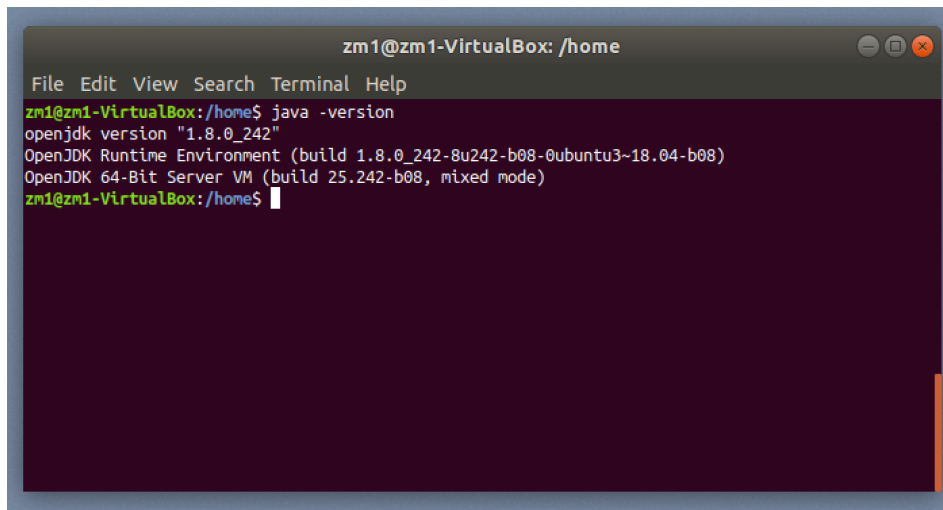


MILESTONE 2 - STORE DATA INTO HIVE DATA WAREHOUSE

STEP 1 : INSTALL APACHE HADOOP IN UBUNTU

1.1 Pre-requisite

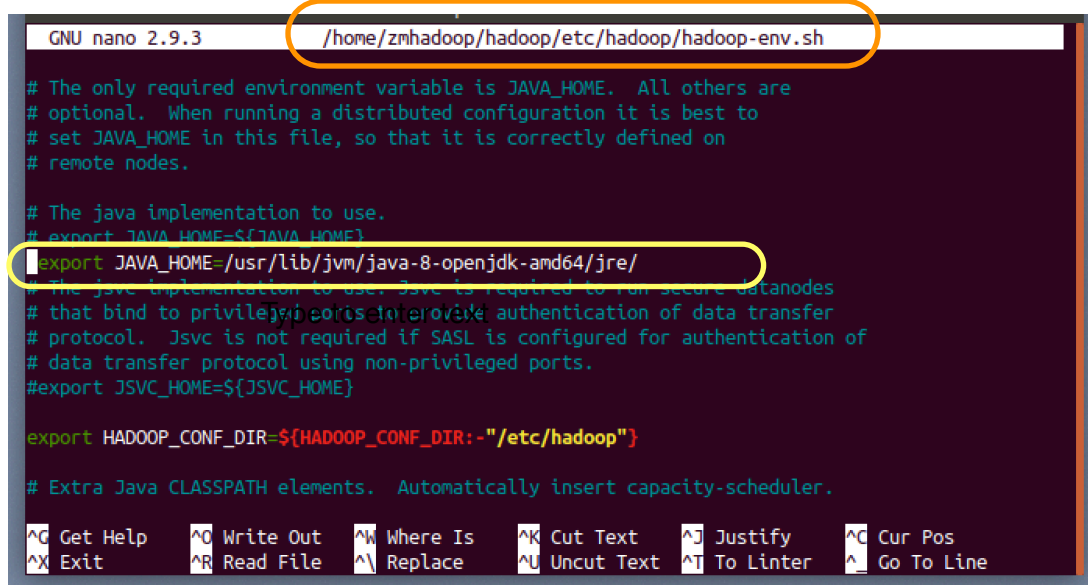
- **sudo apt-get update** - Update latest distribution software
- **sudo apt-get install openssh-server** - SSH is mainly used so that the master node can stay connected with the slave nodes.
- **sudo apt-get install openjdk-8-jdk** - hadoop require java to run
- after install check the java version, type **java -version**.



```
zm1@zm1-VirtualBox: /home
File Edit View Search Terminal Help
zm1@zm1-VirtualBox:/home$ java -version
openjdk version "1.8.0_242"
OpenJDK Runtime Environment (build 1.8.0_242-8u242-b08-0ubuntu3-18.04-b08)
OpenJDK 64-Bit Server VM (build 25.242-b08, mixed mode)
zm1@zm1-VirtualBox:/home$
```

