Milestone 2 Management of Data

WQD7005 – DATA MINING

ZULKANAIN BIN HASAN

WQD180031

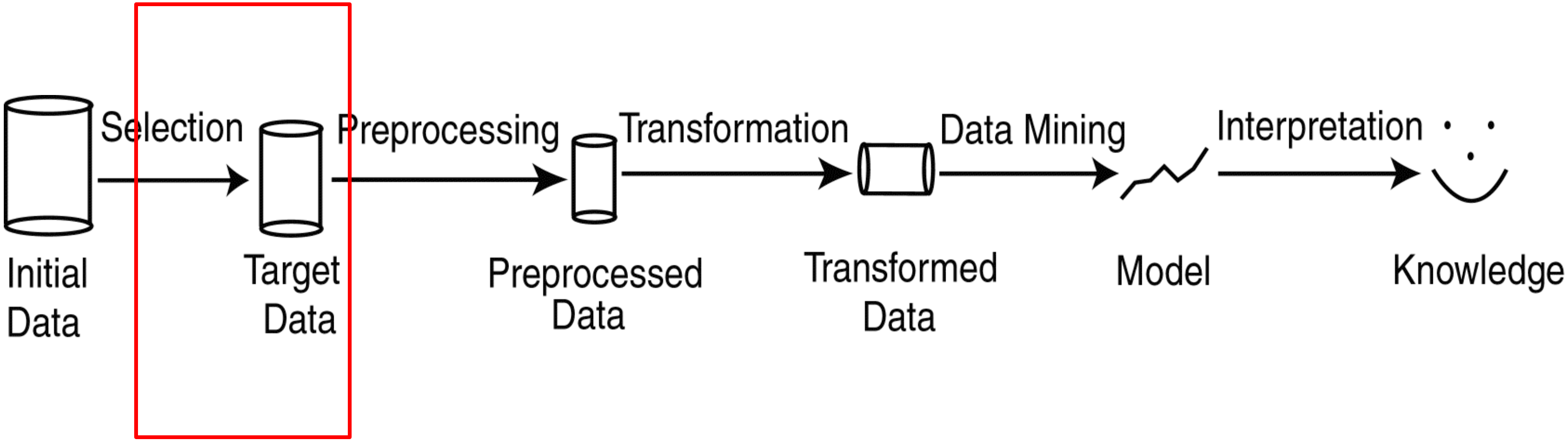
NUR ATHIRAH BT. RAZAK

WQD170072

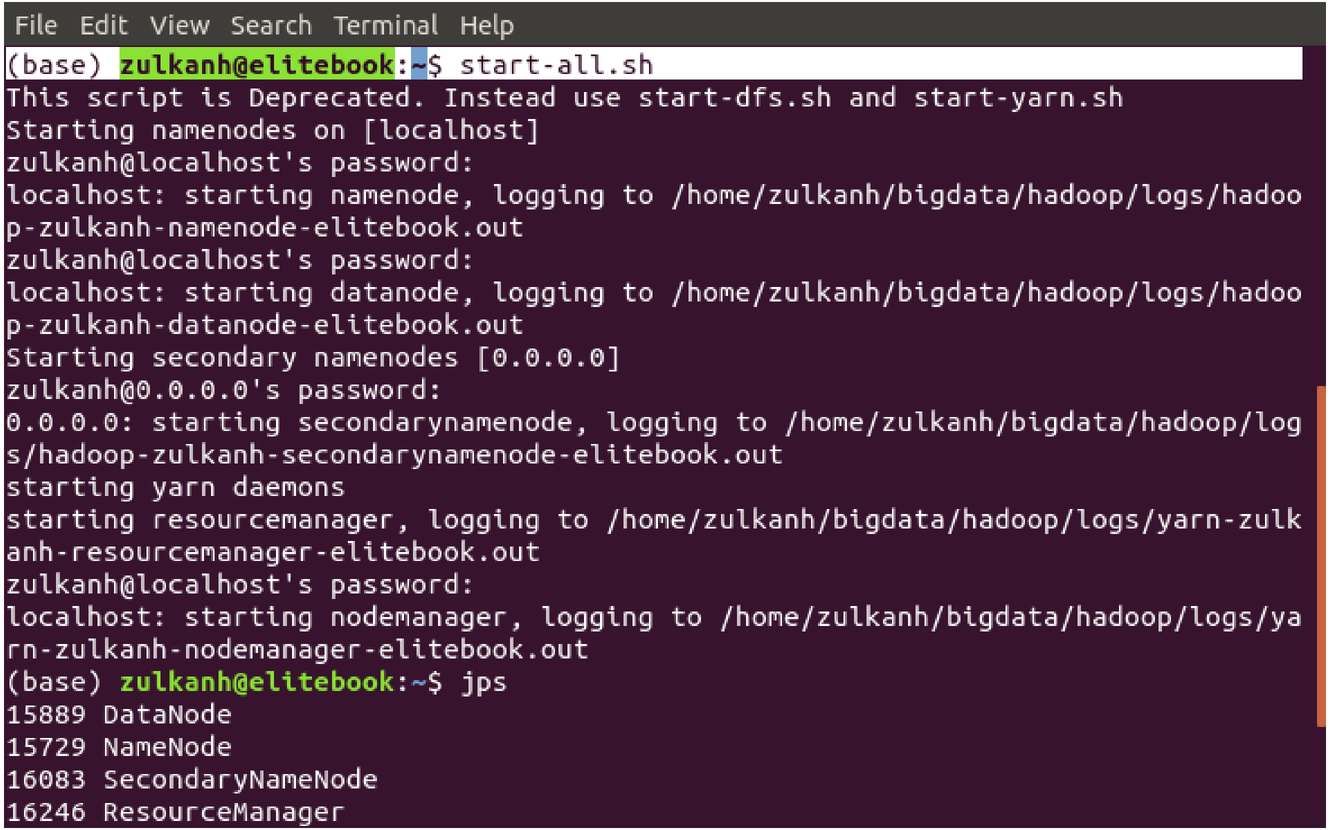
Knowledge Discovery in Databases (KDD): process of finding useful information and patterns in data:

