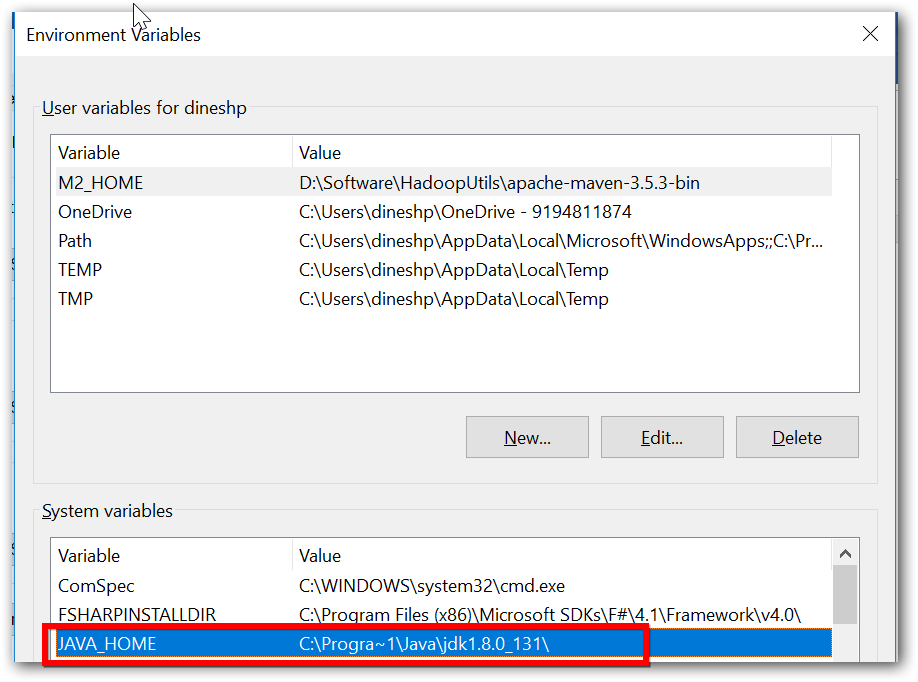
Hands on

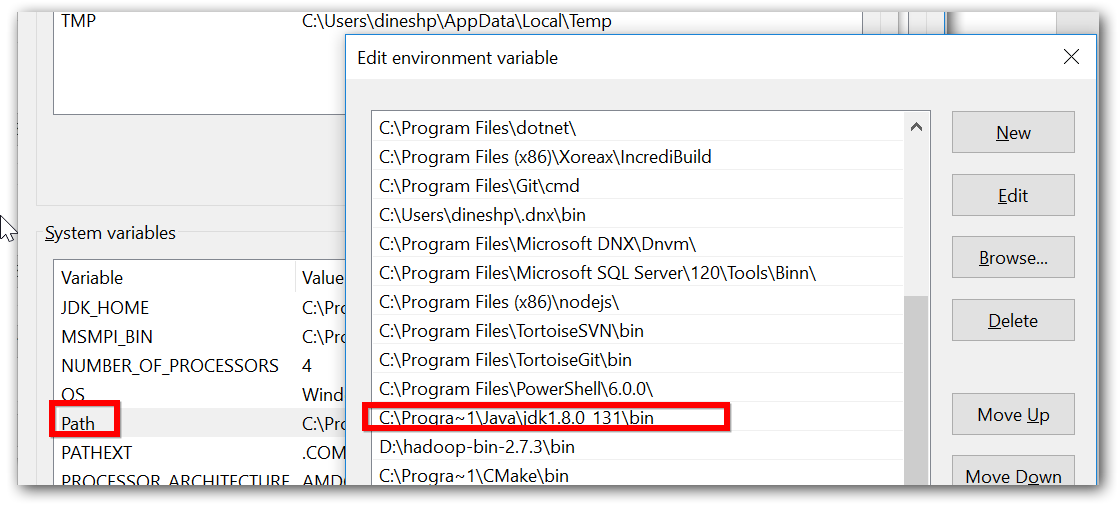
# Pre-requisites

1. Install **7z1801-x64** in your machine.
2. Untar **“hadoop-3.0.1.tar”** using 7zip software installed above.
   1. Try to have the un-compressed folder in any other drive apart from main drive. Like if ‘C:\’ is your main drive, have Hadoop package in ‘D:\’ or some other drive.
   2. E.g. Have it simple like ‘D:\hadoop-3.0.1\bin’
3. Install Java 8 in your machine.
4. Set JAVA\_HOME in environment variables – system variables

