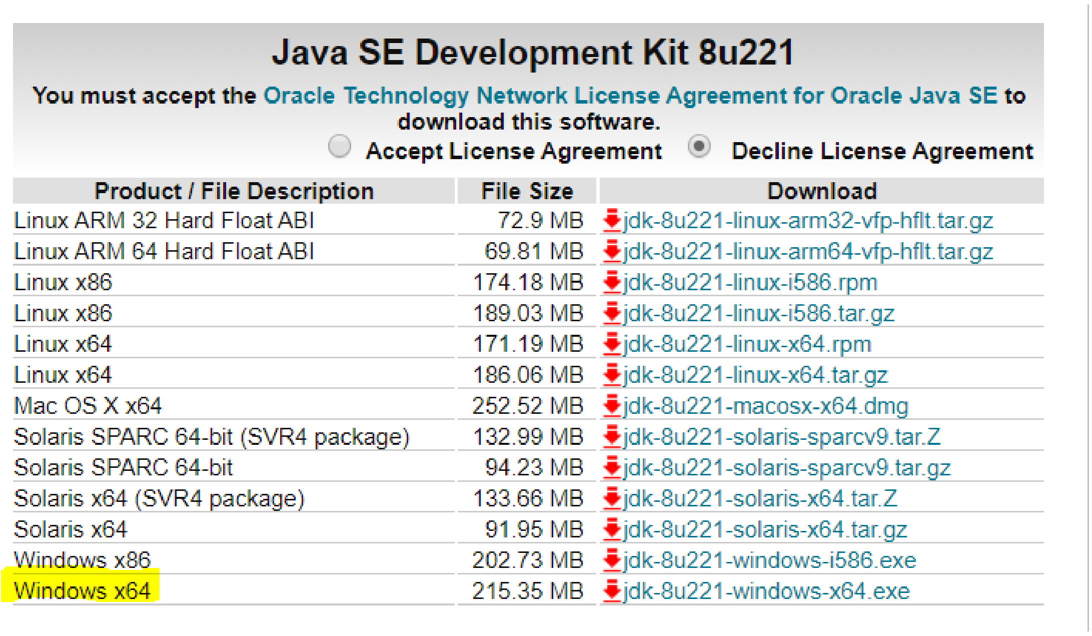
INSTALL HADOOP

1. To install Hadoop, Java JDK 1.8.0 is needed. It can be downloaded at http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html. (need to sign up for an Oracle account)
2. However, download the 64-bit Java version to prevent error with Java when using Hive later.



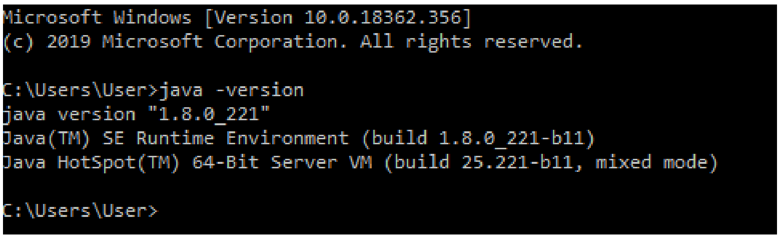
Java SE Development Kit 8u221

You must accept the Oracle Technology Network License Agreement for Oracle Java SE to download this software.

○ Accept License Agreement ● Decline License Agreement

Product / File Description	File Size	Download
Linux ARM 32 Hard Float ABI	72.9 MB	jdk-8u221-linux-arm32-vfp-hflt.tar.gz
Linux ARM 64 Hard Float ABI	69.81 MB	jdk-8u221-linux-arm64-vfp-hflt.tar.gz
Linux x86	174.18 MB	jdk-8u221-linux-i586.rpm
Linux x86	189.03 MB	jdk-8u221-linux-i586.tar.gz
Linux x64	171.19 MB	jdk-8u221-linux-x64.rpm
Linux x64	186.06 MB	jdk-8u221-linux-x64.tar.gz
Mac OS X x64	252.52 MB	jdk-8u221-macosx-x64.dmg
Solaris SPARC 64-bit (SVR4 package)	132.99 MB	jdk-8u221-solaris-sparcv9.tar.Z
Solaris SPARC 64-bit	94.23 MB	jdk-8u221-solaris-sparcv9.tar.gz
Solaris x64 (SVR4 package)	133.66 MB	jdk-8u221-solaris-x64.tar.Z
Solaris x64	91.95 MB	jdk-8u221-solaris-x64.tar.gz
Windows x86	202.73 MB	jdk-8u221-windows-i586.exe
Windows x64	215.35 MB	jdk-8u221-windows-x64.exe

3. Install the at’ C:\Java\jdk1.8.0_221\’ instead of ‘C:\Program Files\Java\’ to prevent error in file path later on due to the space between “Program Files”.
4. Verify the java installation by using cmd and type ‘java –version’