1. Selection (Pre-Mining 1): Obtain data from various sources

Milestone 2

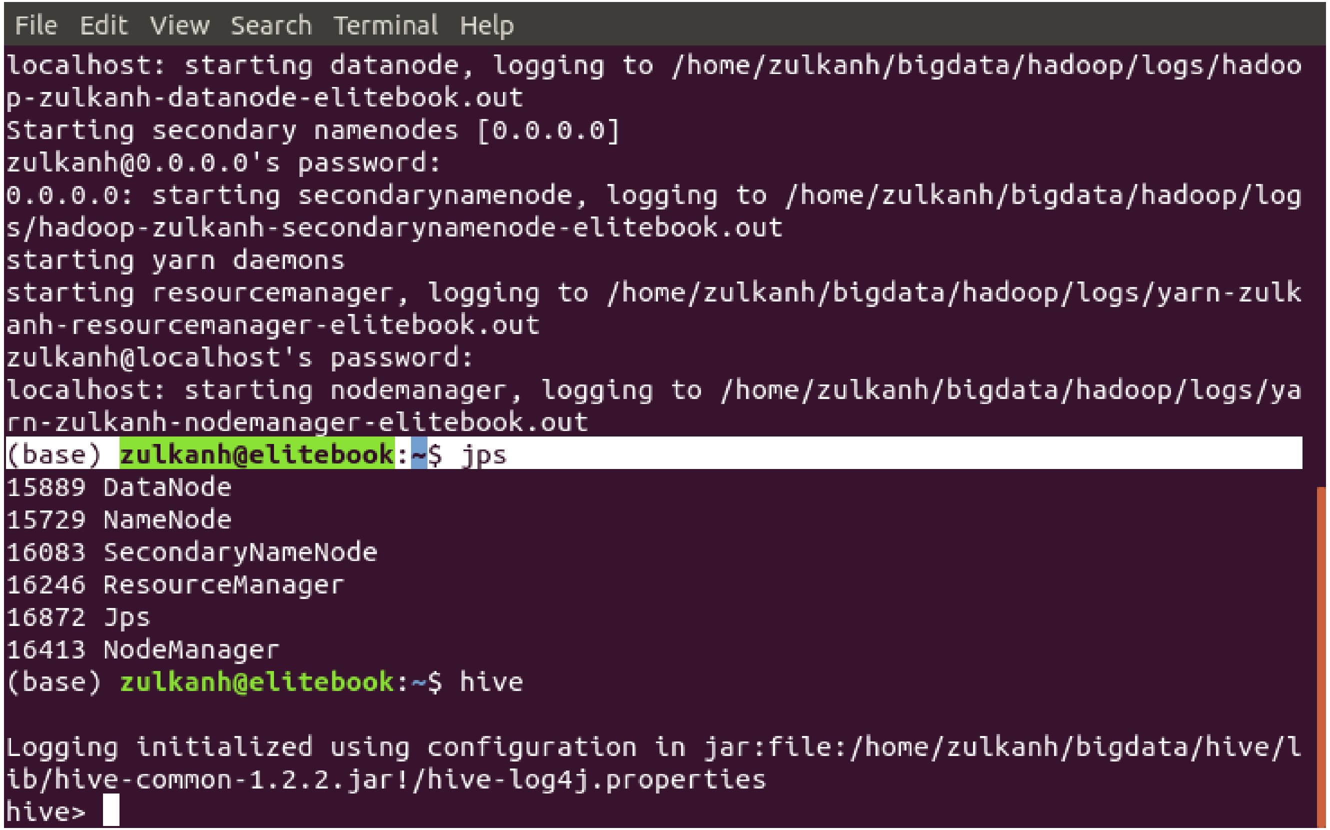


Start hadoop



```
(base) zulkanh@elitebook:~$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
zulkanh@localhost's password:
localhost: starting namenode, logging to /home/zulkanh/bigdata/hadoop/logs/hadoo
p-zulkanh-namenode-elitebook.out
zulkanh@localhost's password:
localhost: starting datanode, logging to /home/zulkanh/bigdata/hadoop/logs/hadoo
p-zulkanh-datanode-elitebook.out
Starting secondary namenodes [0.0.0.0]
zulkanh@0.0.0.0's password:
0.0.0.0: starting secondarynamenode, logging to /home/zulkanh/bigdata/hadoop/log
s/hadoop-zulkanh-secondarynamenode-elitebook.out
starting yarn daemons
starting resourcemanager, logging to /home/zulkanh/bigdata/hadoop/logs/yarn-zulk
anh-resourcemanager-elitebook.out
zulkanh@localhost's password:
localhost: starting nodemanager, logging to /home/zulkanh/bigdata/hadoop/logs/ya
rn-zulkanh-nodemanager-elitebook.out
(base) zulkanh@elitebook:~$ jps
15889 DataNode
