

NAME- Akarsh RAJ

Employee id- 2125994

Cohort- INTCDB22DW043

Coach – Swetha Sri K

BU –AIA-BIG DATA

Technical Big-data Coach – Shrikant Gawande

Topic-Assignments of Big-data

1)Department wise total salary in hadoop (mapreduce)

2)Top 10 Customers by total sales. (using txns1.txt) (pig)

3)Top 10 designation with people count (hive)

4)Create a login table and put data into it and show the output (hbase)

5)Sqoop Program

i)using one mapper

ii)using two mapper

iii)using incremental append

iv)printing data into hive from sqoop

v)printing data into hbase from sqoop

6) Scala programming

- i)Create one empty ArrayBuffer
- ii)Add 10 into ArrayBuffer
- iii)Add 20,30,40 into ArrayBuffer
- iv)Defing arr={1,2,3,4,5}
- v)Append array into arrayBuffer
- vi)Remove 30
- vii)Add 35 in between 20 and 40
- viii)Check 35 is there in ArrayBuffer or not
- ix)Check ArrayBuffer is empty or not
- x)Remove last 3 elements
- xi)Add 10 to each element
- xii)Get element of ArrayBuffer where element>10

7)Scala Programming

- i)to get keys of the map
- ii)to get values of the map
- iii)To check whether key is present or not.
- iv)To get the value of given key.
- v)If value is not there return 0.0
- vi)If the value is not there,then add 0.0 to c3.

and wordcount program in scala

8)Read data from custs.txt and then sort the

data by age(in acending order)

using scala

9)Spark

Sub topic:-> using spark doing joining,getting top result

ASSIGNMENT NO 1.