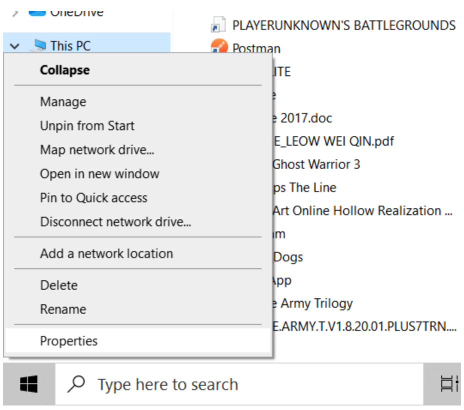


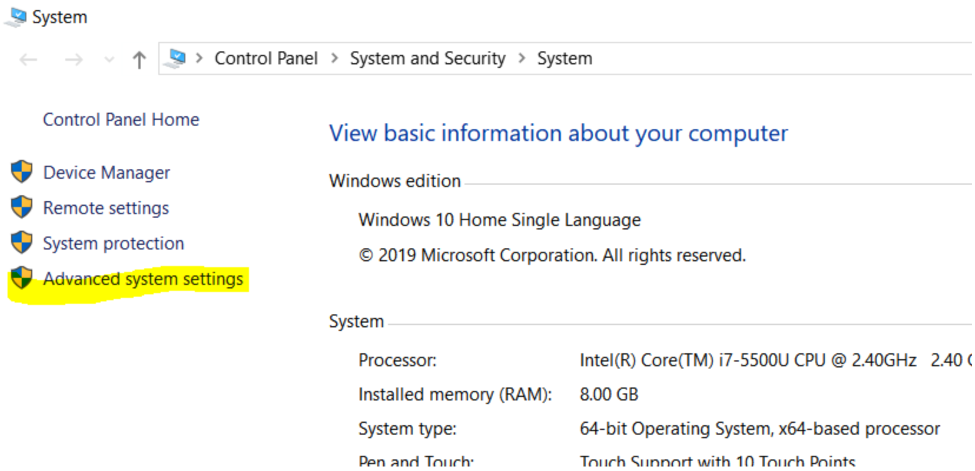
```
Microsoft Windows [Version 10.0.18362.356]
(c) 2019 Microsoft Corporation. All rights reserved.

C:\Users\User>java -version
java version "1.8.0_221"
Java(TM) SE Runtime Environment (build 1.8.0_221-b11)
Java HotSpot(TM) 64-Bit Server VM (build 25.221-b11, mixed mode)

C:\Users\User>
```

5. Download Hadoop (version 2.8.0) is used. The file can be downloaded at http://archive.apache.org/dist/hadoop/core//hadoop-2.8.0/hadoop-2.8.0.tar.gz. The format of the file is tar.gz. Use Git Bash (must run as administrator) to extract the files by typing ‘tar xzvf hadoop-2.8.0.tar.gz’ where xxx is the file name.

hadoop-2.8.0.tar.gz 4/10/2019 8:11 PM WinRAR archive 419,853 KB

6. After extracting, configure the environment variables. This PC - > Right Click - > Properties - > Advanced System Settings - > Advanced - > Environment Variables





View basic information about your computer

Windows edition

Windows 10 Home Single Language

© 2019 Microsoft Corporation. All rights reserved.

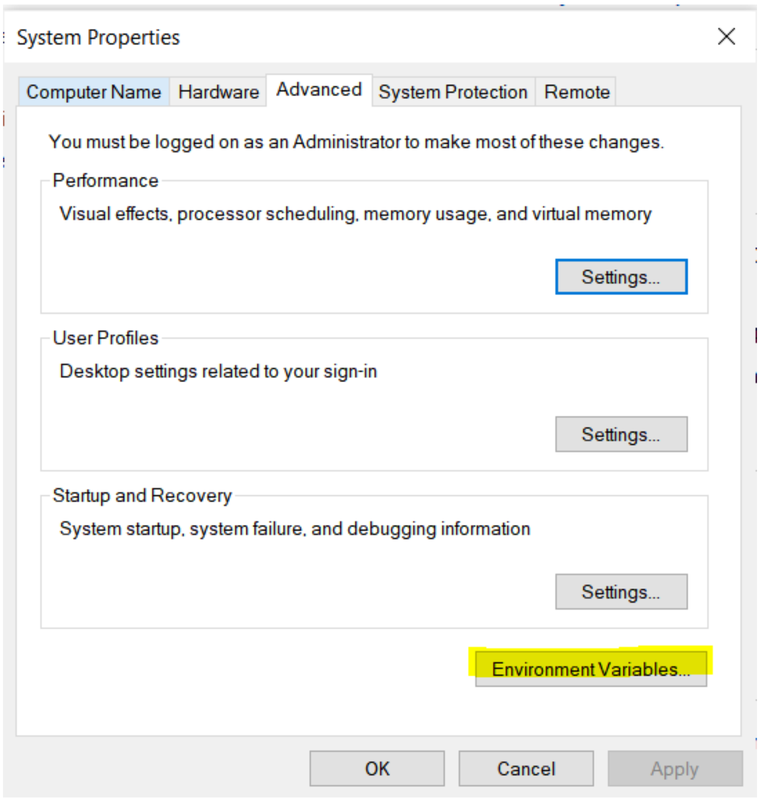
System

Processor: Intel(R) Core(TM) i7-5500U CPU @ 2.40GHz 2.40 (

Installed memory (RAM): 8.00 GB

System type: 64-bit Operating System, x64-based processor

Pen and Touch: Touch Support with 10 Touch Points



You must be logged on as an Administrator to make most of these changes.

Performance

Visual effects, processor scheduling, memory usage, and virtual memory

User Profiles

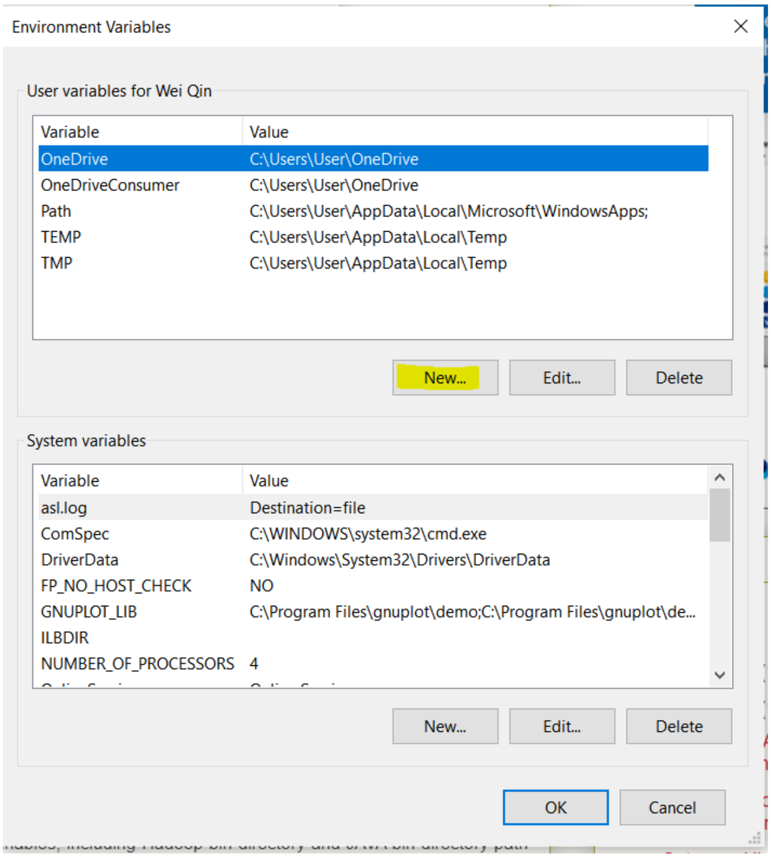
Desktop settings related to your sign-in