1.2 Download and Install Hadoop

- to download type : **wget https://www-eu.apache.org/dist/hadoop/common/hadoop-2.7.7/hadoop-2.7.7.tar.gz**
- to unzip & install type : **tar -xzf hadoop-2.7.7.tar.gz**
- make sure the hadoop is inside your home directory : **/home/{yourname}/hadoop/**
- set JAVA_HOME for hadoop. Edit file hadoop-env.sh.
- to edit using nano editor type : **nano /home/{yourname}/hadoop/etc/hadoop/hadoop-env.sh**
- type inside the file : **export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64/jre/**



```
GNU nano 2.9.3 /home/zmhadoop/hadoop/etc/hadoop/hadoop-env.sh

# The only required environment variable is JAVA_HOME. All others are
# optional. When running a distributed configuration it is best to
# set JAVA_HOME in this file, so that it is correctly defined on
# remote nodes.

# The java implementation to use.
# export JAVA_HOME=${JAVA_HOME}
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64/jre/
# the java implementation to use. Jsvc is required to run secure datanodes
# that bind to privileged ports to provide authentication of data transfer
# protocol. Jsvc is not required if SASL is configured for authentication of
# data transfer protocol using non-privileged ports.
#export JSVC_HOME=${JSVC_HOME}

export HADOOP_CONF_DIR=${HADOOP_CONF_DIR:-"/etc/hadoop"}

# Extra Java CLASSPATH elements. Automatically insert capacity-scheduler.
^G Get Help ^O Write Out ^W Where Is ^K Cut Text ^J Justify ^C Cur Pos
^X Exit ^R Read File ^\ Replace ^U Uncut Text ^T To Linter ^_ Go To Line
```

1.3 Run Hadoop

- to allow easy hadoop function access, edit file .bashrc as below.
- to open and edit file type : nano /home/{yourname}/.bashrc

```

GNU nano 2.9.3 /home/zm1/.bashrc Modified

# this, if it's already enabled in /etc/bash.bashrc and /etc/profile
# sources /etc/bash.bashrc).
if ! shopt -oq posix; then
  if [ -f /usr/share/bash-completion/bash_completion ]; then
    . /usr/share/bash-completion/bash_completion
  elif [ -f /etc/bash_completion ]; then
    . /etc/bash_completion
  fi
fi

# for hadoop configuration
export PATH=$PATH:/home/zmhadoop/hadoop/bin
export PATH=$PATH:/home/zmhadoop/hadoop/sbin
  
```

- type **hadoop** in the terminal to run the application

```

File Edit View Search Terminal Help
zm1@zm1-VirtualBox:/home$ hadoop
Usage: hadoop [--config confdir] [COMMAND | CLASSNAME]
  CLASSNAME                run the class named CLASSNAME
or
where COMMAND is one of:
  fs                        run a generic filesystem user client
  version                   print the version
  jar <jar>                 run a jar file
                           note: please use "yarn jar" to launch
                           YARN applications, not this command.
  checknative [-a|-h]      check native hadoop and compression libraries availability
  distcp <srcurl> <desturl> copy file or directories recursively
  archive -archiveName NAME -p <parent path> <src>* <dest> create a hadoop archive
  classpath                 prints the class path needed to get the
                           interact with credential providers
                           Hadoop jar and the required libraries
  daemonlog                 get/set the log level for each daemon
  trace                     view and modify Hadoop tracing settings

Most commands print help when invoked w/o parameters.
zm1@zm1-VirtualBox:/home$
  
```

- screen above shown that hadoop is running in a good condition.
- to check the version of hadoop type : **hadoop version**