PROBLEM STATEMENT->Department wise total

salary in hadoop

SUB TOPIC-USING MAPREDUCE

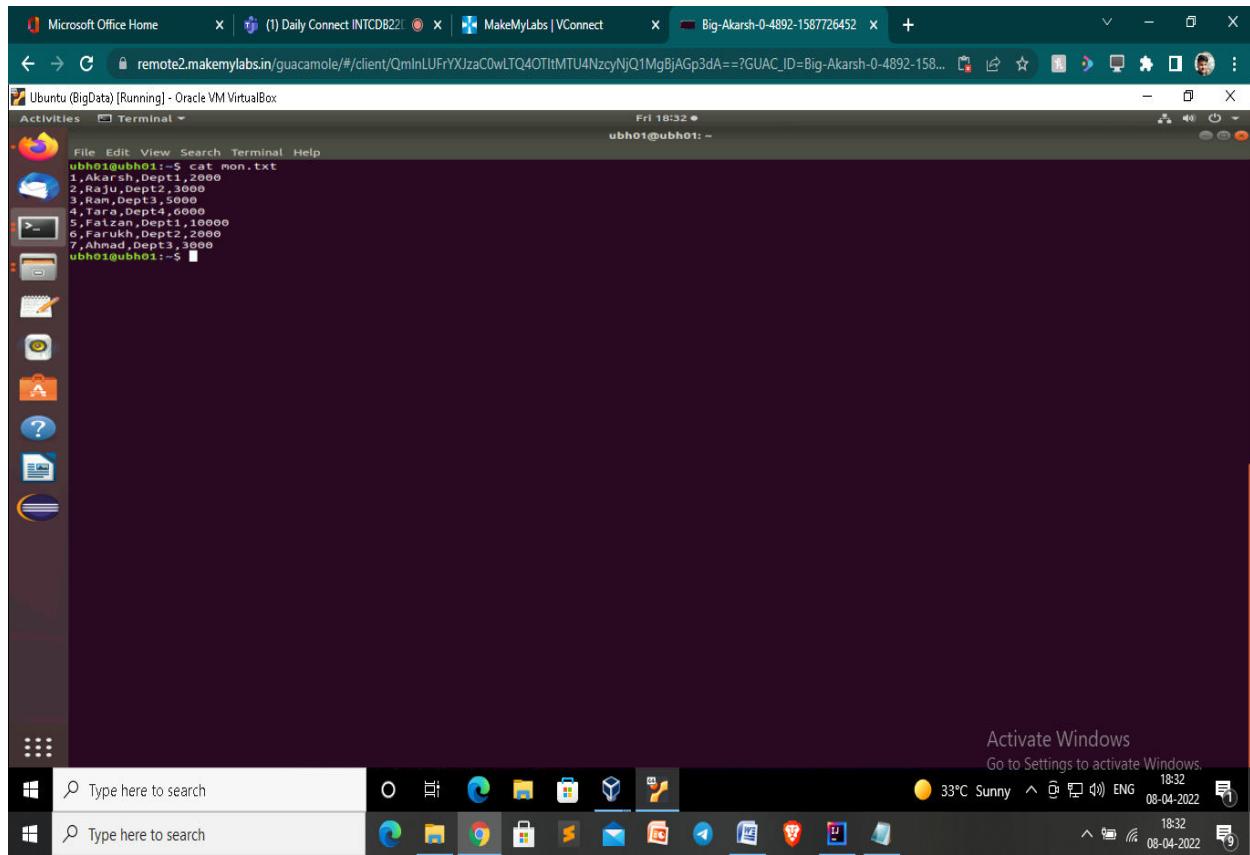
DATA SCHEMA-

*EID

*ENAME

*DEPARTMENT

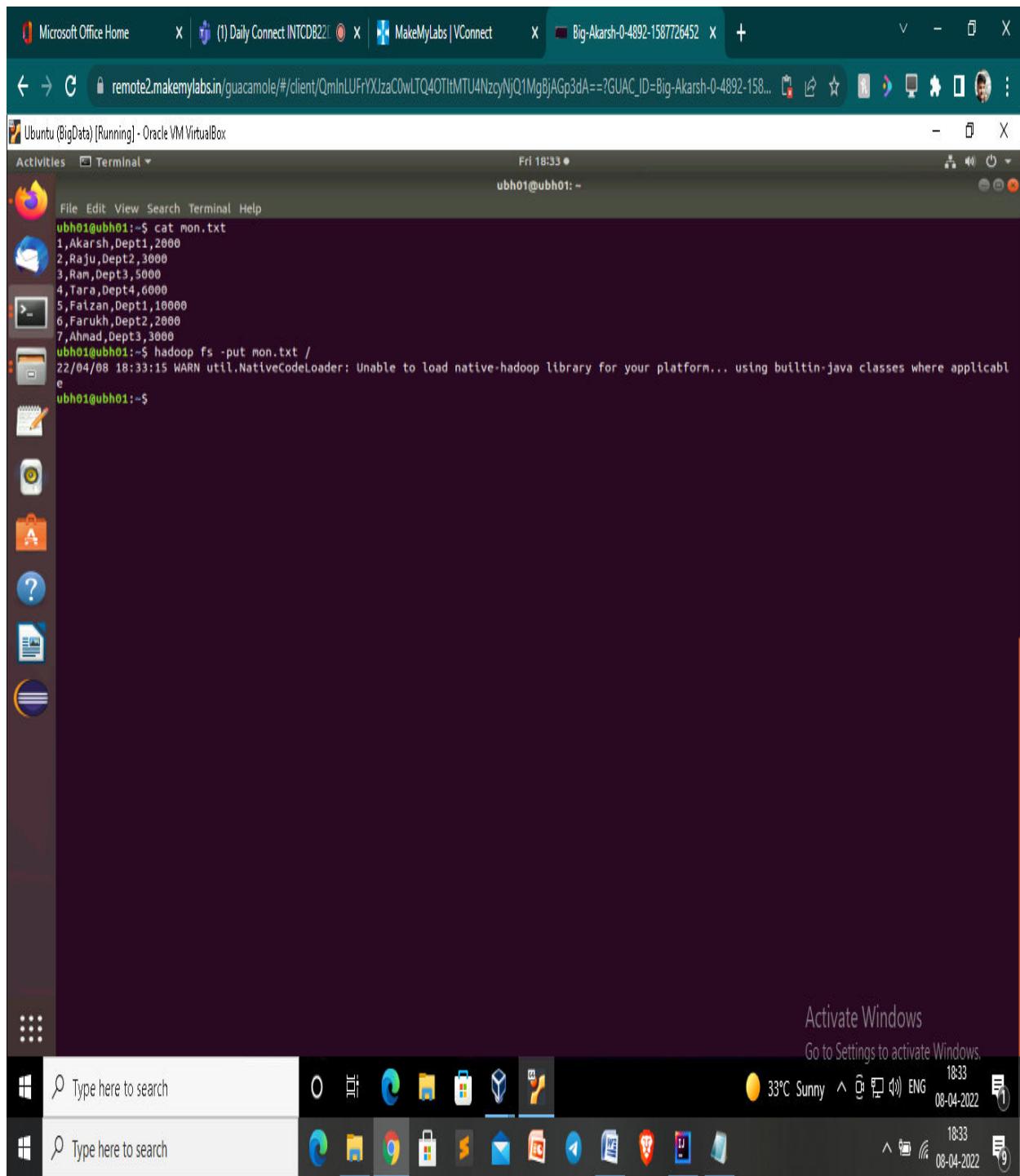
*SALARY



Step 1-> create one file name 'mon.txt'

Step 2->Add some data into it

Step 3->cat mon.txt (command to see the file content)



Step 4-> we have to put the mon.txt data into hdfs.
by using (hadoop fs –put mon.txt /)

```
Microsoft Office Home | (1) Daily Connect INTCD22 | MakeMyLabs | VConnect | Big-Akash-0-4892-1587726452 +  
Ubuntu (BigData) [Running] - Oracle VM VirtualBox  
Activities Terminal Fri 18:33 ubho1@ubho1:  
ubho1@ubho1:~$ cat mon.txt  
1,Akash,Dept1,2000  
2,Raju,Dept2,3000  
3,Ram,Dept3,5000  
4,Tara,Dept4,6000  
5,Faizan,Dept1,10000  
6,Farukh,Dept2,2000  
7,Ahmad,Dept3,3000  
ubho1@ubho1:~$ hadoop fs -put mon.txt /  
22/04/08 18:33:15 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable  
ubho1@ubho1:~$ hadoop fs -ls /  
22/04/08 18:33:29 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable  
Found 15 items  
-rw-r--r-- 1 ubh01 supergroup 45 2022-04-08 18:17 /a.txt  
-rw-r--r-- 1 ubh01 supergroup 64 2022-04-08 09:02 /f1.txt  
drwxr-xr-x 1 ubh01 supergroup 8 2022-04-08 18:25 /finalresult  
-rw-r--r-- 1 ubh01 supergroup 133 2022-04-08 18:33 /mon.txt  
-rw-r--r-- 1 ubh01 supergroup 51 2022-04-08 18:24 /money.txt  
drwxr-xr-x 1 ubh01 supergroup 0 2022-04-07 18:18 /out1  
drwxr-xr-x 1 ubh01 supergroup 0 2022-04-08 09:03 /out2  
drwxr-xr-x 1 ubh01 supergroup 0 2022-04-08 09:07 /out3  
drwxr-xr-x 1 ubh01 supergroup 0 2022-04-08 18:10 /result  
drwxr-xr-x 1 ubh01 supergroup 0 2022-04-08 18:18 /result1  
-rw-r--r-- 1 ubh01 supergroup 0 2022-04-08 18:09 /salary.txt  
drwxr-xr-x 1 ubh01 supergroup 0 2022-04-05 08:54 /ti  
drwxr-xr-x 1 ubh01 supergroup 0 2022-04-05 08:57 /time  
drwxrwx-wx 1 ubh01 supergroup 0 2022-04-07 18:15 /tmp  
drwxr-xr-x 1 ubh01 supergroup 0 2022-01-20 17:54 /user  
ubho1@ubho1:~$  
Activate Windows  
Go to Settings to activate Windows.  
18:33 33°C Sunny ENG 08-04-2022  
Type here to search | O E F G S V H  
18:33 08-04-2022
```

Step 5-> we are able to see that file in hadoop by

(hadoop fs –ls /)

```
Microsoft Office Home | (1) Daily Connect INTCDB22 | MakeMyLabs | VConnect | Big-Akarsh-0-4892-1587726452 | + | X
Ubuntu (BigData) [Running] - Oracle VM VirtualBox | Fri 18:35 • | ubh01@ubh01: ~
Activities Terminal Fri 18:35 •
ubh01@ubh01: ~
File Edit View Search Terminal Help
Map-Reduce Framework
  Map input records=7
  Map output records=7
  Map output bytes=70
  Map output materialized bytes=90
  Input split bytes=94
  Combine input records=0
  Combine output records=0
  Reduce input groups=4
  Reduce shuffle bytes=90
  Reduce input records=7
  Reduce output records=4
  Spilled Records=14
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=1250
  CPU time spent (ms)=1250
  Physical memory (bytes) snapshot=439609280
  Virtual memory (bytes) snapshot=3832619008
  Total committed heap usage (bytes)=347602944
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=133
File Output Format Counters
  Bytes Written=45
ubh01@ubh01:~$ hadoop fs -ls /
Found 16 items
-rw-r--r--  1 ubh01 supergroup      45 2022-04-07 18:17 /a.txt
-rw-r--r--  1 ubh01 supergroup     64 2022-04-08 09:02 /f1.txt
drwxr-xr-x  - ubh01 supergroup      0 2022-04-08 18:25 /finalresult
drwxr-xr-x  - ubh01 supergroup      0 2022-04-08 18:34 /fresult
-rw-r--r--  1 ubh01 supergroup    133 2022-04-08 18:33 /mon.txt
-rw-r--r--  1 ubh01 supergroup     51 2022-04-08 18:24 /money.txt
drwxr-xr-x  - ubh01 supergroup      0 2022-04-07 18:18 /out1
drwxr-xr-x  - ubh01 supergroup      0 2022-04-08 09:03 /out2
drwxr-xr-x  - ubh01 supergroup      0 2022-04-08 09:07 /out3
drwxr-xr-x  - ubh01 supergroup      0 2022-04-08 18:10 /result
drwxr-xr-x  - ubh01 supergroup      0 2022-04-08 18:18 /result1
-rw-r--r--  1 ubh01 supergroup     60 2022-04-08 18:09 /salary.txt
drwxr-xr-x  - ubh01 supergroup      0 2022-04-05 08:54 /t1
drwxr-xr-x  - ubh01 supergroup      0 2022-04-05 08:57 /time
drwxrwx-wx  - ubh01 supergroup      0 2022-04-07 18:15 /tmp
drwxr-xr-x  - ubh01 supergroup     60 2022-01-20 17:54 /user
ubh01@ubh01:~$
```