15729 NameNode
16083 SecondaryNameNode
16246 ResourceManager
```

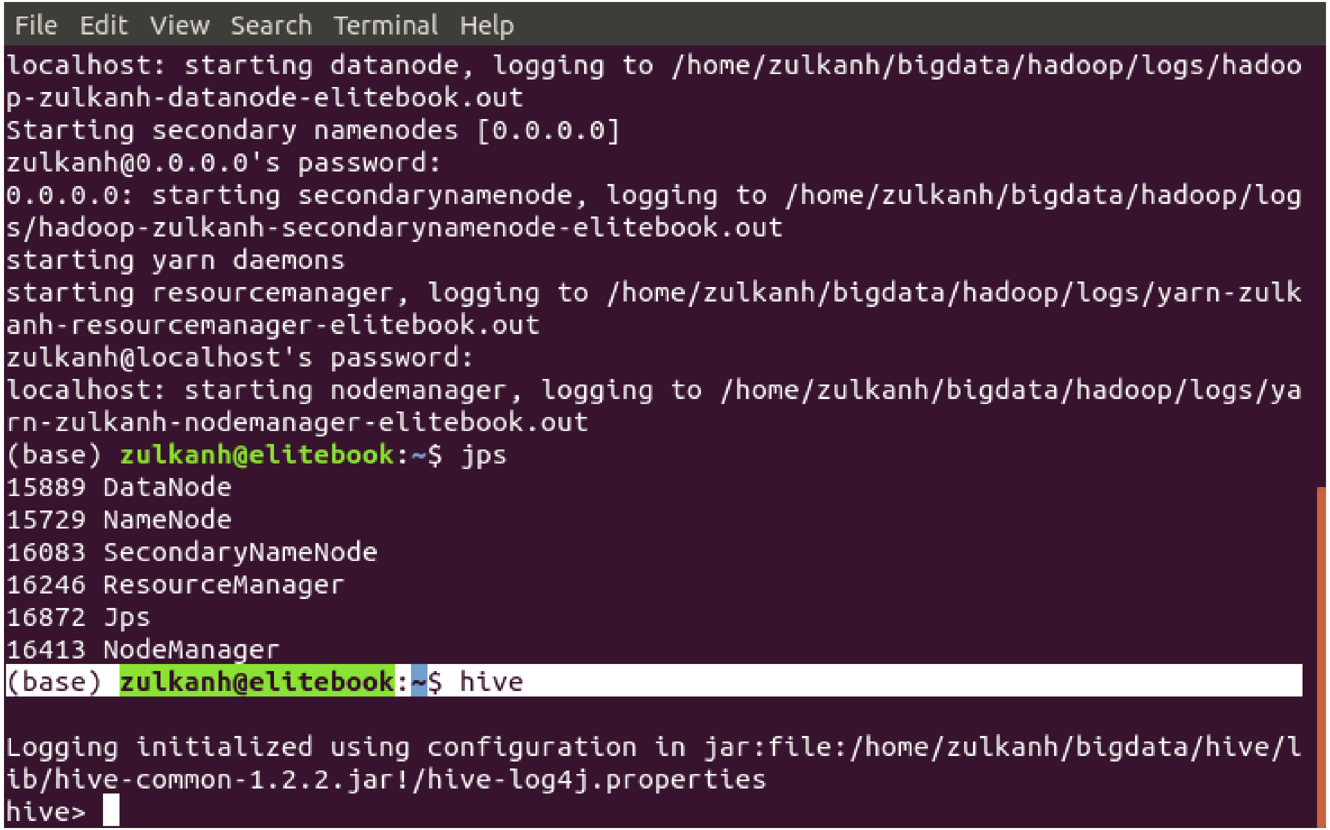
Check hadoop status



```
localhost: starting datanode, logging to /home/zulkanh/bigdata/hadoop/logs/hadoo
p-zulkanh-datanode-elitebook.out
Starting secondary namenodes [0.0.0.0]
zulkanh@0.0.0.0's password:
0.0.0.0: starting secondarynamenode, logging to /home/zulkanh/bigdata/hadoop/log
s/hadoop-zulkanh-secondarynamenode-elitebook.out
starting yarn daemons
starting resourcemanager, logging to /home/zulkanh/bigdata/hadoop/logs/yarn-zulk
anh-resourcemanager-elitebook.out
zulkanh@localhost's password:
localhost: starting nodemanager, logging to /home/zulkanh/bigdata/hadoop/logs/ya
rn-zulkanh-nodemanager-elitebook.out
(base) zulkanh@elitebook:~$ jps
15889 DataNode
15729 NameNode
16083 SecondaryNameNode
16246 ResourceManager
16872 Jps
16413 NodeManager
(base) zulkanh@elitebook:~$ hive

Logging initialized using configuration in jar:file:/home/zulkanh/bigdata/hive/l
ib/hive-common-1.2.2.jar!/hive-log4j.properties
hive>
```