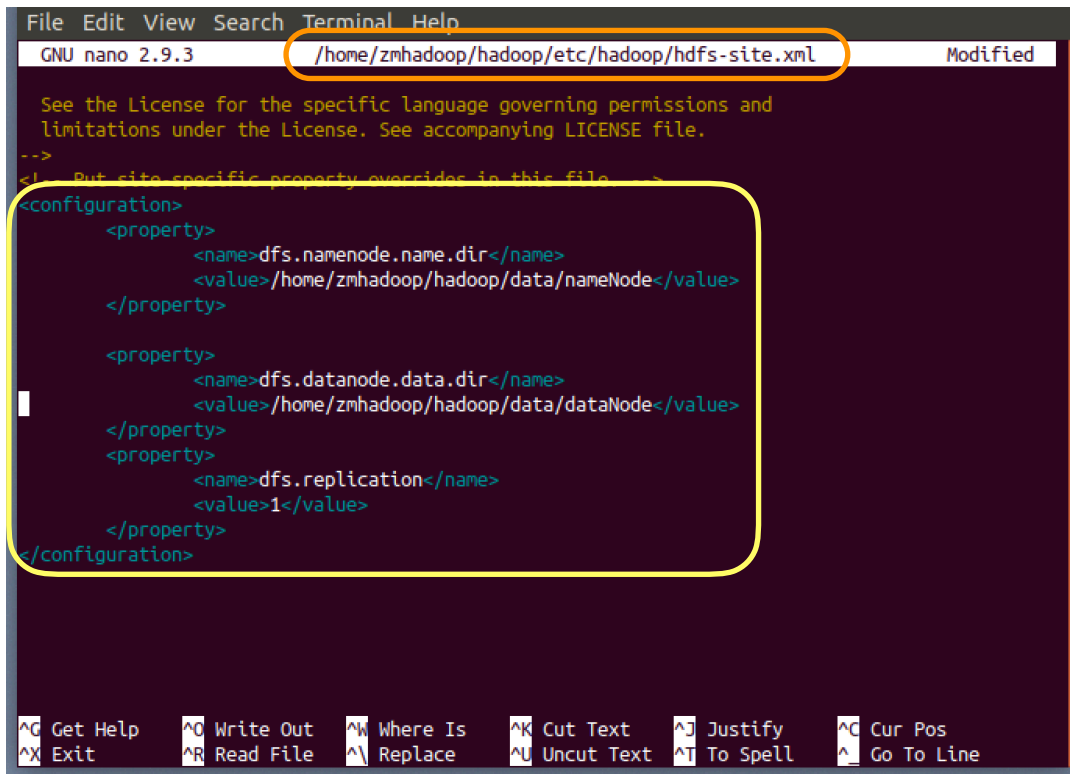
```

zm1@zm1-VirtualBox:/home$ hadoop version
Hadoop 2.7.7
Subversion Unknown -r c1aad84bd27cd79c3d1a7dd58202a8c3ee1ed3ac
Compiled by stevel on 2018-07-18T22:47Z
Compiled with protoc 2.5.0
From source with checksum 792e15d20b12c74bd6f19a1fb886490
This command was run using /home/zmhadoop/hadoop/share/hadoop/common/hadoop-common-2.7.7.jar
zm1@zm1-VirtualBox:/home$
  
```

STEP 2 : CONFIGURE HDFS, MAPREDUCE & YARN

2.1 Configure HDFS

- Update properties inside hfs-site.xml.
- type : **nano /home/{yourname}/hadoop/etc/hadoop/hdfs-site.xml**
- Enter the property as below. Change your path accordingly.