Startup and Recovery

System startup, system failure, and debugging information

7. Add two new User Variable:
    a. HADOOP_HOME (path: the directory you extracted the tar.gz file)

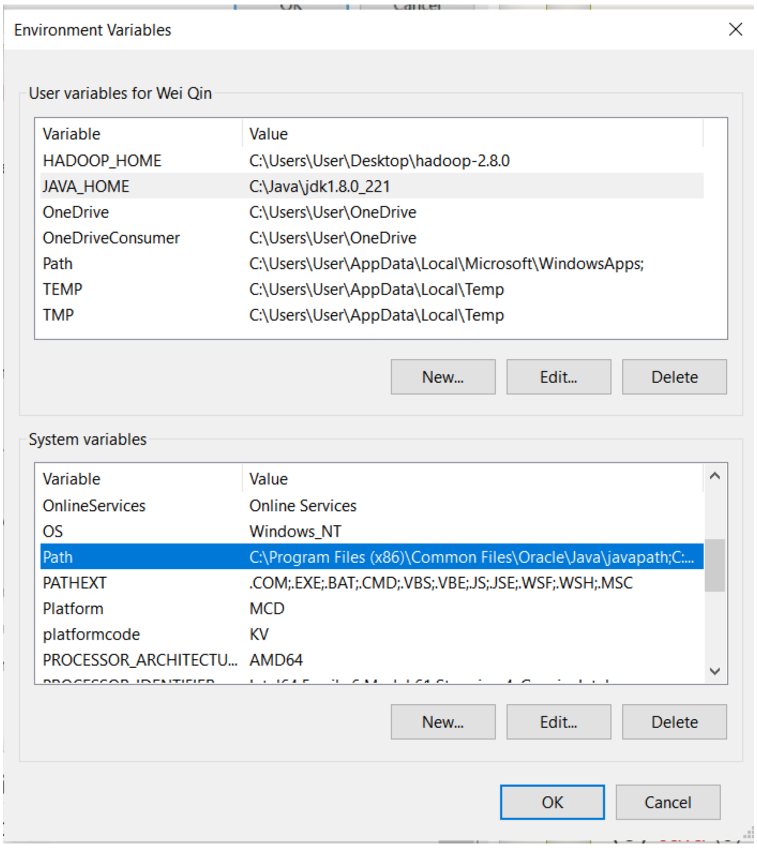
b. JAVA_HOME (path: C:\Java\jdk1.8.0_221)



Variable	Value
OneDrive	C:\Users\User\OneDrive
OneDriveConsumer	C:\Users\User\OneDrive
Path	C:\Users\User\AppData\Local\Microsoft\WindowsApps;
TEMP	C:\Users\User\AppData\Local\Temp
TMP	C:\Users\User\AppData\Local\Temp

Variable	Value
asl.log	Destination=file
ComSpec	C:\WINDOWS\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
FP_NO_HOST_CHECK	NO
GNUPLOT_LIB	C:\Program Files\gnuplot\demo;C:\Program Files\gnuplot\de...
ILBDIR	
NUMBER_OF_PROCESSORS	4

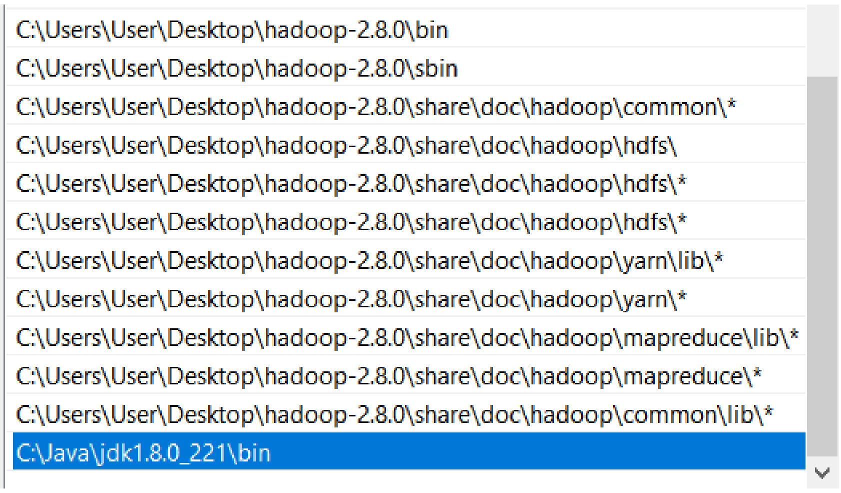
8. Edit the Path variable under System Variable



Variable	Value
HADOOP_HOME	C:\Users\User\Desktop\hadoop-2.8.0
JAVA_HOME	C:\Java\jdk1.8.0_221
OneDrive	C:\Users\User\OneDrive
OneDriveConsumer	C:\Users\User\OneDrive
Path	C:\Users\User\AppData\Local\Microsoft\WindowsApps;
TEMP	C:\Users\User\AppData\Local\Temp
TMP	C:\Users\User\AppData\Local\Temp

Variable	Value
OnlineServices	Online Services
OS	Windows_NT
Path	C:\Program Files (x86)\Common Files\Oracle\Java\javapath;C:...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
Platform	MCD
platformcode	KV
PROCESSOR_ARCHITECTU...	AMD64

9. Add the following path (edit if your directory you extract the tar.gz file is different) and press ok.



```
C:\Users\User\Desktop\hadoop-2.8.0\bin
C:\Users\User\Desktop\hadoop-2.8.0\sbin
C:\Users\User\Desktop\hadoop-2.8.0\share\doc\hadoop\common\*
C:\Users\User\Desktop\hadoop-2.8.0\share\doc\hadoop\hdfs\
C:\Users\User\Desktop\hadoop-2.8.0\share\doc\hadoop\hdfs\*
C:\Users\User\Desktop\hadoop-2.8.0\share\doc\hadoop\hdfs\*
C:\Users\User\Desktop\hadoop-2.8.0\share\doc\hadoop\yarn\lib\*
C:\Users\User\Desktop\hadoop-2.8.0\share\doc\hadoop\yarn\*
C:\Users\User\Desktop\hadoop-2.8.0\share\doc\hadoop\mapreduce\lib\*
C:\Users\User\Desktop\hadoop-2.8.0\share\doc\hadoop\mapreduce\*
C:\Users\User\Desktop\hadoop-2.8.0\share\doc\hadoop\common\lib\*
C:\Java\jdk1.8.0_221\bin
```