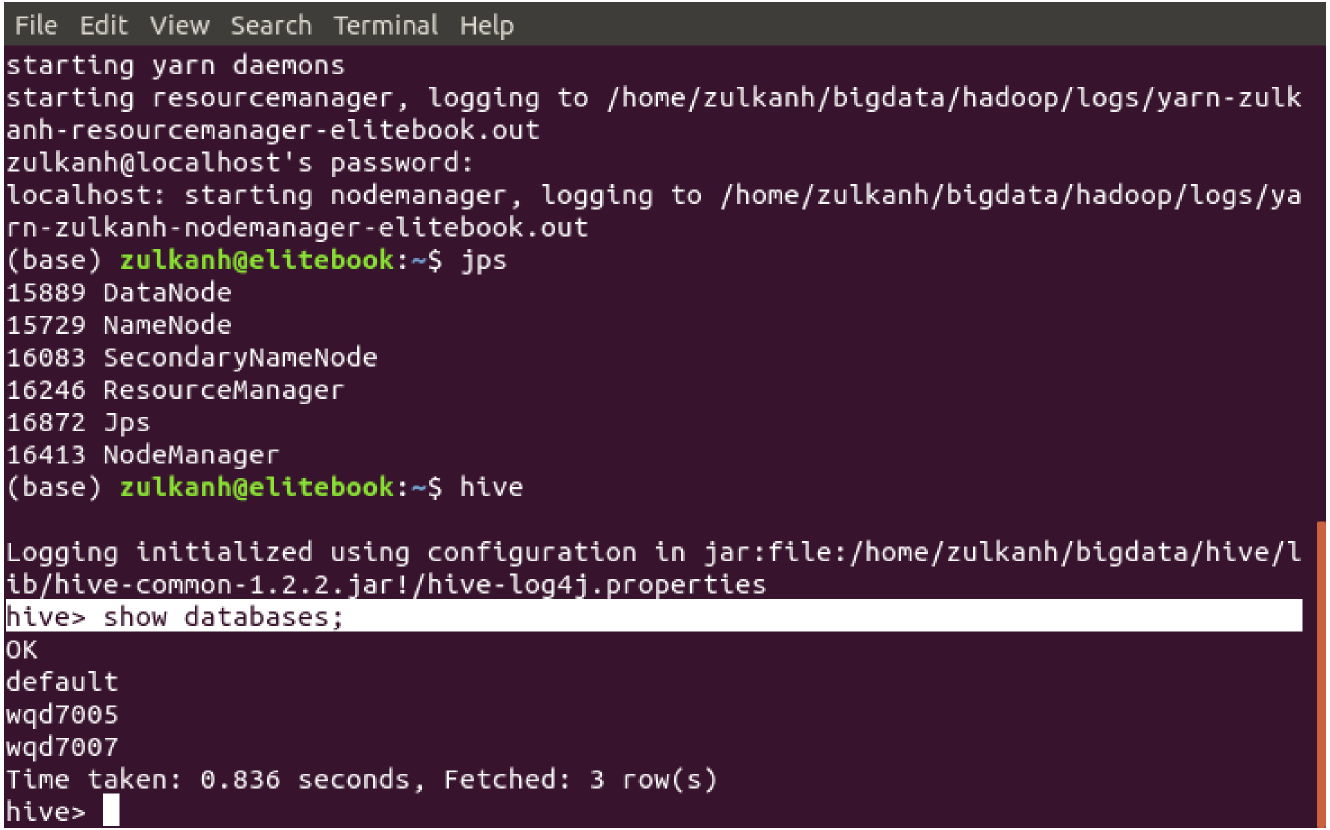
Entering hive



```
localhost: starting datanode, logging to /home/zulkanh/bigdata/hadoop/logs/hadoo
p-zulkanh-datanode-elitebook.out
Starting secondary namenodes [0.0.0.0]
zulkanh@0.0.0.0's password:
0.0.0.0: starting secondarynamenode, logging to /home/zulkanh/bigdata/hadoop/log
s/hadoop-zulkanh-secondarynamenode-elitebook.out
starting yarn daemons
starting resourcemanager, logging to /home/zulkanh/bigdata/hadoop/logs/yarn-zulk
anh-resourcemanager-elitebook.out
zulkanh@localhost's password:
localhost: starting nodemanager, logging to /home/zulkanh/bigdata/hadoop/logs/ya
rn-zulkanh-nodemanager-elitebook.out
(base) zulkanh@elitebook:~$ jps
15889 DataNode
15729 NameNode
16083 SecondaryNameNode
16246 ResourceManager
16872 Jps
16413 NodeManager
(base) zulkanh@elitebook:~$ hive

Logging initialized using configuration in jar:file:/home/zulkanh/bigdata/hive/l
ib/hive-common-1.2.2.jar!/hive-log4j.properties
hive> 
```