The screenshot shows the nano 2.9.3 editor with the file `/home/zmhadoop/hadoop/etc/hadoop/hdfs-site.xml` open. The file content is as follows:

```

File Edit View Search Terminal Help
GNU nano 2.9.3 /home/zmhadoop/hadoop/etc/hadoop/hdfs-site.xml Modified

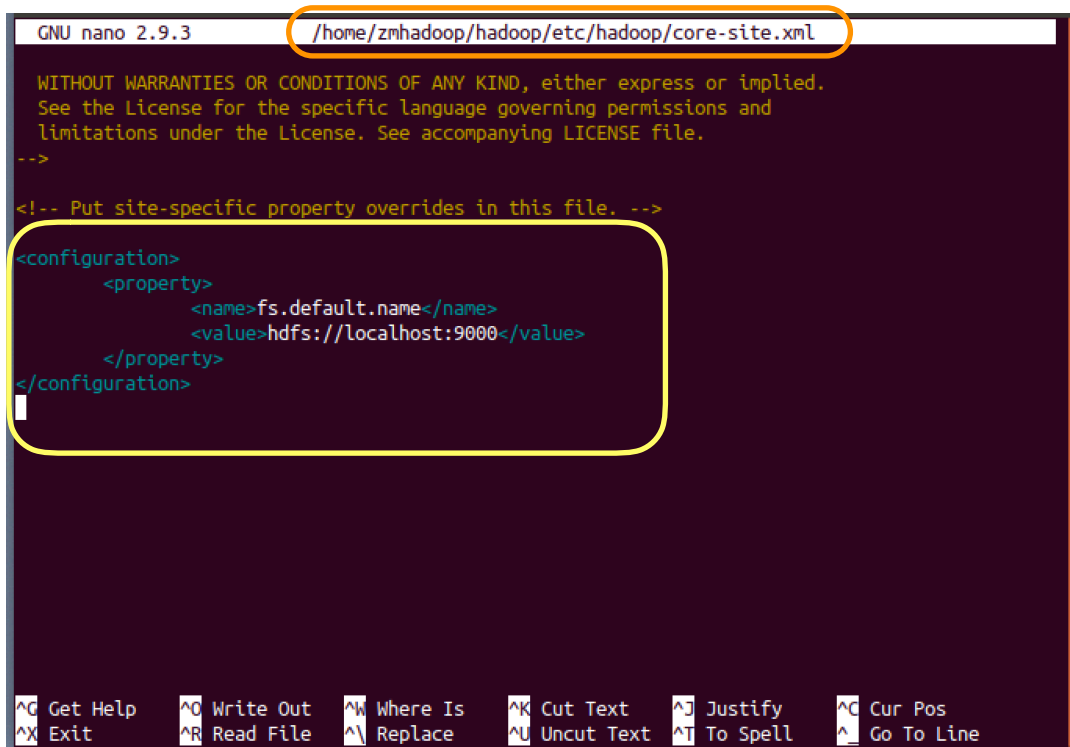
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->
<!-- Put site-specific property overrides in this file -->
<configuration>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>/home/zmhadoop/hadoop/data/nameNode</value>
  </property>

  <property>
    <name>dfs.datanode.data.dir</name>
    <value>/home/zmhadoop/hadoop/data/dataNode</value>
  </property>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
</configuration>

```

The bottom status bar shows various nano editor shortcuts: ^G Get Help, ^O Write Out, ^W Where Is, ^K Cut Text, ^J Justify, ^C Cur Pos, ^X Exit, ^R Read File, ^\ Replace, ^U Uncut Text, ^T To Spell, and ^_ Go To Line.

- Update properties inside core-site.xml.
- type : **nano /home/{yourname}/hadoop/etc/hadoop/core-site.xml**



The screenshot shows the nano 2.9.3 editor with the file `/home/zmhadoop/hadoop/etc/hadoop/core-site.xml` open. The file content is as follows:

```

GNU nano 2.9.3 /home/zmhadoop/hadoop/etc/hadoop/core-site.xml Modified

WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->
<!-- Put site-specific property overrides in this file. -->
<configuration>
  <property>
    <name>fs.default.name</name>
    <value>hdfs://localhost:9000</value>
  </property>
</configuration>

```

The bottom status bar shows various nano editor shortcuts: ^G Get Help, ^O Write Out, ^W Where Is, ^K Cut Text, ^J Justify, ^C Cur Pos, ^X Exit, ^R Read File, ^\ Replace, ^U Uncut Text, ^T To Spell, and ^_ Go To Line.