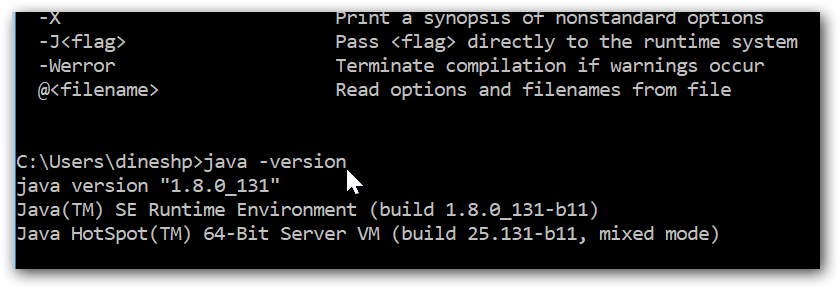
<https://confluence.atlassian.com/doc/setting-the-java_home-variable-in-windows-8895.html>



1. Add Java’s bin location PATH



1. Ensure java home and path is set properly.
   1. Open command prompt
   2. Execute ‘javac’ or ‘java -version’ command.



# Installation

Do the following changes in file specified in directory “..\..\hadoop-3.0.1\etc\hadoop”

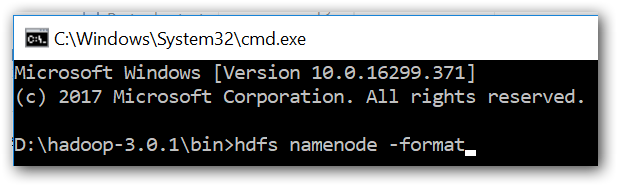
|  |  |  |
| --- | --- | --- |
| File name | Properties | Comments |
| hadoop-env.cmd | set HADOOP\_PREFIX=D:\hadoop-3.0.1  set HADOOP\_CONF\_DIR=%HADOOP\_PREFIX%\etc\hadoop  set YARN\_CONF\_DIR=%HADOOP\_CONF\_DIR%  set PATH=%PATH%;%HADOOP\_PREFIX%\bin | HADOOP\_PREFIX is the Hadoop package unzipped. |
| core-site.xml | <configuration>  <property>  <name>fs.defaultFS</name>  <value>hdfs://localhost:9000</value>  </property>  </configuration> |  |
| hdfs-site.xml | <configuration>  <property>  <name>dfs.replication</name>  <value>1</value>  </property>  <property>  <name>dfs.namenode.name.dir</name>  <value>file:///d:/Metadata/NameNode</value>  </property>  <property>  <name>dfs.datanode.data.dir</name>  <value>file:///d:/Metadata/DataNode</value>  </property>  </configuration> | Specify the directory where there won’t be any permission issues. |
| mapred-site.xml | <configuration>  <property>  <name>mapreduce.framework.name</name>  <value>local</value>  </property>  <property>  <name>yarn.app.mapreduce.am.env</name>  <value>HADOOP\_MAPRED\_HOME=/hadoop-3.0.1</value>  </property>  <property>  <name>mapreduce.map.env</name>  <value>HADOOP\_MAPRED\_HOME=/hadoop-3.0.1</value>  </property>  <property>  <name>mapreduce.reduce.env</name>  <value>HADOOP\_MAPRED\_HOME=/hadoop-3.0.1</value>  </property> </configuration> | Specify the directory like in Linux format itself. |
| yarn-site.xml | <configuration>  <property>  <name>yarn.nodemanager.aux-services</name>  <value>mapreduce\_shuffle</value>  </property>  <property>  <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>  <value>org.apache.hadoop.mapred.ShuffleHandler</value>  </property>  <property>  <description>NM Webapp address.</description>  <name>yarn.nodemanager.webapp.address</name>  <value>${yarn.nodemanager.hostname}:8042</value>  </property>  <property>  <description>hostname of the machine</description>  <name>yarn.nodemanager.hostname</name>  <value>{yourhostname}</value>  </property>  </configuration> | Replace with your machine’s hostname in the property value  {yourhostname} |