Activate Windows
Go to Settings to activate Windows.
18:35 33°C Sunny 08-04-2022 ENG
18:35 08-04-2022

Step 6->

we can put our java code into /fresult

The screenshot shows a Linux desktop environment with a terminal window open. The terminal window title is "Ubuntu (BigData) [Running] - Oracle VM VirtualBox". The terminal session is as follows:

```

Microsoft Office Home                               X | (1) Daily Connect INTCD22[ ] X | MakeMyLabs | VConnect X | Big-Akarsh-0-4892-1587726452 X + 
Fri 18:36
ubh01@ubh01: ~
File Edit View Search Terminal Help
Combine output records=0
Reduce input groups=4
Reduce shuffle bytes=90
Reduce input records=7
Reduce output records=4
Spilled Records=14
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ns)=504
CPU time spent (ms)=1250
Physical memory (bytes) snapshot=4390009280
Virtual memory (bytes) snapshot=3832619008
Total committed heap usage (bytes)=347602944
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=133
File Output Format Counters
Bytes Written=45
ubh01@ubh01:~$ hadoop fs -ls /
22/04/08 18:35:09 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 16 items
-rw-r--r-- 1 ubh01 supergroup 45 2022-04-07 18:17 /a.txt
-rw-r--r-- 1 ubh01 supergroup 64 2022-04-08 09:02 /f1.txt
drwxr-xr-x - ubh01 supergroup 0 2022-04-08 18:25 /finalresult
drwxr-xr-x - ubh01 supergroup 0 2022-04-08 18:34 /fresult
-rw-r--r-- 1 ubh01 supergroup 133 2022-04-08 18:33 /mon.txt
-rw-r--r-- 1 ubh01 supergroup 51 2022-04-08 18:24 /money.txt
drwxr-xr-x - ubh01 supergroup 0 2022-04-07 18:18 /out1
drwxr-xr-x - ubh01 supergroup 0 2022-04-08 09:03 /out2
drwxr-xr-x - ubh01 supergroup 0 2022-04-08 09:07 /out3
drwxr-xr-x - ubh01 supergroup 0 2022-04-08 18:10 /result
drwxr-xr-x - ubh01 supergroup 0 2022-04-08 18:18 /result1
-rw-r--r-- 1 ubh01 supergroup 0 2022-04-08 18:09 /salary.txt
drwxr-xr-x - ubh01 supergroup 0 2022-04-05 08:54 /t1
drwxr-xr-x - ubh01 supergroup 0 2022-04-05 08:57 /time
drwxrwxrwx - ubh01 supergroup 0 2022-04-07 18:15 /tmp
drwxr-xr-x - ubh01 supergroup 0 2022-01-28 17:54 /user
ubh01@ubh01:~$ hadoop fs -cat /fresult/*
22/04/08 18:35:52 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
e
Dept1 12000
Dept2 5000
Dept3 8000
Dept4 6000
ubh01@ubh01:~$
```

The desktop interface includes a taskbar at the bottom with a search bar, pinned icons for Microsoft Edge, File Explorer, Task View, File History, File Explorer, and File History, and a system tray showing weather (33°C Sunny), battery level (18:35), and date (08-04-2022).

Step 7-> we are able to see our output by using
(hadoop fs –cat /fresult/*)

Input->

```
ubh01@ubh01:~$ cat mon.txt
1,Akarsh,Dept1,2000
2,Raju,Dept2,3000
3,Ram,Dept3,5000
4,Tara,Dept4,6000
5,Faizan,Dept1,10000
6,Farukh,Dept2,2000
7,Ahmad,Dept3,3000
```

Output->

```
ubh01@ubh01:~$ hadoop fs -cat /fresult/*
22/04/08 18:35:52 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
Dept1 12000
Dept2 5000
Dept3 8000
Dept4 6000
```

Codes->

MyReducer.java

```
package com.example;
```

```
import org.apache.hadoop.io.IntWritable;
```

```
import org.apache.hadoop.io.Text;
```

```
import org.apache.hadoop.mapreduce.Reducer;
```

```
import java.io.IOException;

public class MyReducer extends Reducer<Text,
IntWritable,Text,IntWritable> { @Override

    protected void reduce(Text key,
Iterable<IntWritable> values, Reducer<Text,
IntWritable, Text, IntWritable>.Context context)
throws IOException, InterruptedException {

    int sum =0;for(IntWritable val :
values){sum+=val.get();}      context.write(key,new
IntWritable(sum));}}
```

The screenshot shows the IntelliJ IDEA interface with the project 'Departmentwisesalary' open. The code editor displays the `MyReducer.java` file, which contains the following code:

```
package com.example;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;
import java.io.IOException;
public class MyReducer extends Reducer<Text, IntWritable, Text, IntWritable> {
    @Override
    protected void reduce(Text key, Iterable<IntWritable> values, Reducer<Text, IntWritable, Text, IntWritable>.Context context) throws IOException, InterruptedException {
        int sum = 0;
        for(IntWritable val : values){
            sum+=val.get();
        }
        context.write(key,new IntWritable(sum));
    }
}
```

The project structure on the left shows the directory tree: `src/main/java/com/example` containing `Driver`, `MyMapper`, and `MyReducer`. The `pom.xml` file is also visible.

Driver.java

The screenshot shows the IntelliJ IDEA interface with the project 'Departmentwisesalary' open. The code editor displays the `Driver.java` file, which contains the following code:

```
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import java.io.IOException;
public class Driver {
    public static void main(String[] args) throws IOException, InterruptedException, ClassNotFoundException {
        Configuration conf = new Configuration(); //will give access to config files
        //Like core-site.xml and yarm-site.xml, hdfs-site.xml
        Job job = Job.getInstance(conf);
        //submit the job the RH and monitor the Job
        job.setJobName("Department wise Salary");
        job.setJarByClass(Driver.class);
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);
        job.setMapperClass(MyMapper.class);
        job.setCombinerClass(MyReducer.class);
        job.setReducerClass(MyReducer.class);
        //set Input path to read the data from
        FileInputFormat.addInputPath(job, new Path(args[0]));
        //set output path to store the output..i.e. wordcount
        FileOutputFormat.setOutputPath(job, new Path(args[1]));
        job.waitForCompletion(true);
        //hadoop jar name-of-jar.jar com.example.Driver /mydata/words.txt /wordcount_out
    }
}
```