2.2 Configure MAPREDUCE & YARN

- Update properties inside mapred-site.xml.

```
GNU nano 2.9.3 /home/zmhadoop/hadoop/etc/hadoop/mapred-site.xml

WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
  </property>
  <property>
    <name>yarn.app.mapreduce.am.resource.mb</name>
    <value>512</value>
  </property>
  <property>
    <name>mapreduce.map.memory.mb</name>
    <value>256</value>
  </property>
  <property>
    <name>mapreduce.reduce.memory.mb</name>
    <value>256</value>
  </property>
</configuration>

^G Get Help  ^O Write Out  ^W Where Is  ^K Cut Text  ^J Justify   ^C Cur Pos
^X Exit      ^R Read File  ^\ Replace   ^U Uncut Text ^T To Spell  ^_ Go To Line
```

- Update properties inside yarn-site.xml.

```
GNU nano 2.9.3 /home/zmhadoop/hadoop/etc/hadoop/yarn-site.xml Modified

<?xml version="1.0"?>

<configuration>
  <property>
    <name>yarn.acl.enable</name>
    <value>0</value>
  </property>

  <property>
    <name>yarn.resourcemanager.hostname</name>
    <value>localhost</value>
  </property>

  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>
  <property>
    <name>yarn.nodemanager.resource.memory-mb</name>
    <value>1536</value>
  </property>
  <property>
    <name>yarn.scheduler.maximum-allocation-mb</name>
    <value>1536</value>
  </property>
  <property>
    <name>yarn.scheduler.minimum-allocation-mb</name>
    <value>128</value>
  </property>
  <property>
    <name>yarn.nodemanager.vmem-check-enabled</name>
    <value>>false</value>
  </property>

  <!-- Site specific YARN configuration properties -->
</configuration>
```

2.3 Starting all services (HDFS & YARN)

- type : **run `hdfs namenode -format`**
- type : **run `start-all.sh` (or run `start-dfs.sh` and `start-yarn.sh` separately)**
- to check all the services are running properly type : **`jps`**

```

File Edit View Search Terminal Help
zm1@zm1-VirtualBox:~$ jps
1957 DataNode
2181 SecondaryNameNode
2678 NodeManager
1768 NameNode
3038 Jps
2335 ResourceManager
zm1@zm1-VirtualBox:~$

```

- Access namenode and secondary node using browser :
<http://localhost:50070> and <http://localhost:50090>

Left Screenshot: <http://localhost:50070/dfshealth.html#tab-overview>

Overview 'localhost:9000' (active)

Started:	Sun Mar 22 17:24:16 MYT 2020
Version:	2.7.7, rc1aad84bd27cd79c3d1a7dd58202a8c3ee1ed3ac
Compiled:	2018-07-18T22:47Z by stevel from branch-2.7.7
Cluster ID:	CID-afc159b5-ccca-41df-8aa3-6a9529371e8b
Block Pool ID:	BP-702544374-127.0.1.1-1584806429957

Summary

Security is off.
 Safemode is off.
 101 files and directories, 56 blocks = 157 MB
 Heap Memory used 27.33 MB of 60.06 MB
 Non Heap Memory used 38.04 MB of 38.88 MB

Configured Capacity:	
DFS Used:	
Non DFS Used:	
DFS Remaining:	
Block Pool Used:	
DataNodes usages% (Min/Median/Max)	

[Live Nodes](#)

Right Screenshot: <http://localhost:50090/status.html>

Overview

Version	2.7.7
Compiled	2018-07-18T22:47Z by stevel from branch-2.7.7
NameNode Address	localhost:9000
Started	3/22/2020, 5:24:30 PM
Last Checkpoint	Never
Checkpoint Period	3600 seconds
Checkpoint Transactions	1000000

Checkpoint Image URI

- file:///tmp/hadoop-zm1/dfs/namesecondary

Checkpoint Editlog URI

- file:///tmp/hadoop-zm1/dfs/namesecondary

Hadoop, 2018.