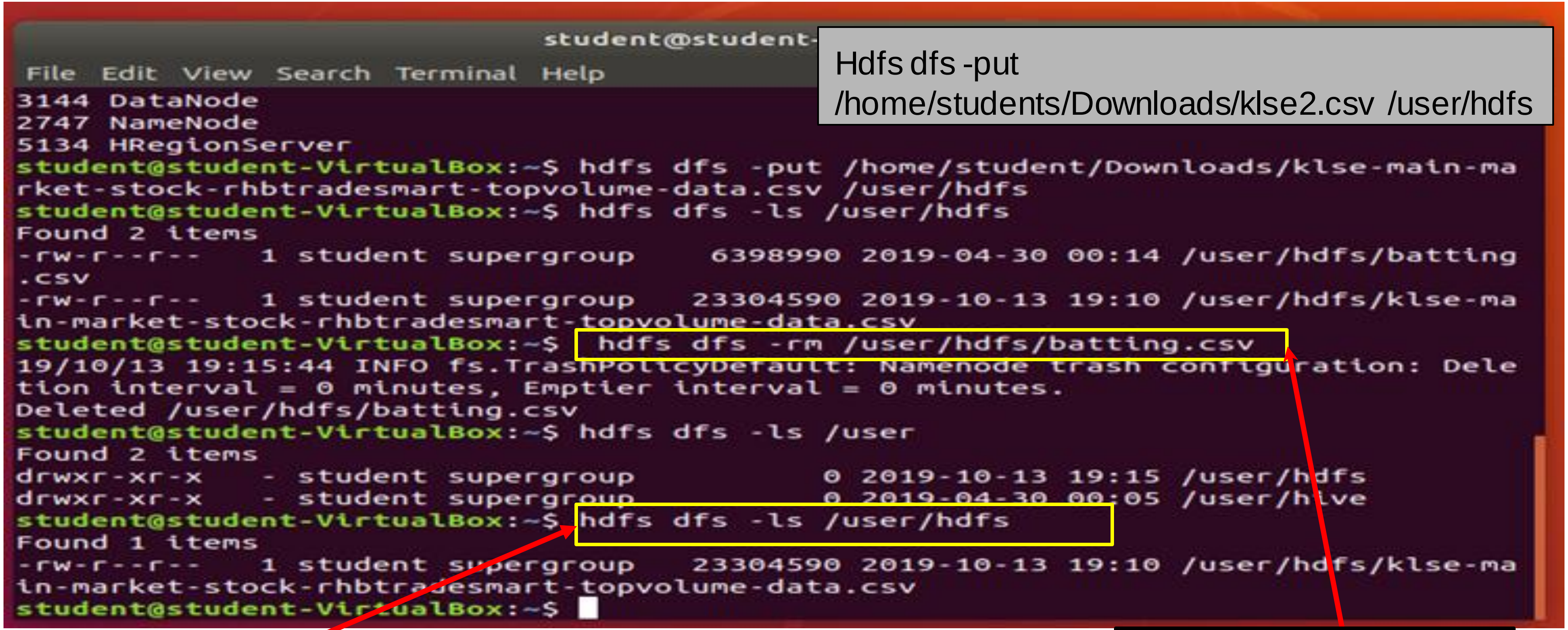
Show databases inside hive



```
starting yarn daemons
starting resourcemanager, logging to /home/zulkanh/bigdata/hadoop/logs/yarn-zulk
anh-resourcemanager-elitebook.out
zulkanh@localhost's password:
localhost: starting nodemanager, logging to /home/zulkanh/bigdata/hadoop/logs/ya
rn-zulkanh-nodemanager-elitebook.out
(base) zulkanh@elitebook:~$ jps
15889 DataNode
15729 NameNode
16083 SecondaryNameNode
16246 ResourceManager
16872 Jps
16413 NodeManager
(base) zulkanh@elitebook:~$ hive

Logging initialized using configuration in jar:file:/home/zulkanh/bigdata/hive/l
ib/hive-common-1.2.2.jar!/hive-log4j.properties
hive> show databases;
OK
default
wqd7005
wqd7007
Time taken: 0.836 seconds, Fetched: 3 row(s)
hive>
```