The project structure on the left shows the directory tree: `src/main/java/com/example` containing `Driver`, `MyMapper`, and `MyReducer`. The `pom.xml` file is also visible.

package com.example;

```
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import
org.apache.hadoop.mapreduce.lib.input.FileInputFo
rmat;
import
org.apache.hadoop.mapreduce.lib.output.FileOutpu
tFormat;
import java.io.IOException;
public class Driver {
```

```
public static void main(String[] args) throws  
IOException, InterruptedException,  
ClassNotFoundException {  
  
    Configuration conf = new Configuration();//will  
give access to config files  
  
    //like core-site.xml and yarm-site.xml, hdfs-  
site.xml
```

```
Job job = Job.getInstance(conf);  
  
//submit the job the RM and monitor the Job  
  
job.setJobName("Department wise Salary");  
job.setJarByClass(Driver.class);  
job.setOutputKeyClass(Text.class);  
job.setOutputValueClass(IntWritable.class);
```

```
job.setMapperClass(MyMapper.class);

//    job.setCombinerClass(MyReducer.class);

job.setReducerClass(MyReducer.class);

//set Input path to read the data from
FileInputFormat.addInputPath(job, new
Path(args[0]));

//set output path to store the output..i.e.
wordcount.

FileOutputFormat.setOutputPath(job, new
Path(args[1]));
```

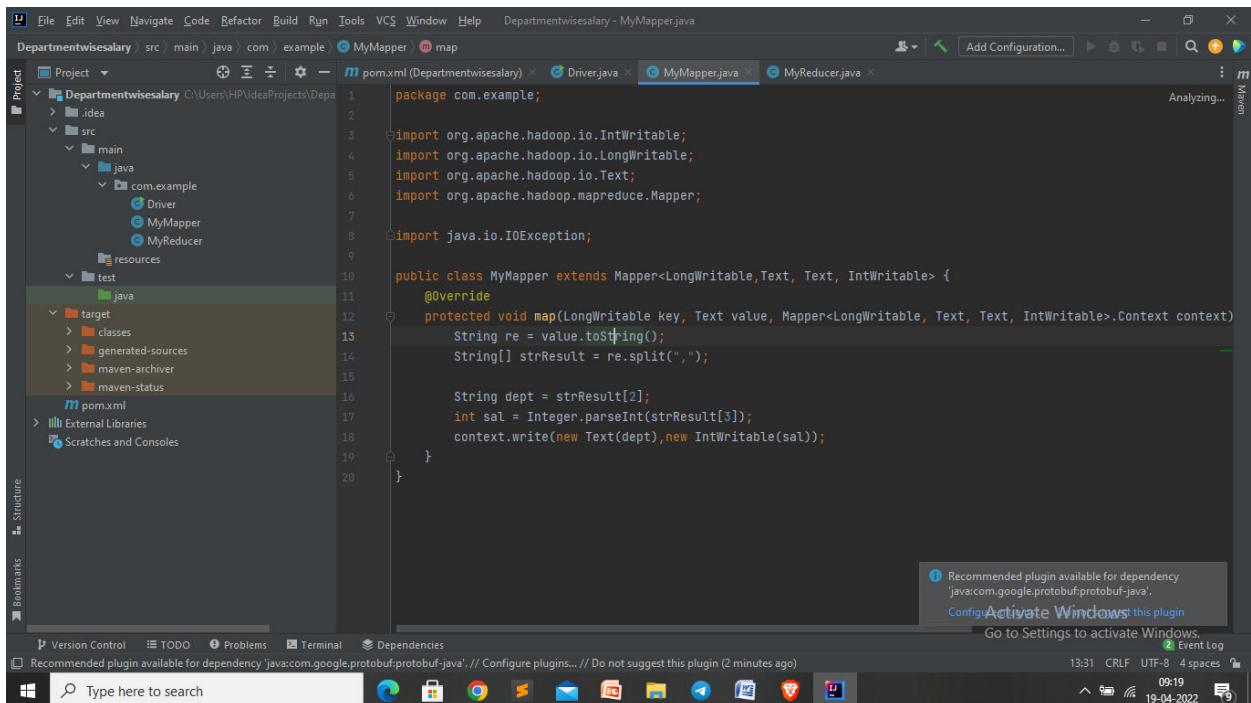
```
        job.waitForCompletion(true);

        //hadoop jar name-of-jar.jar
com.example.Driver /mydata/words.txt
/wordcount_out
```

```
}
```

```
}
```

MyMapper.java



The screenshot shows the IntelliJ IDEA interface with the 'MyMapper.java' file open in the editor. The code implements a custom Mapper class:

```
package com.example;

import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;

import java.io.IOException;

public class MyMapper extends Mapper<LongWritable, Text, Text, IntWritable> {
    @Override
    protected void map(LongWritable key, Text value, Mapper<LongWritable, Text, Text, IntWritable>.Context context)
            throws IOException {
        String re = value.toString();
        String[] strResult = re.split(",");
        String dept = strResult[2];
        int sal = Integer.parseInt(strResult[3]);
        context.write(new Text(dept), new IntWritable(sal));
    }
}
```

The project structure on the left shows a Maven project named 'Departmentwisesalary' with modules 'Driver', 'MyMapper', and 'MyReducer'. The 'MyMapper' module is currently selected. A tooltip at the bottom right indicates a recommended plugin for protobuf support.

```
package com.example;
```

```
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;
import java.io.IOException;

public class MyMapper extends
Mapper<LongWritable,Text, Text, IntWritable> {

@Override
protected void map(LongWritable key, Text value,
Mapper<LongWritable, Text, Text,
IntWritable>.Context context) throws IOException,
InterruptedException {
String re = value.toString();
String[] strResult = re.split(",");
}
```