# Format your cluster’s HDFS file system to make it use

1. Open command prompt as ‘Administrator’ and navigate to Hadoop bin path,

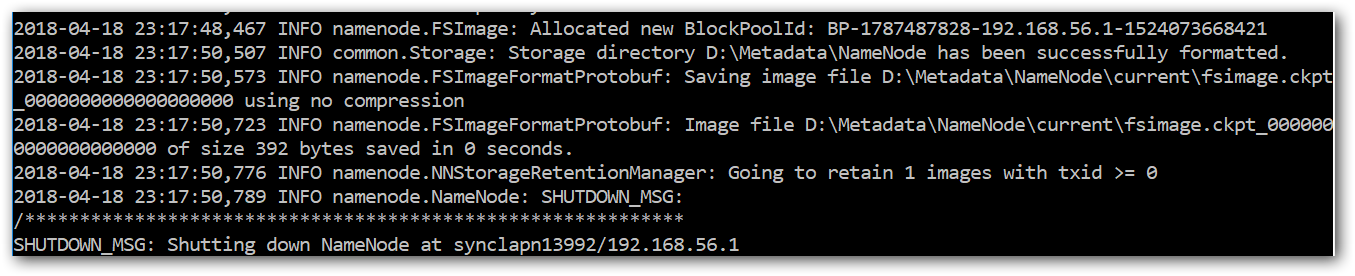
**Command:**

\bin>hdfs namenode -format



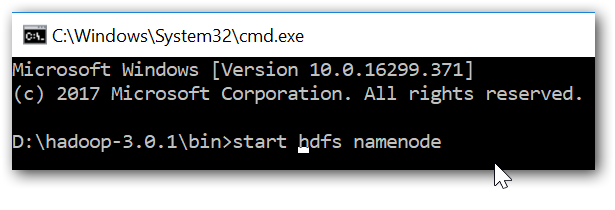
**Result:**

Ensure that format got successfully completed with similar message in below image.

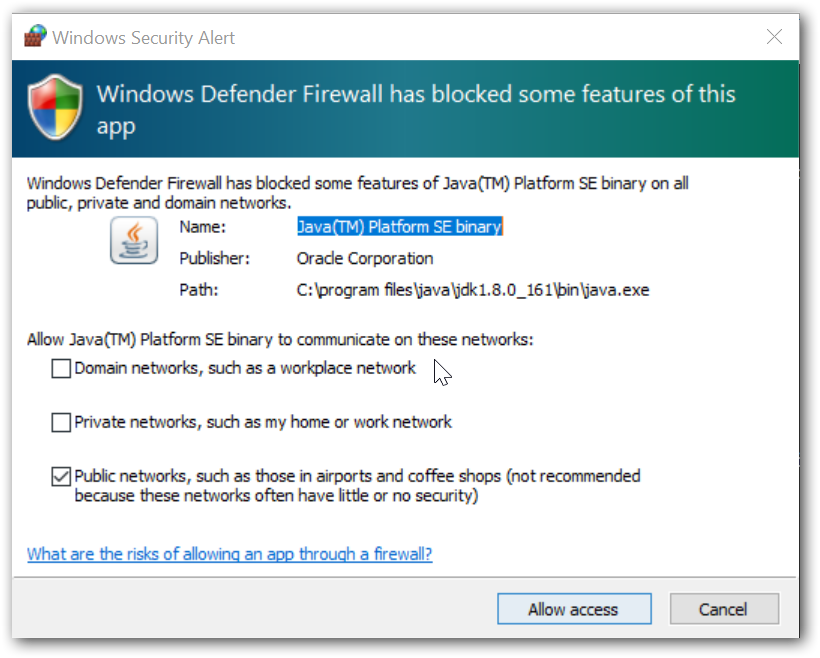


# Start Hadoop services

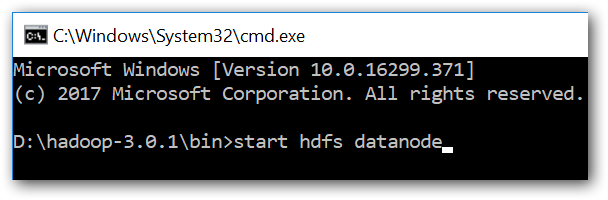
1. Open command prompt as ‘Administrator’ and navigate to Hadoop bin path,
2. Start NameNode