STEP 3 : INSTALL APACHE HIVE

3.1 Download and Install Hive

- to download type : <https://downloads.apache.org/hive/hive-3.1.2/apache-hive-3.1.2-bin.tar.gz>
- to unzip & install type : **tar -xzf apache-hive-3.1.2-bin.tar.gz**
- similar to hadoop, to allow easy hive function access, edit file .bashrc as below.
- to open and edit file type : nano /home/{yourname}/.basic
- to get sync type : source .bashrc

```
GNU nano 2.9.3 /home/zm1/.bashrc

if ! shopt -oq posix; then
  if [ -f /usr/share/bash-completion/bash_completion ]; then
    . /usr/share/bash-completion/bash_completion
  elif [ -f /etc/bash_completion ]; then
    . /etc/bash_completion
  fi
fi

# for hadoop configuration
export PATH=$PATH:/home/zmhadoop/hadoop/bin
export PATH=$PATH:/home/zmhadoop/hadoop/sbin

# for hive configuration
export HIVE_HOME=/home/zmhadoop/apache-hive-3.1.2-bin
export PATH=$PATH:/home/zmhadoop/apache-hive-3.1.2-bin/bin

^G Get Help  ^O Write Out  ^W Where Is   ^K Cut Text   ^J Justify    ^C Cur Pos
^X Exit      ^R Read File  ^\ Replace    ^U Uncut Text ^T To Spell   ^_ Go To Line
```

- create warehouse folder inside HDFS

```
File Edit View Search Terminal Help
zm1@zm1-VirtualBox:~$ hdfs dfs -mkdir -p /user/hive/warehouse
zm1@zm1-VirtualBox:~$
```

- change folder permission to the directory /user/hive/warehouse and /tmp

```
File Edit View Search Terminal Help
zm1@zm1-VirtualBox:/home/zmhadoop$ hdfs dfs -chmod g+w /user/hive/warehouse
zm1@zm1-VirtualBox:/home/zmhadoop$ hdfs dfs -chmod g+w /tmp
zm1@zm1-VirtualBox:/home/zmhadoop$
```

- initiate derby and create metastore_db type : schema tool-dbType derby -initSchema

```
File Edit View Search Terminal Help
zm1@zm1-VirtualBox:/home/zmhadoop$ schematool -dbType derby initSchema
```


STEP 4 : CREATE DATABASE & TABLE FROM HIVE

- to run hive type : **hive**

```
File Edit View Search Terminal Help
zm1@zm1-VirtualBox:/home/zmhadoop$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/zmhadoop/apache-hive-3.1.2-bin/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/zmhadoop/hadoop/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]
Hive Session ID = 6ca6a7d1-df63-4b4e-a29c-0da08608a160

Logging initialized using configuration in jar:file:/home/zmhadoop/apache-hive-3.1.2-bin/lib/hive-common-3.1.2.jar!/hive-log4j2.properties Async: true
Sun Mar 22 18:24:22 MYT 2020 Thread[6ca6a7d1-df63-4b4e-a29c-0da08608a160 main,5,main] java.io.
FileNotFoundException: derby.log (Permission denied)
Sun Mar 22 18:24:22 MYT 2020 Thread[6ca6a7d1-df63-4b4e-a29c-0da08608a160 main,5,main] Ignored
duplicate property derby.module.replication.master in jar:file:/home/zmhadoop/apache-hive-3.1.2-bin/lib/hive-druid-handler-3.1.2.jar!/org/apache/derby/modules.properties
Sun Mar 22 18:24:22 MYT 2020 Thread[6ca6a7d1-df63-4b4e-a29c-0da08608a160 main,5,main] Ignored
duplicate property derby.module.resultSetStatisticsFactory in jar:file:/home/zmhadoop/apache-hive-3.1.2-bin/lib/hive-druid-handler-3.1.2.jar!/org/apache/derby/modules.properties
Sun Mar 22 18:24:22 MYT 2020 Thread[6ca6a7d1-df63-4b4e-a29c-0da08608a160 main,5,main] Ignored
duplicate property derby.module.NoneAuthentication in jar:file:/home/zmhadoop/apache-hive-3.1.2-bin/lib/hive-druid-handler-3.1.2.jar!/org/apache/derby/modules.properties
Sun Mar 22 18:24:22 MYT 2020 Thread[6ca6a7d1-df63-4b4e-a29c-0da08608a160 main,5,main] Ignored
duplicate property derby.module.lockManagerJ6 in jar:file:/home/zmhadoop/apache-hive-3.1.2-bin/lib/hive-druid-handler-3.1.2.jar!/org/apache/derby/modules.properties
```

- to show database type : **show databases;**
- to create database type : **create database crudeoil;** or use the default database.

```
File Edit View Search Terminal Help
hive> show databases;
OK
crudeoil
default
Time taken: 0.024 seconds, Fetched: 2 row(s)
```