10. Create some dedicated folders -
    a. Create folder "data" under "C:\Users\User\Desktop\hadoop-2.8.0".
    b. Create folder "datanode" under “C:\Users\User\Desktop\hadoop-2.8.0\data".
    c. Create folder "namenode" under “C:\Users\User\Desktop\hadoop-2.8.0\data”
    d. Create a folder to store temporary data during execution of a project, such as “C:\Users\User\Desktop\hadoop-2.8.0\temp.”
    e. Create a log folder, such as “C:\Users\User\Desktop\hadoop-2.8.0\userlog”

11. Now need to configure four key files with minimal required details –
    a. core-site.xml
    b. hdfs-site.xml
    c. mapred.xml
    d. yarn.xml

    Edit file C:\Users\User\Desktop\hadoop-2.8.0\etc\hadoop\core-site.xml, paste below xml paragraph and save this file.

```
<configuration>

  <property>
     <name>fs.defaultFS</name>
     <value>hdfs://localhost:9000</value>
  </property>
</configuration>
```

[2] Rename "mapred-site.xml.template" to "mapred-site.xml" and edit this file C:\Users\User\Desktop\hadoop-2.8.0\etc\hadoop\mapred-site.xml, paste below xml paragraph and save this file.

```
<configuration>

  <property>
     <name>mapreduce.framework.name</name>
     <value>yarn</value>
  </property>
</configuration>
```

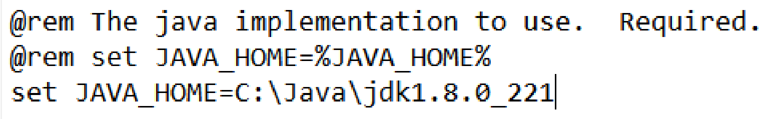
[3] Edit file C:\Users\User\Desktop\hadoop-2.8.0\etc\hadoop\hdfs-site.xml, paste below xml paragraph and save this file.

```
<configuration>

  <property>
     <name>dfs.replication</name>
     <value>1</value>
  </property>
  <property>
     <name>dfs.namenode.name.dir</name>
     <value>/C:/Users/User/Desktop/hadoop-2.8.0/data/namenode</value>
  </property>
  <property>
     <name>dfs.datanode.data.dir</name>
     <value>/C:/Users/User/Desktop/hadoop-2.8.0/data/datanode</value>
  </property>
</configuration>
```

[4] Edit file C:\Users\User\Desktop\hadoop-2.8.0\etc\hadoop\yarn-site.xml, paste below xml paragraph and save this file.

```
<configuration>

  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>
  <property>
     <name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
 <value>org.apache.hadoop.mapred.ShuffleHandler</value>
  </property>
  <property>
  <name>yarn.nodemanager.log-dirs</name>
  <value>/ C:/Users/User/Desktop/hadoop-2.8.0/userlog</value><final>true</final>
 </property>
  <property><name>yarn.nodemanager.local-dirs</name>
 <value>/ C:/Users/User/Desktop/hadoop-2.8.0/temp/nm-localdir</value>
 </property>
</configuration>
```

[5] Edit file D:/Hadoop/hadoop-2.8.0/etc/hadoop/hadoop-env.cmd as below:

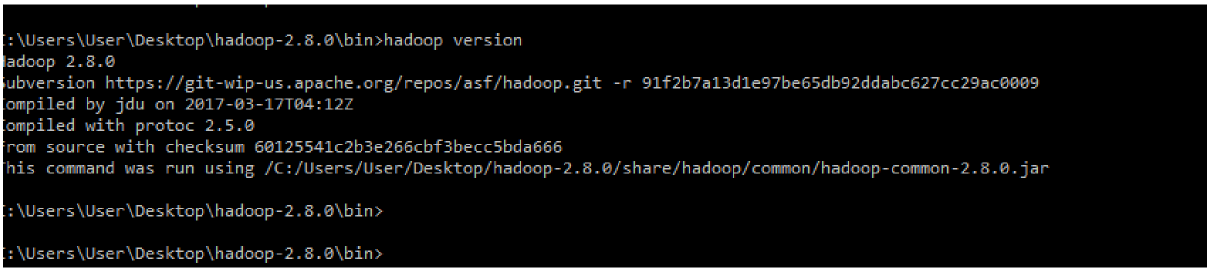


```
@rem The java implementation to use.  Required.
@rem set JAVA_HOME=%JAVA_HOME%
set JAVA_HOME=C:\Java\jdk1.8.0_221
```

12. Download Hadoop Configuration Zip from https://github.com/MuhammadBilalYar/HADOOP-INSTALLATION-ON-WINDOW-10/blob/master/Hadoop%20Configuration.zip

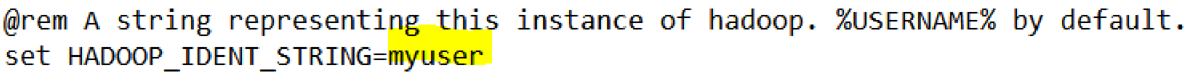
13. Delete file bin on C:\Users\User\Desktop\hadoop-2.8.0\bin, replaced by file bin on file just download (from Hadoop Configuration.zip).