Note: If such a popup is showed at any of below service, enter ‘Yes’



1. Start DataNode



# Work-out

1. Create a file with the following content and do SaveAs with filename ‘Trip.csv’

|  |
| --- |
| vendorid,passenger\_count,trip\_distance,payment\_type,fare\_amount,tip\_amount,total\_amount  2,1,1.55,2,7.5,0,8.8  1,3,1.2,2,7,0,8.3  1,1,2.4,2,10.5,0,11.8  1,1,7.3,2,21.5,0,22.8  1,1,0.4,2,3.5,0,4.8  2,1,0.71,2,5,0,6.3  1,1,1.6,1,7.5,1.2,10  2,1,0.52,1,4,0.9,6.2  2,3,6.23,1,18.5,3.8,23.6 |

1. Create a file with following content and do SaveAs with filename ‘Vendor.csv’

|  |
| --- |
| vendorid,name  1,nyc  2,noc |

1. Create a file with following content and do SaveAs with filename ‘Payment.csv’

|  |
| --- |
| payment\_type,name  1,cash  2,creditcard  3,debitcard |

1. Open HDFS bin location in command prompt,
2. Create a folder called ‘Data’,
   1. Command: hdfs dfs -mkdir /Data
   2. Result:
3. Upload data into HDFS from local file system,
   1. Command: hdfs dfs -put C:/Trip.csv /Data
   2. Result:
4. Word count,

Note: Replace the jar file location properly with your Hadoop installed path

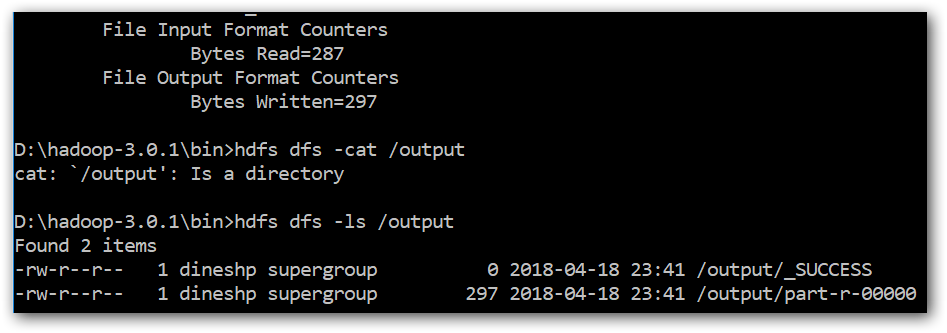
* 1. Upload any content file to know the number of times each word is present.

bin> hdfs dfs -put C:\Users\dineshp\Desktop\Journal.txt /WordCountData

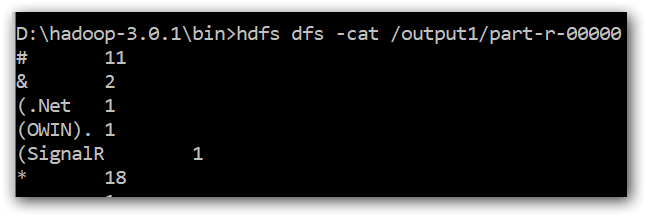
* 1. Execute the in-built sample,

bin> yarn jar "D:\hadoop-3.0.1\share\hadoop\mapreduce\hadoop-mapreduce-examples-3.0.1.jar" wordcount / WordCountData /output1

* 1. Result:



* 1. Check the result stored in file /output1/part-r-00000
  2. Command to list the “/output1” directory: hdfs dfs -ls /output1
  3. Command to check the output of job: hdfs dfs -cat /output1/part-r-00000



1. Copy data between clusters,
   1. Command: hadoop distcp /Data /Data-Copy
   2. Result:
2. Web UI - HDFS - <http://localhost:9870/dfshealth.html#tab-overview>