Import dataset to HDFS



```
3144 DataNode
2747 NameNode
5134 HRegionServer
student@student-VirtualBox:~$ hdfs dfs -put /home/student/Downloads/klse-main-ma
rket-stock-rhbtradesmart-topvolume-data.csv /user/hdfs
student@student-VirtualBox:~$ hdfs dfs -ls /user/hdfs
Found 2 items
-rw-r--r--   1 student supergroup    6398990 2019-04-30 00:14 /user/hdfs/batting
.csv
-rw-r--r--   1 student supergroup   23304590 2019-10-13 19:10 /user/hdfs/klse-ma
in-market-stock-rhbtradesmart-topvolume-data.csv
student@student-VirtualBox:~$  hdfs dfs -rm /user/hdfs/batting.csv
19/10/13 19:15:44 INFO fs.TrashPolicyDefault: Namenode trash configuration: Dele
tion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /user/hdfs/batting.csv
student@student-VirtualBox:~$ hdfs dfs -ls /user
Found 2 items
drwxr-xr-x   - student supergroup          0 2019-10-13 19:15 /user/hdfs
drwxr-xr-x   - student supergroup          0 2019-04-30 00:05 /user/hive
student@student-VirtualBox:~$ hdfs dfs -ls /user/hdfs
Found 1 items
-rw-r--r--   1 student supergroup   23304590 2019-10-13 19:10 /user/hdfs/klse-ma
in-market-stock-rhbtradesmart-topvolume-data.csv
student@student-VirtualBox:~$
```