14. At the cmd prompt, cd into the hadoop directory and type ‘hadoop version’



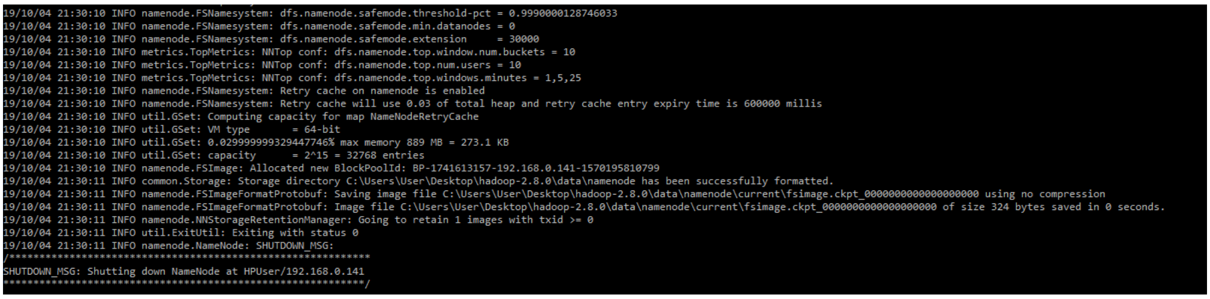
```
:\Users\User\Desktop\hadoop-2.8.0\bin>hadoop version
adoop 2.8.0
ubversion https://git-wip-us.apache.org/repos/asf/hadoop.git -r 91f2b7a13d1e97be65db92ddabc627cc29ac0009
ompiled by jdu on 2017-03-17T04:12Z
ompiled with protoc 2.5.0
rom source with checksum 60125541c2b3e266cbf3becc5bda666
his command was run using /C:/Users/User/Desktop/hadoop-2.8.0/share/hadoop/common/hadoop-common-2.8.0.jar

:\Users\User\Desktop\hadoop-2.8.0\bin>

:\Users\User\Desktop\hadoop-2.8.0\bin>
```

If you have the following error: Error: Could not find or load main class M
edit the D:/Hadoop/hadoop-2.8.0/etc/hadoop/hadoop-env.cmd by changing %username% to anything string without space e.g. myuser



```
@rem A string representing this instance of hadoop. %USERNAME% by default.
set HADOOP_IDENT_STRING=myuser
```

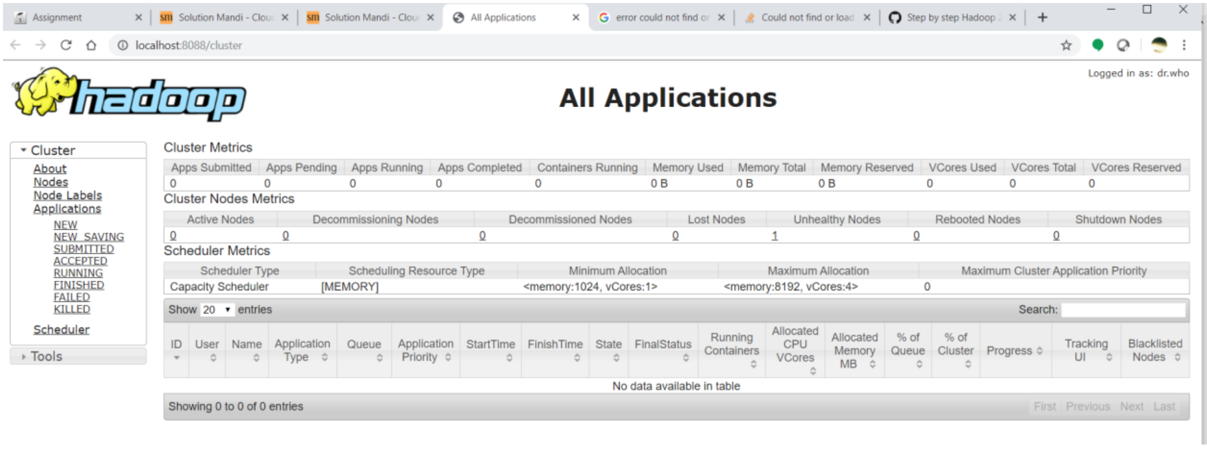
15. Execute the namenode by typing ‘hdfs namenode –format’ Make sure it ended with status 0. If not, try to read the log to see where the error is coming from.



```
19/10/04 21:30:10 INFO namenode.FSNamesystem: dfs.namenode.safemode.threshold-pct = 0.9990000128746033
19/10/04 21:30:10 INFO namenode.FSNamesystem: dfs.namenode.safemode.min.datanodes = 0
19/10/04 21:30:10 INFO namenode.FSNamesystem: dfs.namenode.safemode.extension     = 30000
19/10/04 21:30:10 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
19/10/04 21:30:10 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
19/10/04 21:30:10 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
19/10/04 21:30:10 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
19/10/04 21:30:10 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry time is 600000 millis
19/10/04 21:30:10 INFO util.GSet: Computing capacity for map NameNodeRetryCache
19/10/04 21:30:10 INFO util.GSet: VM type       = 64-bit
19/10/04 21:30:10 INFO util.GSet: 0.029999999329447746% max memory 889 MB = 273.1 KB
19/10/04 21:30:10 INFO util.GSet: capacity      = 2^15 = 32768 entries
19/10/04 21:30:10 INFO namenode.FSImage: Allocated new BlockPoolId: BP-1741613157-192.168.0.141-1570195810799
19/10/04 21:30:11 INFO common.Storage: Storage directory C:\Users\User\Desktop\hadoop-2.8.0\data\namenode has been successfully formatted.
19/10/04 21:30:11 INFO namenode.FSImageFormatProtobuf: Saving image file C:\Users\User\Desktop\hadoop-2.8.0\data\namenode\current\fsimage.ckpt_0000000000000000000 using no compression
19/10/04 21:30:11 INFO namenode.FSImageFormatProtobuf: Image file C:\Users\User\Desktop\hadoop-2.8.0\data\namenode\current\fsimage.ckpt_0000000000000000000 of size 324 bytes saved in 0 seconds.
19/10/04 21:30:11 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
19/10/04 21:30:11 INFO util.ExitUtil: Exiting with status 0
19/10/04 21:30:11 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at HPUser/192.168.0.141
************************************************************/
```

16. Cd to C:\Users\User\Desktop\hadoop-2.8.0\sbin and start hadoop by typing ‘start-all.cmd’

17. Hadoop can be verified via browser also as –
    - Namenode (hdfs) - http://localhost:50070
    - Datanode - http://localhost:50075
    - All Applications (cluster) - http://localhost:8088 etc.