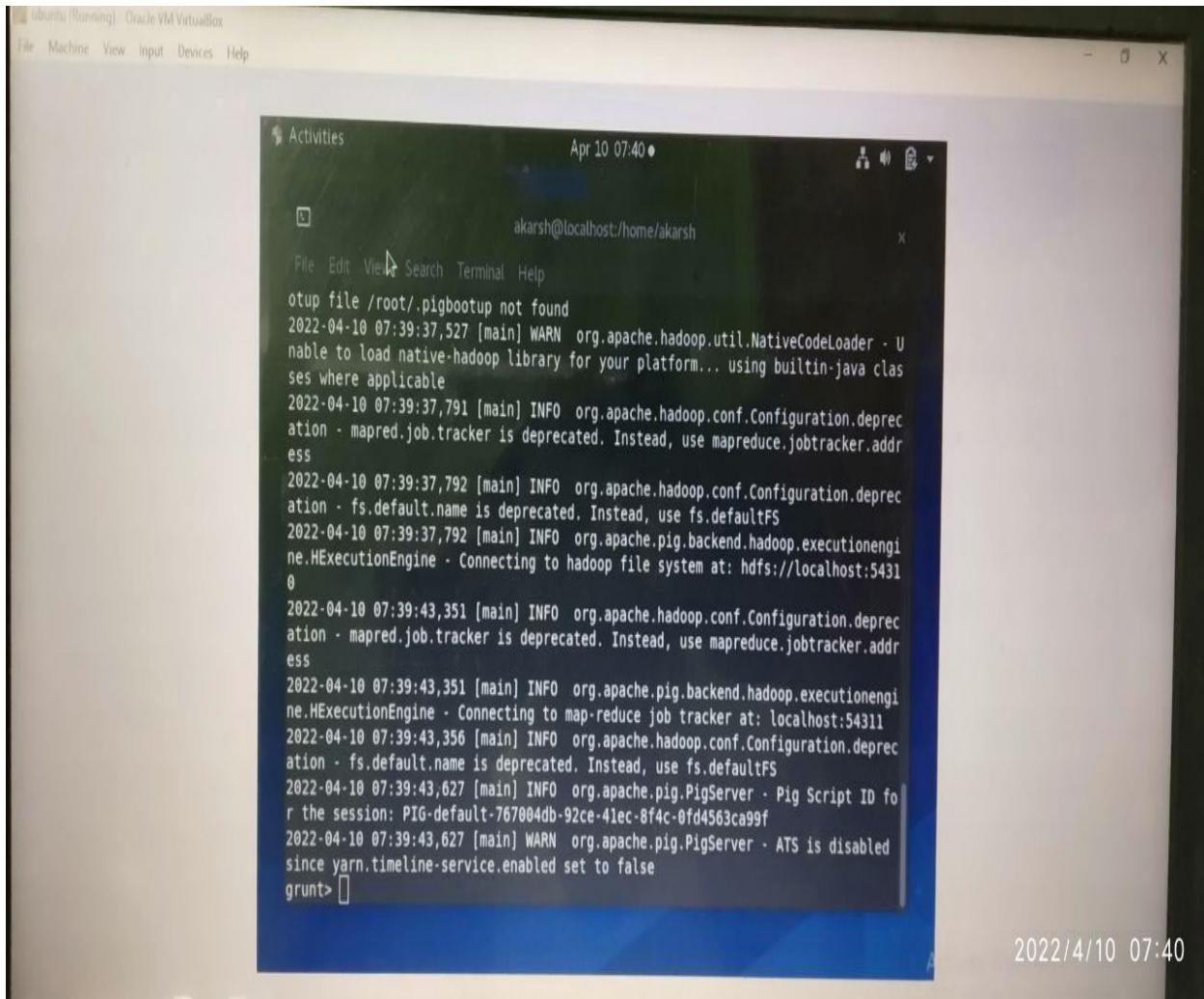
```
String dept = strResult[2];  
int sal = Integer.parseInt(strResult[3]);  
context.write(new Text(dept),new  
IntWritable(sal));  
}  
}
```

ASSIGNMENT NO. 2

PROBLEM STATEMENT ->Top 10 Customers by total sales.

SUB TOPIC -> USING PIG

Data – ‘txns1.txt’



The screenshot shows a terminal window titled 'Activities' running on an Ubuntu system. The window title bar says 'ubuntu (Running) Oracle VM VirtualBox'. The terminal window has a dark background and displays the following text:

```
akarsh@localhost:~$ pig -x local
2022-04-10 07:39:37,527 [main] WARN org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2022-04-10 07:39:37,791 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-04-10 07:39:37,792 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-04-10 07:39:37,792 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:54310
2022-04-10 07:39:43,351 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-04-10 07:39:43,351 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to map-reduce job tracker at: localhost:54311
2022-04-10 07:39:43,356 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-04-10 07:39:43,627 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-767004db-92ce-41ec-8f4c-0fd4563ca99f
2022-04-10 07:39:43,627 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt>
```

The terminal window is titled 'akarsh@localhost:~\$'. The date and time 'Apr 10 07:40' are visible at the top right of the terminal window. The bottom right corner of the terminal window shows the date and time '2022/4/10 07:40'.

ming - Oracle VM VirtualBox

View Input Devices Help

Activities Terminal Apr 10 08:16

akarsh@localhost:/home/akarsh

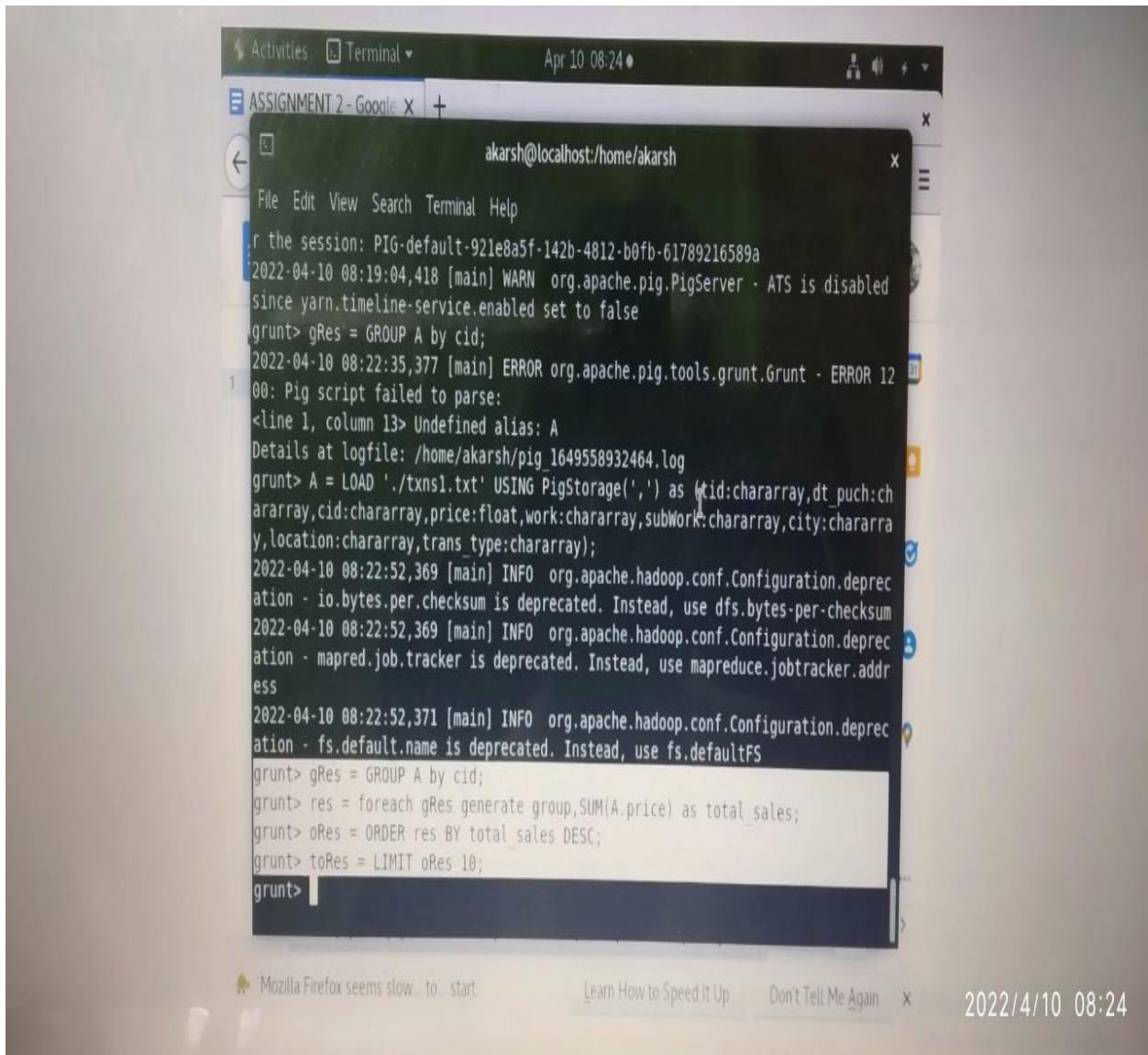
```
File Edit View Search Terminal Help
inputFormat - Total input files to process : 1
2022-04-10 08:12:58,422 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(4009485,1973.3000144958496)
(4006425,1732.0900039672852)
(4000221,1671.4700012207031)
(4003228,1640.629981994629)
(4006606,1628.9400157928467)
(4006467,1605.9499855041504)
(4004927,1576.710018157959)
(4008321,1560.7899951934814)
(4000815,1557.8200035095215)
(4001051,1488.6699829101562)
grunt> A = LOAD './txns1.txt' USING PigStorage(',') as (tid:chararray,dt:puch:chararray,cid:chararray,price:float,work:chararray,subWork:chararray,city:chararray,location:chararray,trans_type:chararray);
2022-04-10 08:15:29,129 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-04-10 08:15:29,130 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-04-10 08:15:29,130 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt>
```

