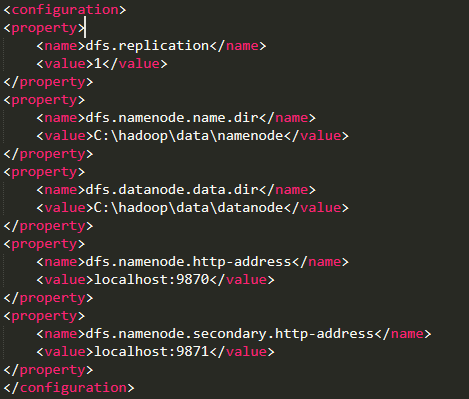
4).Hdfs-site.xml

For this , create a folder ‘data’ in the hadoop directory.

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



Now edit the file hdfs-site.xml and add below property in the configuration.



5). Hadoop-env.cmd

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

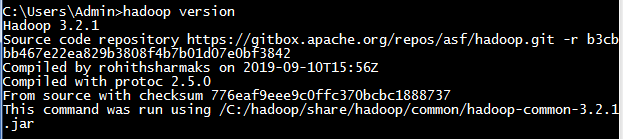
Then, Hadoop needs windows OS specific files which does not come with default download of hadoop.

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

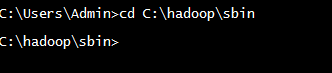
<https://github.com/s911415/apache-hadoop-3.1.0-winutils>

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-



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



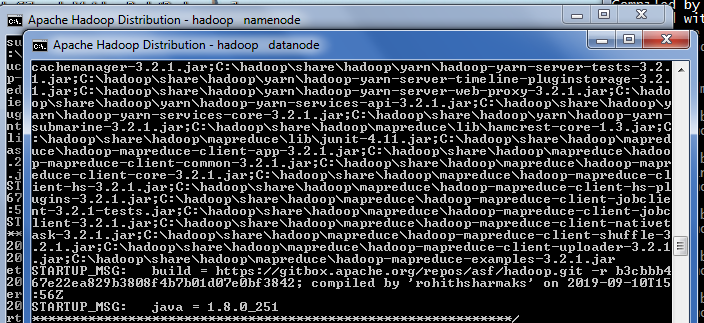
Start namenode and datanode with this command –

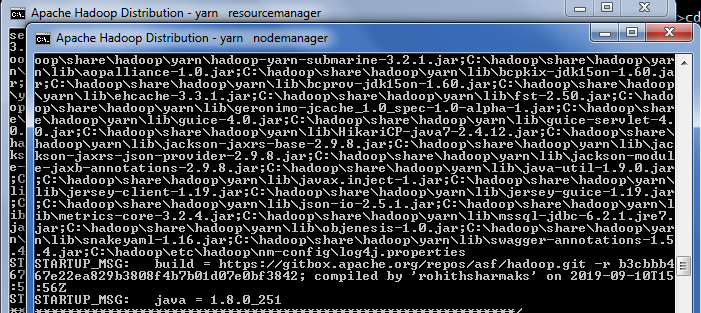
C:\Users\Admin\Desktop\13.PNG

Now start yarn through this command-

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

Make sure all the 4 Apche Hadoop Distribution windows are up n running.



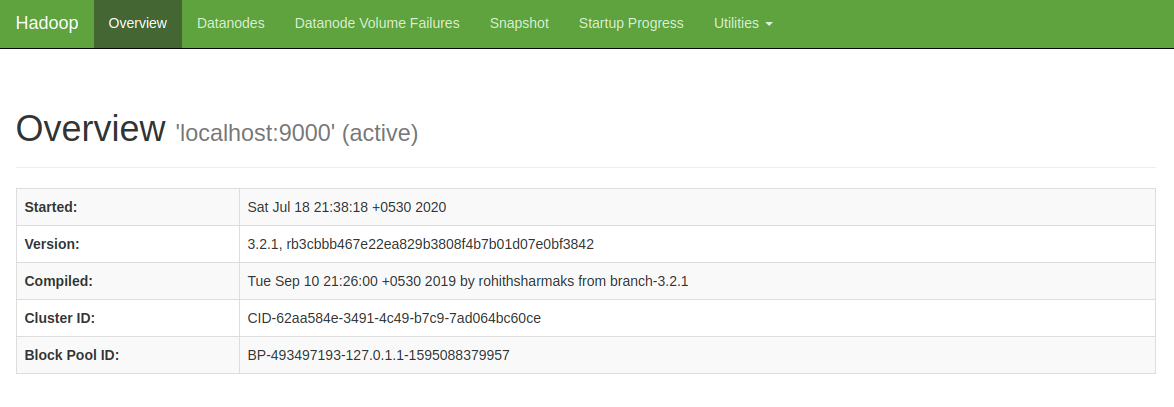


To access information about resource manager current jobs, successful and failed jobs, go this link in browser-

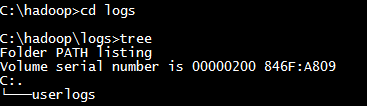
<http://localhost:8088/cluster>

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

<http://localhost:9870/>



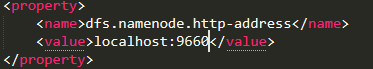
Task 2: Locate the log directory and refer for troubleshooting.