Logged in as: dr.who

All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved
0	0	0	0	0	0 B	0 B	0 B	0	0	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
0	0	0	0	1	0	0

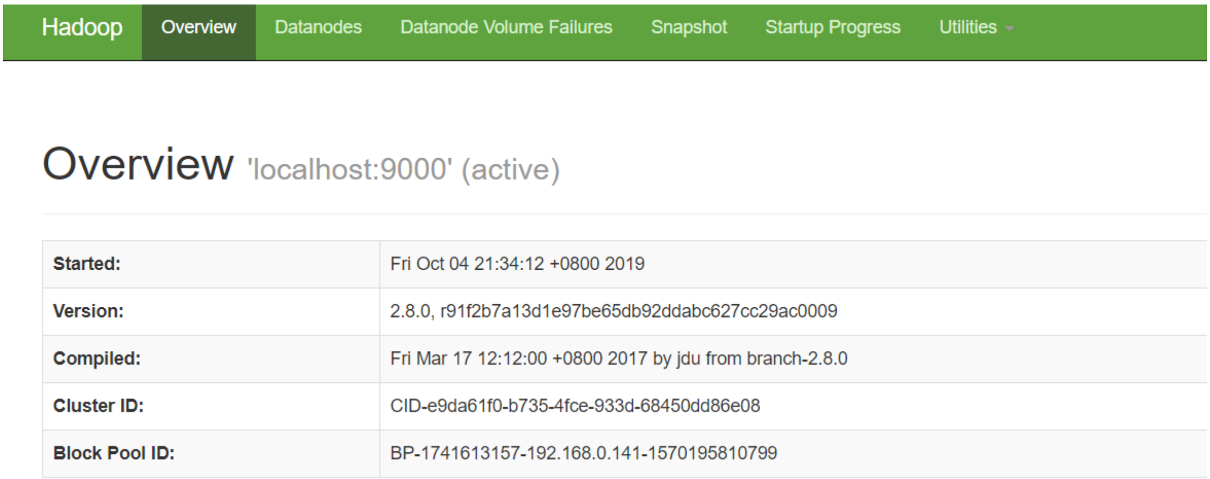
Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority
Capacity Scheduler	[MEMORY]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0

Show 20 entries Search:

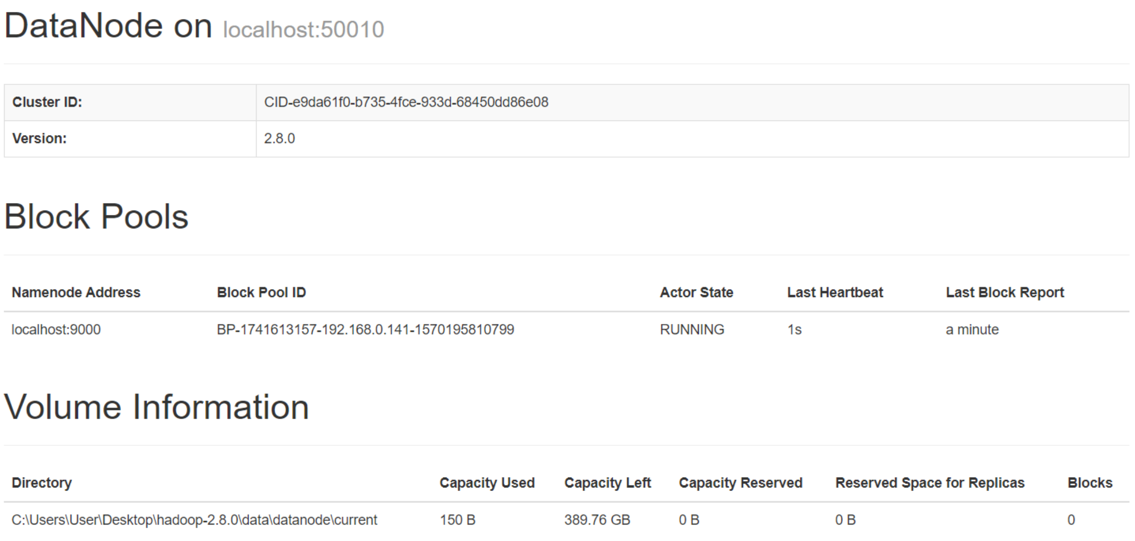
ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes
No data available in table																	

Showing 0 to 0 of 0 entries



Overview 'localhost:9000' (active)

Started:	Fri Oct 04 21:34:12 +0800 2019
Version:	2.8.0, r91f2b7a13d1e97be65db92ddabc627cc29ac0009
Compiled:	Fri Mar 17 12:12:00 +0800 2017 by jdu from branch-2.8.0
Cluster ID:	CID-e9da61f0-b735-4fce-933d-68450dd86e08
Block Pool ID:	BP-1741613157-192.168.0.141-1570195810799



DataNode on localhost:50010

Cluster ID:	CID-e9da61f0-b735-4fce-933d-68450dd86e08
Version:	2.8.0

Block Pools

Namenode Address	Block Pool ID	Actor State	Last Heartbeat	Last Block Report
localhost:9000	BP-1741613157-192.168.0.141-1570195810799	RUNNING	1s	a minute

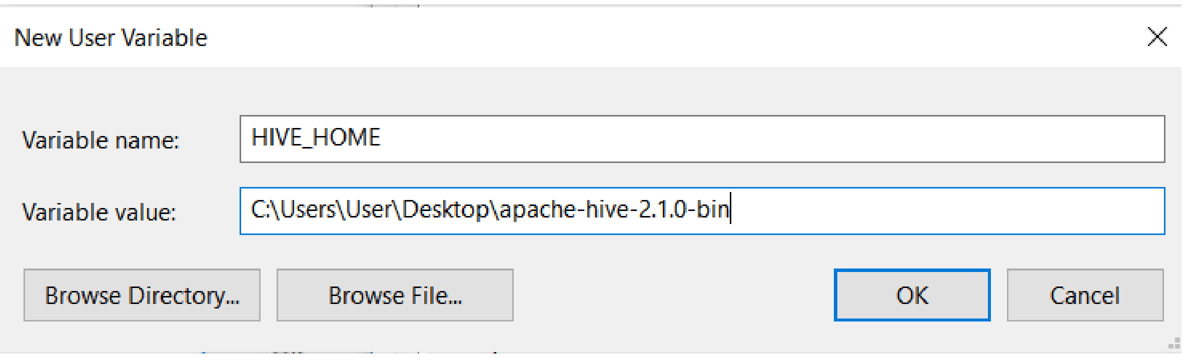
Volume Information

Directory	Capacity Used	Capacity Left	Capacity Reserved	Reserved Space for Replicas	Blocks
C:\Users\User\Desktop\hadoop-2.8.0\data\datanode\current	150 B	389.76 GB	0 B	0 B	0