txns1.txt var Vid 100% selected of 4 MB

2022/4/10 08:16

Loading the data :-

```
grunt> A = LOAD './txns1.txt' USING PigStorage(',') as  
(tid:chararray,dt_puch:chararray,cid:chararray,price:float,work:chararray,subWork:chararray,city  
:chararray,location:chararray,trans_type:chararray);
```



The screenshot shows a terminal window titled "ASSIGNMENT 2 - Google" with the user "akarsh@localhost:/home/akarsh". The terminal displays the following Apache Pig session:

```
File Edit View Search Terminal Help  
r the session: PIG-default-921e8a5f-142b-4812-b0fb-61789216589a  
2022-04-10 08:19:04,418 [main] WARN org.apache.pig.PigServer - ATS is disabled  
since yarn.timeline-service.enabled set to false  
grunt> gRes = GROUP A by cid;  
2022-04-10 08:22:35,377 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 12  
00: Pig script failed to parse:  
<line 1, column 13> Undefined alias: A  
Details at logfile: /home/akarsh/pig_1649558932464.log  
grunt> A = LOAD './txns1.txt' USING PigStorage(',') as (tid:chararray,dt_puch:ch  
ararray,cid:chararray,price:float,work:chararray,subWork:chararray,city:chararra  
y,location:chararray,trans_type:chararray);  
2022-04-10 08:22:52,369 [main] INFO org.apache.hadoop.conf.Configuration.deprec  
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum  
2022-04-10 08:22:52,369 [main] INFO org.apache.hadoop.conf.Configuration.deprec  
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.add  
ress  
2022-04-10 08:22:52,371 [main] INFO org.apache.hadoop.conf.Configuration.deprec  
ation - fs.default.name is deprecated. Instead, use fs.defaultFS  
grunt> gRes = GROUP A by cid;  
grunt> res = foreach gRes generate group,SUM(A.price) as total_sales;  
grunt> oRes = ORDER res BY total_sales DESC;  
grunt> toRes = LIMIT oRes 10;  
grunt>
```

Doing some operation:-

```
grunt> gRes = GROUP A by cid;  
grunt> res = foreach gRes generate group,SUM(A.price) as total_sales;  
grunt> oRes = ORDER res BY total_sales DESC;  
grunt> toRes = LIMIT oRes 10;  
grunt> dump toRes
```

```
File Edit View Search Terminal Help  
2022-04-10 08:25:10,575 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS  
2022-04-10 08:25:10,575 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address  
2022-04-10 08:25:10,576 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum  
2022-04-10 08:25:10,577 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized  
2022-04-10 08:25:10,831 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1  
2022-04-10 08:25:10,831 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1  
(4009485,1973.3000144958496)  
(4006425,1732.0900039672852)  
(4000221,1671.4700012207031)  
(4003228,1640.629981994629)  
(4006606,1628.9400157928467)  
(4006467,1605.9499855041504)  
(4004927,1576.710018157959)  
(4008321,1560.7899951934814)  
(4000815,1557.8200035095215)  
(4001051,1488.6699829101562)  
grunt>
```