- create table for type : **create table crudeoilbrent (dateprice string, closingprice double, openprice double, dailyhigh double, dailylow double) row format delimited fields terminated by ',' tblproperties('skip.header.line.count'='1');**

```
File Edit View Search Terminal Help
hive> create table crudeoilbrent (dateprice string, closingprice double, openprice double, dailyhigh double, dailylow double) row format delimited fields terminated by ',' tblproperties('skip.header.line.count'='1');
```

- for this project we have created two tables crudeoilbrent and crudeoilwti.
- to show table lists type : **show tables;**
- to show table properties type : **describe crudeoilbrent;**

```
File Edit View Search Terminal Help
hive> show tables;
OK
crudeoilbrent
Time taken: 0.047 seconds, Fetched: 1 row(s)
hive> describe crudeoilbrent;
OK
dateprice          string
closingprice       double
openprice          double
dailyhigh          double
dailylow           double
Time taken: 0.084 seconds, Fetched: 5 row(s)
hive>
```

STEP 5 & 6 : LOAD CSV FILE INTO HIVE DATA WAREHOUSE & RETRIEVE THE DATA

- Using milestone 1 data that we have scrape and put inside csv file.
- Load the csv file into the hive tables that we have created using this command : **load data local inpath '/home/zm1/mywebscrapBrentFinal.csv' overwrite into table crudeoilbrent;**

```
File Edit View Search Terminal Help
hive> load data local inpath '/home/zm1/mywebscrapBrentFinal.csv' overwrite into table crudeoilbrent;
Loading data to table default.crudeoilbrent
OK
Time taken: 0.336 seconds
hive> █
```

- Change the path accordingly refer to where the csv file is located.
- For this project we load two types of record into two different table named 'crudeoilbrent' and 'crudeoilwti' that contained crude oil brent price and crude oil wti price.
- To show the contents of the table type : **select * from crudeoilbrent;**

```
File Edit View Search Terminal Help
2/22/2006      61.53      61.53      61.53      61.53
2/21/2006      61.48      61.48      61.48      61.48
2/20/2006      59.92      59.92      59.92      59.92
2/17/2006      58.76      58.76      58.76      58.76
2/16/2006      59.46      59.46      59.46      59.46
2/15/2006      60.27      60.27      60.27      60.27
2/14/2006      59.55      59.55      59.55      59.55
2/13/2006      60.53      60.53      60.53      60.53
2/10/2006      61.37      61.37      61.37      61.37
2/9/2006       61.5       61.5       61.5       61.5
2/8/2006       62.46      62.46      62.46      62.46
2/7/2006       63.96      63.96      63.96      63.96
2/6/2006       63.18      63.18      63.18      63.18
2/3/2006       64.24      64.24      64.24      64.24
2/2/2006       65.85      65.85      65.85      65.85
2/1/2006       66.14      66.14      66.14      66.14
1/31/2006      66.15      66.15      66.15      66.15
1/30/2006      65.99      65.99      65.99      65.99
1/27/2006      64.71      64.71      64.71      64.71
1/26/2006      64.69      64.69      64.69      64.69
1/25/2006      65.64      65.64      65.64      65.64
1/24/2006      66.15      66.15      66.15      66.15
1/23/2006      66.11      66.11      66.11      66.11
1/20/2006      64.73      64.73      64.73      64.73
1/19/2006      65.0       65.0       65.0       65.0
1/18/2006      64.28      64.28      64.28      64.28
1/17/2006      62.92      62.92      62.92      62.92
1/16/2006      62.56      62.56      62.56      62.56
1/13/2006      63.02      63.02      63.02      63.02
1/12/2006      61.87      61.87      61.87      61.87
1/11/2006      62.18      62.18      62.18      62.18
1/10/2006      62.71      62.71      62.71      62.71
1/9/2006       62.24      62.24      62.24      62.24
1/6/2006       61.62      61.62      61.62      61.62
1/5/2006       61.05      61.05      61.05      61.05
1/4/2006       60.37      60.37      60.37      60.37
1/3/2006       58.16      58.16      58.16      58.16
Time taken: 0.23 seconds, Fetched: 3675 row(s)
hive> █
```

- Compare the rows inserted into the table with the content of the csv files. The csv file was successfully loaded into the hive data warehouse.