Reference: https://www.solutionmandi.com/2018/11/hadoop-installation-on-windows-10.html

INSTALL Hive

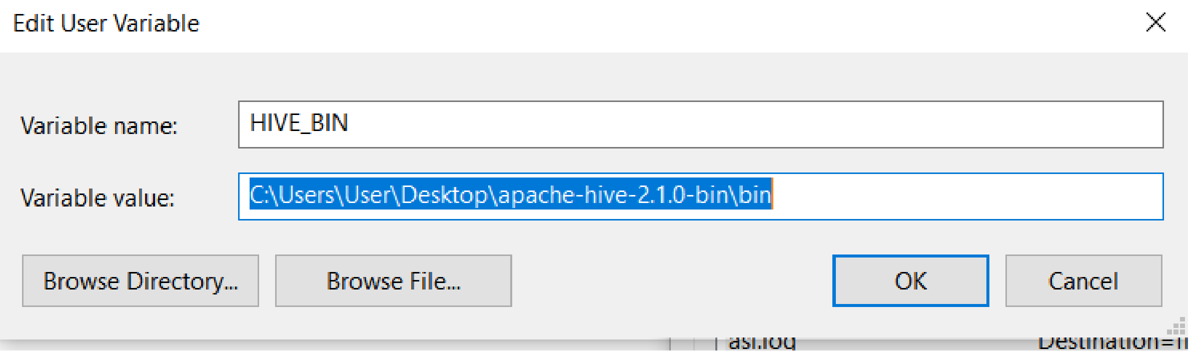
1. Download Hive from https://archive.apache.org/dist/hive/hive-2.1.0/ (version 2.1.0 is used)
2. Download Derby from https://archive.apache.org/dist/db/derby/db-derby-10.12.1.1/ (version 10.12.1.1 is used)
3. Same as Hadoop, extract using git bash as explained.
4. Download hive-site.xml from https://drive.google.com/file/d/1qqAo7RQfr5Q6O-GTom6Rji3TdufP81zd/view?usp=sharing this will be used to define the metastore to Derby.
5. Drop the downloaded file “hive-site.xml” to hive configuration location “C:\Users\User\Desktop\apache-hive-2.1.0-bin\conf”
6. Go to C:\Users\User\Desktop\db-derby-10.12.1.1-bin\lib and copy every files inside and paste it in C:\Users\User\Desktop\apache-hive-2.1.0-bin\lib
7. This PC - > Right Click - > Properties - > Advanced System Settings - > Advanced - > Environment Variables
8. Create new user variable.



New User Variable

Variable name: HIVE_HOME

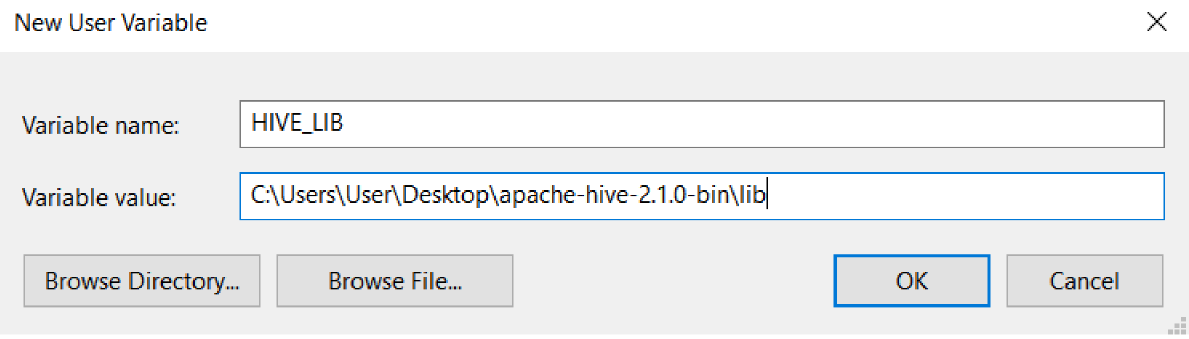
Variable value: C:\Users\User\Desktop\apache-hive-2.1.0-bin



Edit User Variable

Variable name: HIVE_BIN

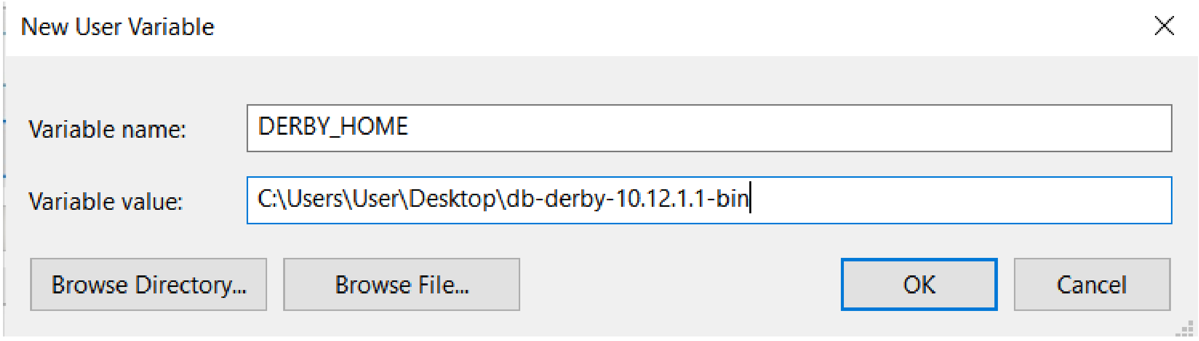
Variable value: C:\Users\User\Desktop\apache-hive-2.1.0-bin\bin