Output:-

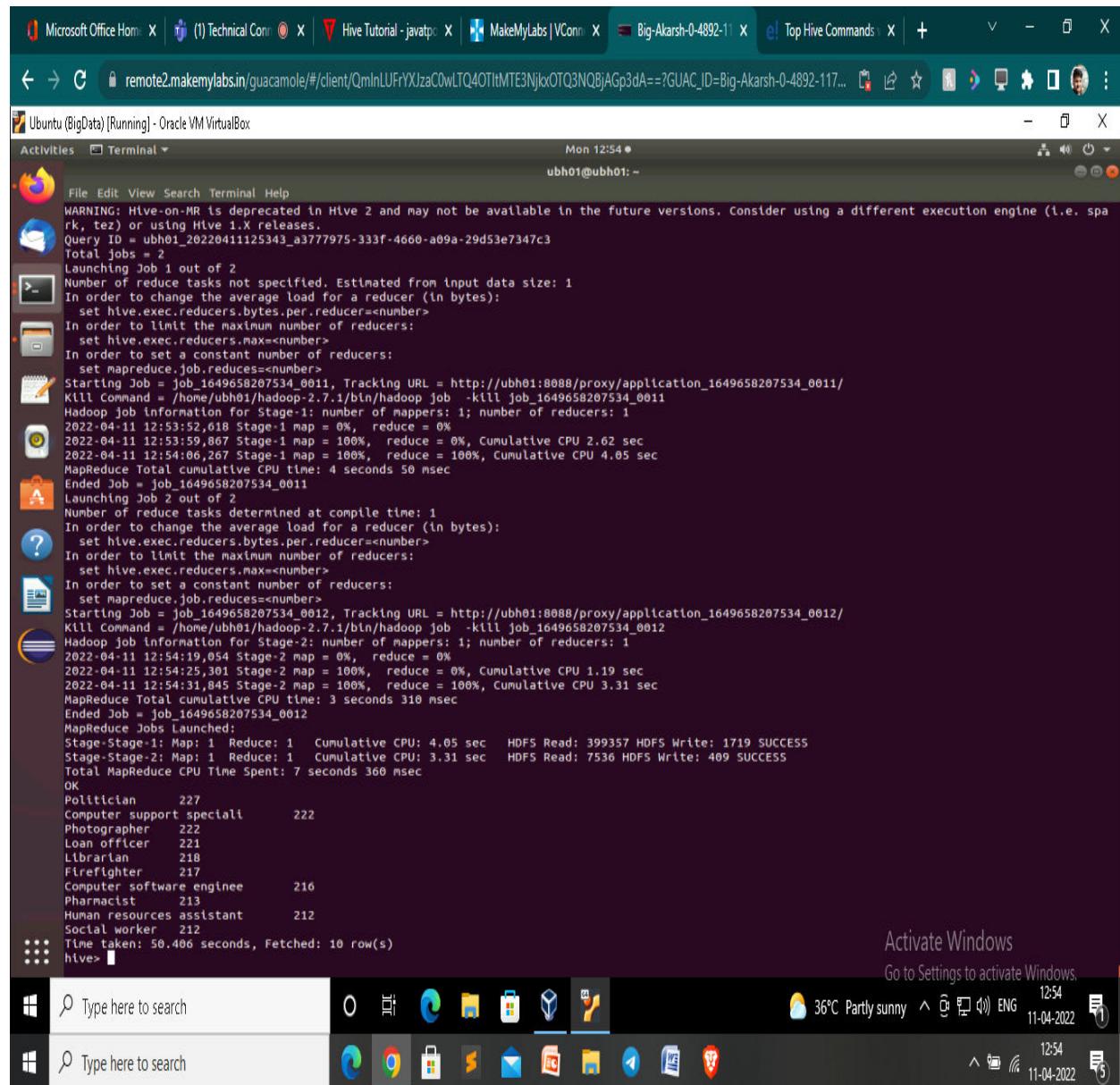
```
2022-04-10 08:12:58,257 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning ACCESSING_NON_EXISTENT_FIELD 8 time(s).
2022-04-10 08:12:58,257 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-04-10 08:12:58,296 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-04-10 08:12:58,296 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-04-10 08:12:58,297 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-04-10 08:12:58,298 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-04-10 08:12:58,421 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-04-10 08:12:58,422 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(4009485,1973.3000144958496)
(4006425,1732.0900039672852)
(4000221,1671.4700012207031)
(4003228,1640.629981994629)
(4006606,1628.9400157928467)
(4006467,1605.9499855041504)
(4004927,1576.710018157959)
(4008321,1560.7899951934814)
(4000815,1557.8200035095215)
(4001051,1488.6699829101562)
```

ASSIGNMENT NO.3

PROBLEM STATEMENT->Top 10 designation with people count

SUB TOPIC->USING HIVE

Data -> 'custs.txt'



The screenshot shows a terminal window on a Linux desktop environment. The terminal is running on an Oracle VM VirtualBox machine named 'Ubuntu (BigData) [Running]'. The window title is 'remote2.makemylabs.in/guacamole/#/client/QmlnLUFrYXJzaC0wLTQ4OTltMTE3NjlxOTQ3NQBjAGp3dA==?GUAC_ID=Big-Akarsh-0-4892-117...'. The terminal session is connected via SSH to the host machine, with the prompt 'ubh01@ubh01: ~'.

The terminal output displays the execution of a Hive query on a dataset named 'custs.txt'. The query retrieves the top 10 rows from the dataset. The output shows the following results:

```
File Edit View Search Terminal Help
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID: ubh01_2022041125343_a3777975-333f-4660-a09a-29d53e7347c3
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1649658207534_0011, Tracking URL = http://ubh01:8088/proxy/application_1649658207534_0011/
Kill Command = /home/ubh01/hadoop-2.7.1/bin/hadoop job -kill job_1649658207534_0011
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-04-11 12:53:52,618 Stage-1 map = 0%, reduce = 0%
2022-04-11 12:53:59,807 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.62 sec
2022-04-11 12:54:06,267 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 4.05 sec
MapReduce Total cumulative CPU time: 4 seconds 50 msec
Ended Job = job_1649658207534_0011
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1649658207534_0012, Tracking URL = http://ubh01:8088/proxy/application_1649658207534_0012/
Kill Command = /home/ubh01/hadoop-2.7.1/bin/hadoop job -kill job_1649658207534_0012
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-04-11 12:54:19,054 Stage-2 map = 0%, reduce = 0%
2022-04-11 12:54:25,301 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.19 sec
2022-04-11 12:54:31,845 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 3.31 sec
MapReduce Total cumulative CPU time: 3 seconds 310 msec
Ended Job = job_1649658207534_0012
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 4.05 sec HDFS Read: 399357 HDFS Write: 1719 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 3.31 sec HDFS Read: 7536 HDFS Write: 409 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 360 msec
OK
Politician      227
Computer support speciali    222
Photographer     222
Loan officer     221
Librarian        218
Firefighter      217
Computer software engineer  216
Pharmacist       213
Human resources assistant 212
Social worker    212
Time taken: 50.406 seconds, Fetched: 10 row(s)
hive>
```

The terminal window is part of a desktop environment with a taskbar at the bottom. The taskbar shows several open applications, including Microsoft Office Home, Technical Conn, Hive Tutorial - javatpoint, MakeMyLabs | VConn, Big-Akarsh-0-4892-1, Top Hive Commands, and a browser window. The system tray indicates the date as 11-04-2022, the time as 12:54, and the weather as 36°C Partly sunny. The language setting is ENG.

```
hive> load data local inpath '/home/ubh01/custs.txt' into table db.cu;
```

```
hive> create table cu(cid integer,fname varchar(15),lname varchar(15),age integer,designation
varchar(25))
      > row format delimited
      > fields terminated by ','
      > stored as textfile;
```

*) Without handling null designation

```
hive> select designation, count(designation) as total from db.cu group by designation order by
total desc limit 10;
```

*) BY handling null designation ->

```
hive> SELECT case WHEN designation = " THEN 'Unknown' ELSE designation END as
designations,COUNT(*) total FROM db.cu GROUP BY designation ORDER BY total DESC
LIMIT 10;
```

```
Kill Command = /home/ubh01/hadoop-2.7.1/bin/hadoop job -kill job_1649658207534_0010
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-04-11 12:46:25,743 Stage-2 map = 0%, reduce = 0%
2022-04-11 12:46:31,122 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.41 sec
2022-04-11 12:46:39,662 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 5.92 sec
MapReduce Total cumulative CPU time: 5 seconds 920 msec
Ended Job = job_1649658207534_0010
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 9.01 sec HDFS Read: 399357 HDFS
Write: 1719 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 5.92 sec HDFS Read: 7536 HDFS
Write: 409 SUCCESS
Total MapReduce CPU Time Spent: 14 seconds 930 msec
OK
Politician 227
Computer support speciali 222
Photographer 222
Loan officer 221
Librarian 218
Firefighter 217
Computer software enginee 216
Pharmacist 213
Human resources assistant 212
```

Social worker 212
Time taken: 52.651 seconds, Fetched: 10 row(s)