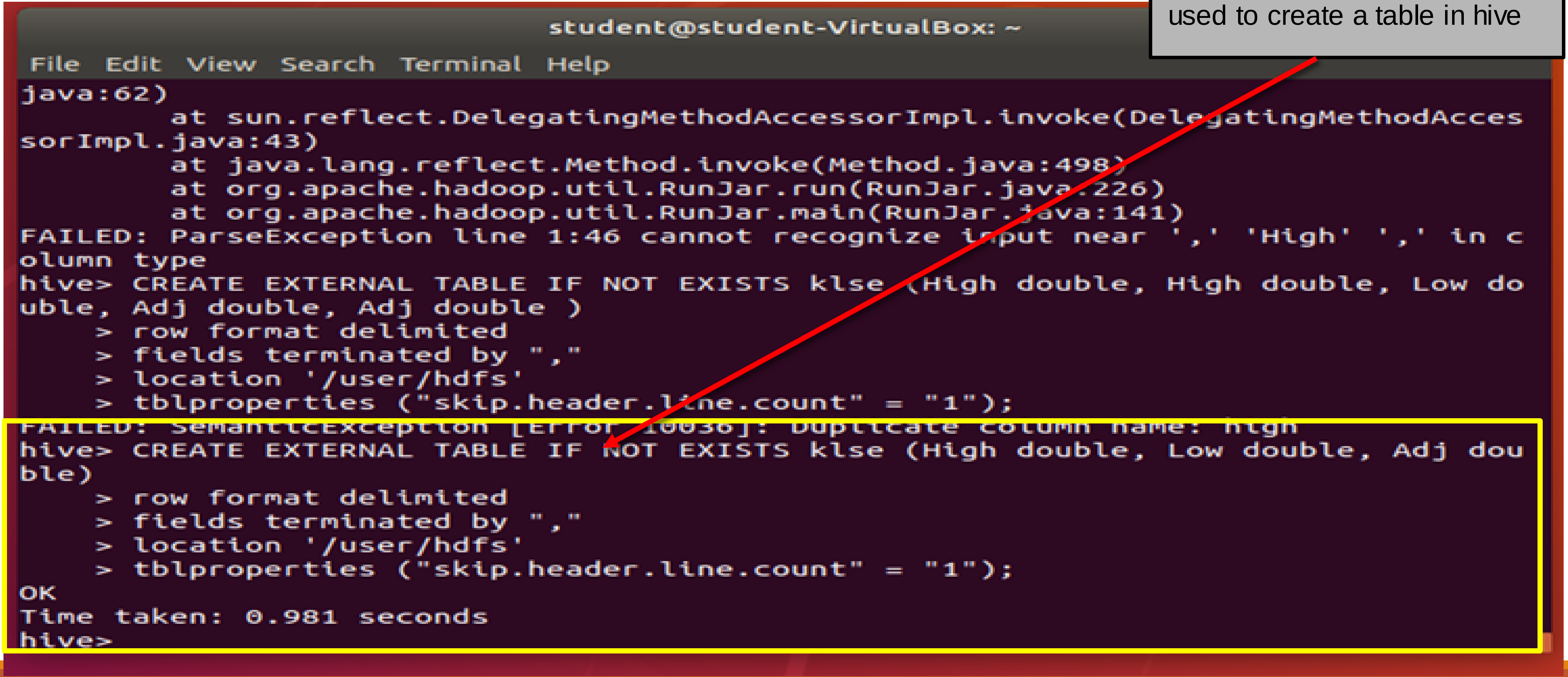
Hdfs dfs -put /home/students/Downloads/klse2.csv /user/hdfs