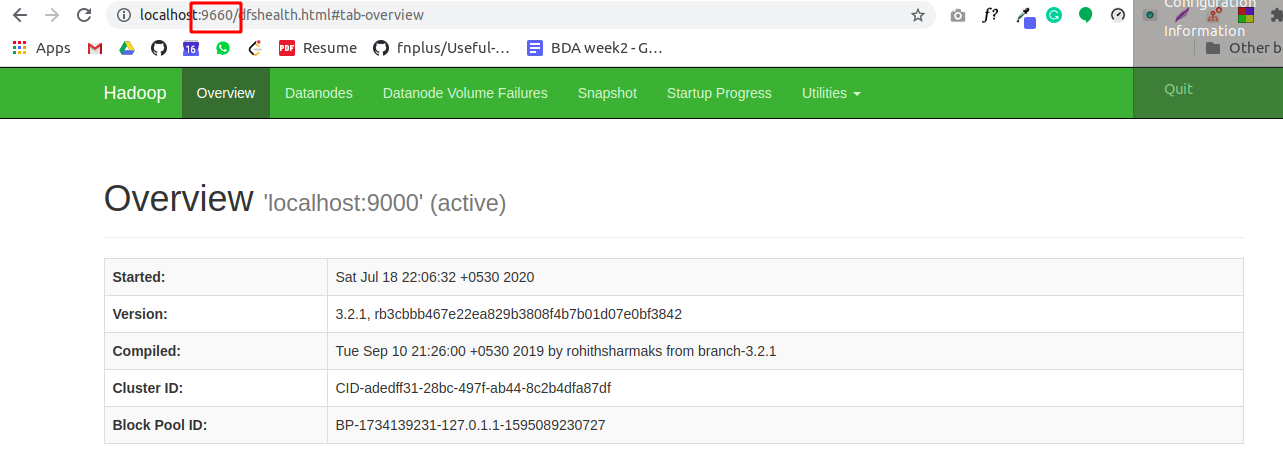


Task 3: Change the URL of HDFS Web UI to localhost:51234 instead of localhost:50070 and the access the web UI using new URL.

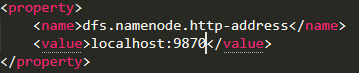
For hadoop 3.2.1 , web Ui is by default localhost:9870.

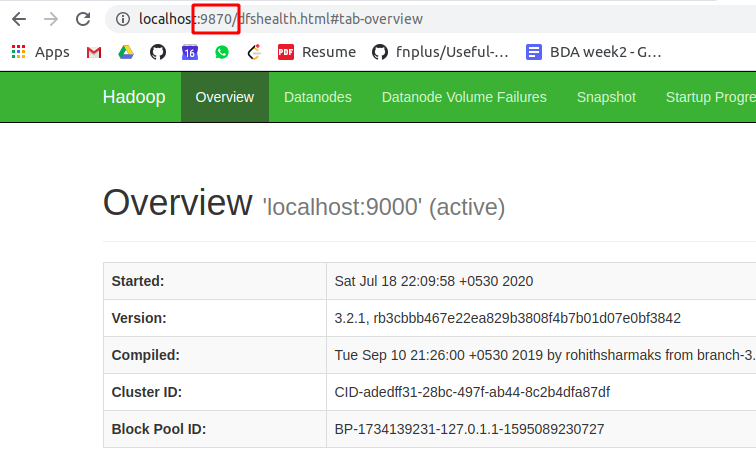
So for this task ,I have changed it to localhost:9660.



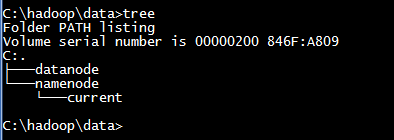


Task 4: Undo changes done in task1.

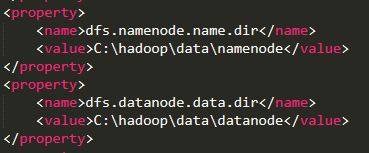


****

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



Also find location of namenode and datanode using hdfs-site.xml file.



Task 6: Analyze the output of following command:

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

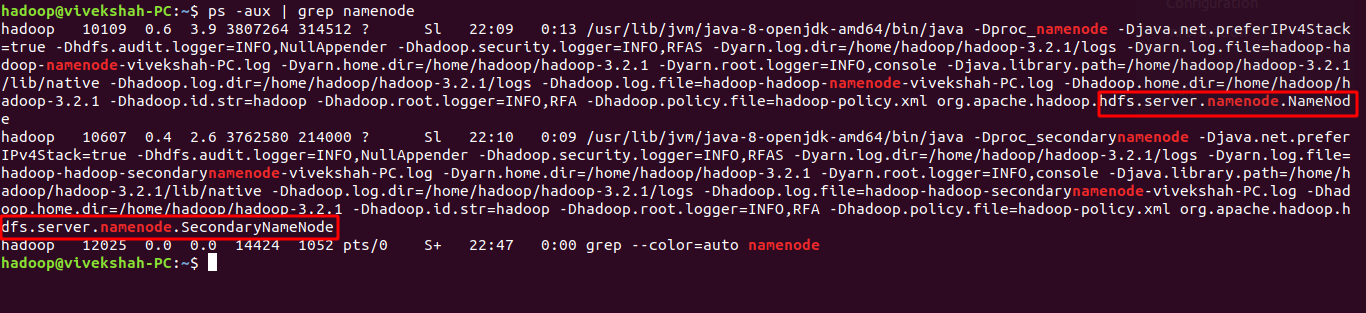
[hadoop@hadoop-clone hadoop]$ ps -ef |grep DataNode

As we know 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 command show the result about which process related to Namenode and Datanode respectively.

1). $ ps -ef |grep NameNode

****

2). $ ps -ef |grep DataNode