New User Variable

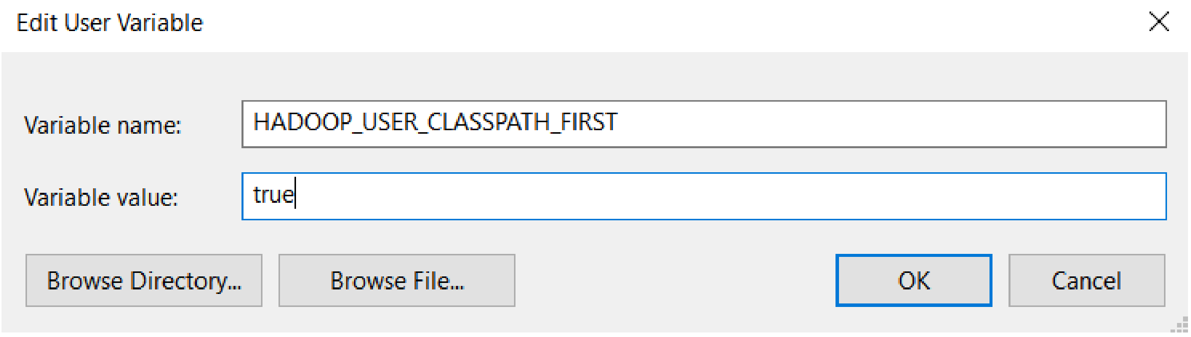
Variable name: HIVE_LIB

Variable value: C:\Users\User\Desktop\apache-hive-2.1.0-bin\lib


New User Variable

Variable name: DERBY_HOME

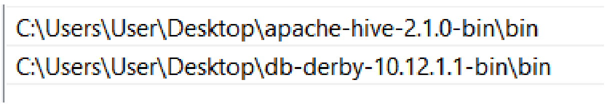
Variable value: C:\Users\User\Desktop\db-derby-10.12.1.1-bin


Edit User Variable

Variable name: HADOOP_USER_CLASSPATH_FIRST

Variable value: true

9. At system variable, edit path, and add 2 path below:

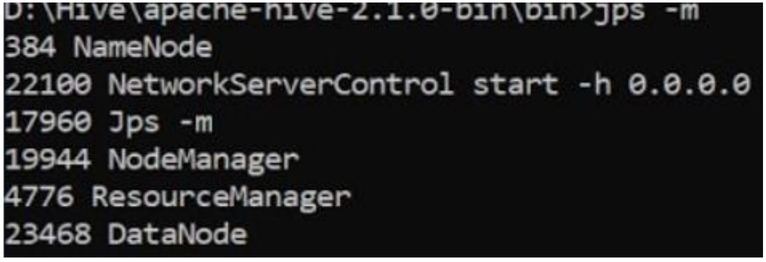

C:\Users\User\Desktop\apache-hive-2.1.0-bin\bin
C:\Users\User\Desktop\db-derby-10.12.1.1-bin\bin

10. Edit C:\Users\User\Desktop\apache-hive-2.1.0-bin\conf/hive-site.xml, paste below xml paragraph and save this file. This is the metastore configuration.

```
<configuration>
 <property>
  <name>javax.jdo.option.ConnectionURL</name>
  <value>jdbc:derby://localhost:1527/metastore_db;create=true</value>
  <description>JDBC connect string for a JDBC metastore</description>
 </property>
 <property>
  <name>javax.jdo.option.ConnectionDriverName</name>
  <value>org.apache.derby.jdbc.ClientDriver</value>
  <description>Driver class name for a JDBC metastore</description>
 </property>
 <property>
  <name>hive.server2.enable.impersonation</name>
  <description>Enable user impersonation for HiveServer2</description>
  <value>true</value>
 </property>
 <property>
  <name>hive.server2.authentication</name>
  <value>NONE</value>
  <description> Client authentication types. NONE: no authentication check LDAP: LDAP/AD
based authentication KERBEROS: Kerberos/GSSAPI authentication CUSTOM: Custom
authentication provider (Use with property hive.server2.custom.authentication.class)
</description>
```

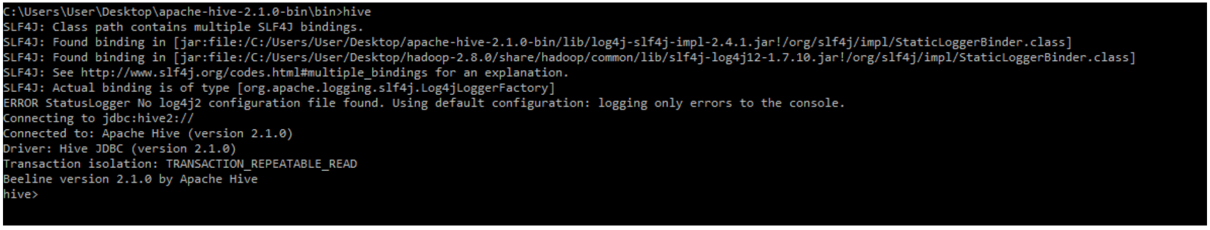
```
 </property>
 <property>
  <name>datanucleus.autoCreateTables</name>
  <value>True</value>
 </property>
</configuration>
```

18. Start hadoop first, at cmd Cd to C:\Users\User\Desktop\hadoop-2.8.0\sbin and start hadoop by typing ‘start-all.cmd’
11. Then start Derby server, at cmd cd to C:\Users\User\Desktop\db-derby-10.12.1.1-bin\bin and type “startNetworkServer -h 0.0.0.0” Derby server will be started.
12. Open a new cmd, cd into C:\Users\User\Desktop\apache-hive-2.1.0-bin\bin and type “jps –m’ to check Network Server Control.



```
D:\Hive\apache-hive-2.1.0-bin\bin>jps -m
384 NameNode
22100 NetworkServerControl start -h 0.0.0.0
17960 Jps -m
19944 NodeManager
4776 ResourceManager
23468 DataNode
```

13. Then, run hive by typing ‘hive’.



```
C:\Users\User\Desktop\apache-hive-2.1.0-bin\bin>hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/C:/Users/User/Desktop/apache-hive-2.1.0-bin/lib/log4j-slf4j-impl-2.4.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/C:/Users/User/Desktop/hadoop-2.8.0/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
ERROR StatusLogger No log4j2 configuration file found. Using default configuration: logging only errors to the console.
Connecting to jdbc:hive2://
Connected to: Apache Hive (version 2.1.0)
Driver: Hive JDBC (version 2.1.0)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 2.1.0 by Apache Hive
hive>
```

14. Congrats!

Reference: https://www.solutionmandi.com/2018/11/hive-installation-on-windows-10.html