Manage to import dataset to HDFS

Remove all irrelevent document

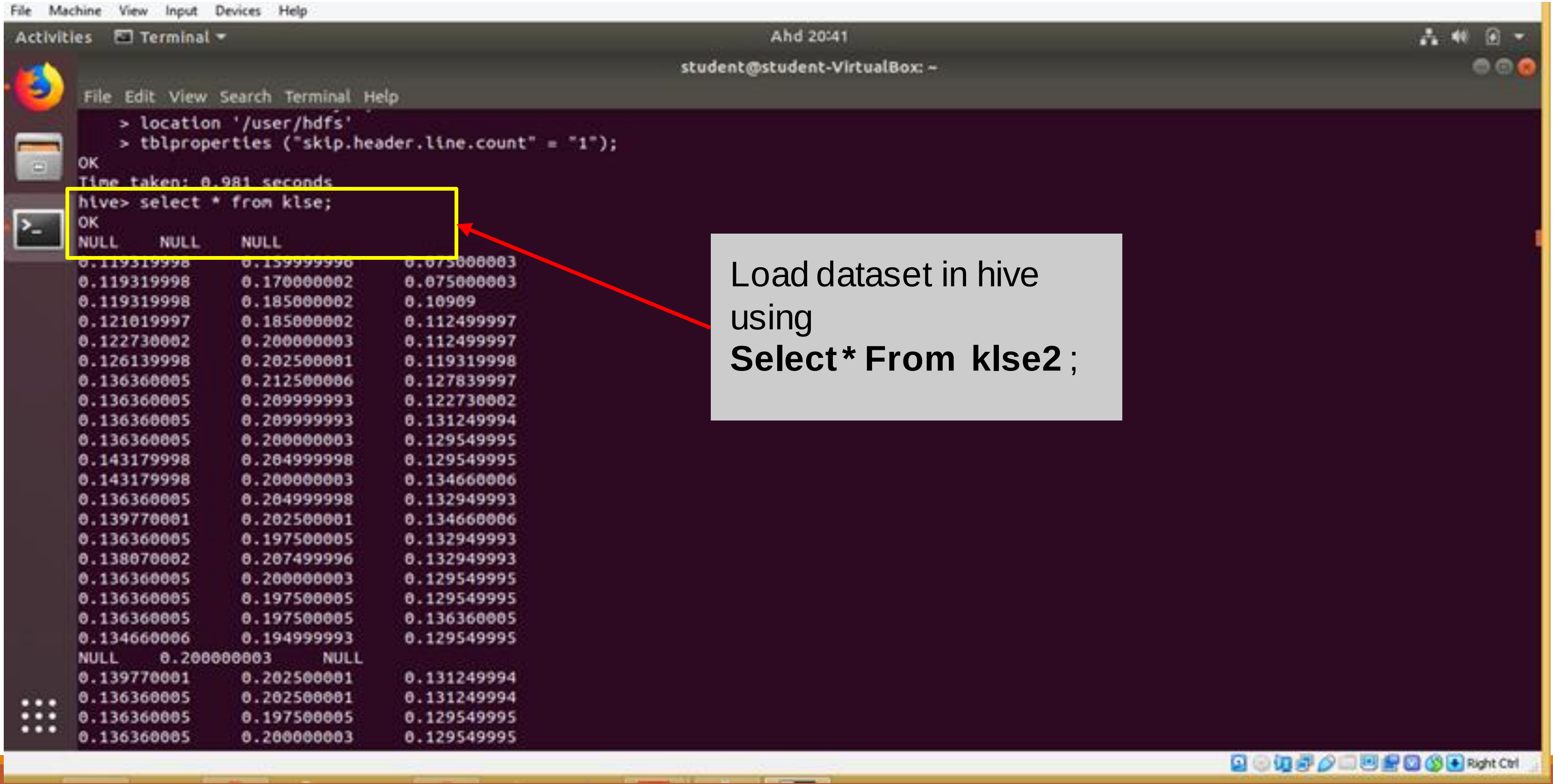
Create **EXTERNAL TABLE** to Hive

Create Table is a statement used to create a table in hive


student@student-VirtualBox: ~
File Edit View Search Terminal Help

```
java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAcces
sorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:226)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:141)
FAILED: ParseException line 1:46 cannot recognize input near ',' 'High' ',' in c
olumn type
hive> CREATE EXTERNAL TABLE IF NOT EXISTS klse (High double, High double, Low do
uble, Adj double, Adj double )
    > row format delimited
    > fields terminated by ","
    > location '/user/hdfs'
    > tblproperties ("skip.header.line.count" = "1");
FAILED: SemanticException [Error 10036]: Duplicate column name: high
hive> CREATE EXTERNAL TABLE IF NOT EXISTS klse (High double, Low double, Adj dou
ble)
    > row format delimited
    > fields terminated by ","
    > location '/user/hdfs'
    > tblproperties ("skip.header.line.count" = "1");
OK
Time taken: 0.981 seconds
hive>
```

LOAD DATASET


File Machine View Input Devices Help
Activities
Terminal
Ahd 20:41
student@student-VirtualBox: ~
File Edit View Search Terminal Help
Right Ctrl

```
    > location '/user/hdfs'
    > tblproperties ("skip.header.line.count" = "1");
OK
Time taken: 0.981 seconds
hive> select * from klse;
OK
NULL    NULL    NULL
0.119319998     0.159999996     0.075000003
0.119319998     0.170000002     0.075000003
0.119319998     0.185000002     0.10909
0.121019997     0.185000002     0.112499997
0.122730002     0.200000003     0.112499997
0.126139998     0.202500001     0.119319998
0.136360005     0.212500006     0.127839997
0.136360005     0.209999993     0.122730002
0.136360005     0.209999993     0.131249994
0.136360005     0.200000003     0.129549995
0.143179998     0.204999998     0.129549995
0.143179998     0.200000003     0.134660006
0.136360005     0.204999998     0.132949993
0.139770001     0.202500001     0.134660006
0.136360005     0.197500005     0.132949993
0.138070002     0.207499996     0.132949993
0.136360005     0.200000003     0.129549995
0.136360005     0.197500005     0.129549995
0.136360005     0.197500005     0.136360005
0.134660006     0.194999993     0.129549995
NULL    0.200000003     NULL
0.139770001     0.202500001     0.131249994
0.136360005     0.202500001     0.131249994
0.136360005     0.197500005     0.129549995
0.136360005     0.200000003     0.129549995
```

Load dataset in hive using
Select * From klse2 ;