ASSIGNMENT NO.4

PROBLEM STATEMENT->Create a login table and put data into it and show the output

SUB TOPIC-> USING hbase

Ans4->

to start the hdfs and hbase shell

```
ubh01@ubh01:~$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
22/04/13 14:32:54 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /home/ubh01/hadoop-2.7.1/logs/hadoop-ubh01-
namenode-ubh01.out
localhost: starting datanode, logging to /home/ubh01/hadoop-2.7.1/logs/hadoop-ubh01-
datanode-ubh01.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/ubh01/hadoop-2.7.1/logs/hadoop-
ubh01-secondarynamenode-ubh01.out
22/04/13 14:33:16 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /home/ubh01/hadoop-2.7.1/logs/yarn-ubh01-
resourcemanager-ubh01.out
localhost: starting nodemanager, logging to /home/ubh01/hadoop-2.7.1/logs/yarn-ubh01-
nodemanager-ubh01.out
ubh01@ubh01:~$ start-hbase.sh
starting master, logging to /home/ubh01/hbase-1.1.2/logs/hbase-ubh01-master-ubh01.out
```

OpenJDK 64-Bit Server VM warning: ignoring option PermSize=128m; support was removed in 8.0

OpenJDK 64-Bit Server VM warning: ignoring option MaxPermSize=128m; support was removed in 8.0

ubh01@ubh01:~\$ hbase shell

SLF4J: Class path contains multiple SLF4J bindings.

SLF4J: Found binding in [jar:file:/home/ubh01/hbase-1.1.2/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]

SLF4J: Found binding in [jar:file:/home/ubh01/hadoop-2.7.1/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.slf4j.impl.Log4jLoggerFactory]

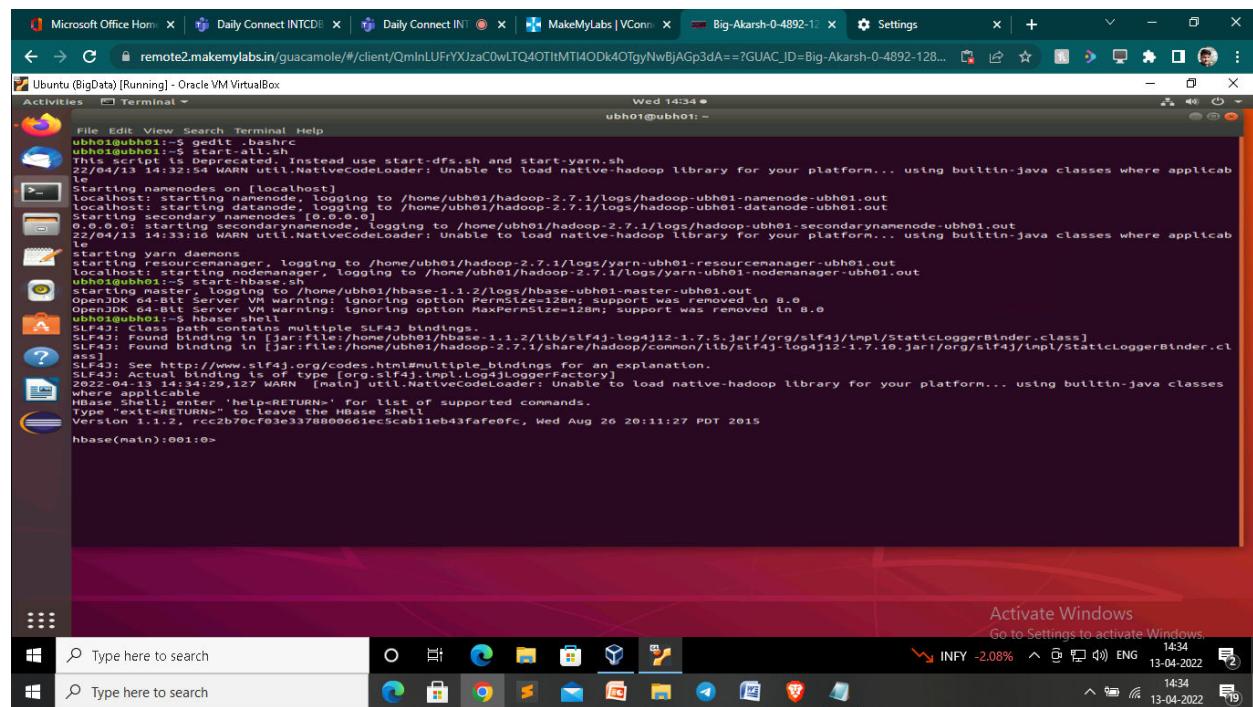
2022-04-13 14:34:29,127 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

HBase Shell; enter 'help<RETURN>' for list of supported commands.

Type "exit<RETURN>" to leave the HBase Shell

Version 1.1.2, rcc2b70cf03e3378800661ec5cab11eb43fafe0fc, Wed Aug 26 20:11:27 PDT 2015

hbase(main):001:0>



```

Microsoft Office Home & Business | Daily Connect INTCD | Daily Connect INT | MakeMyLabs | VConn | Big-Akarsh-0-4892-12 | Settings | + | - | X
← → C 🔒 remote2.makemylabs.in/guacamole/#/client/QmlnLUFrYXJzaC0wLTQ4OTltMTI4ODk4OTgyNwBjAGp3dA==?GUAC_ID=Big-Akarsh-0-4892-12...
Activities Terminal Wed 14:43 ubh01@ubh01 ~
Ubuntu (BigData) [Running] - Oracle VM VirtualBox
File Edit View Search Terminal Help
SLF4J: Found binding in [jar:file:/home/ubh01/hadoop-2.7.1/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/tmpl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
2022-04-13 14:34:29,327 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes w
here applicable
Type "exit<RETURN>" to leave the HBase Shell
Version 1.1.2, rcc2b70cf03e3378800061ec5cab1e43fafe0fc, Wed Aug 26 20:11:27 PDT 2015
hbase(main):001:0> list
TABLE
0 row(s) in 0.3350 seconds
hbase(main):002:0> create 'login','per','prof','misc'
0 row(s) in 1.3090 seconds
==> Hbase::Table - login
hbase(main):003:0> list
TABLE
1 row(s)
1 row(s) in 0.0150 seconds
==> ["login"]
hbase(main):004:0> put 'login','row-001','per:fname','Akarsh'
0 row(s) in 0.2770 seconds
hbase(main):005:0> put 'login','row-001','per:lname','Raj'
0 row(s) in 0.0520 seconds
hbase(main):006:0> put 'login','row-001','per:city','Punjab'
0 row(s) in 0.0140 seconds
hbase(main):007:0> scan 'login'
ROW
  COLUMN+CELL
row-001          column=per:city, timestamp=1649841048264, value=Punjab
row-001          column=per:fname, timestamp=1649841012640, value=Akarsh
row-001          column=per:lname, timestamp=1649841029937, value=Raj
1 row(s) in 0.0750 seconds
hbase(main):008:0> put 'login','row-001','misc:hobby','Playing Table'
0 row(s) in 0.0160 seconds
hbase(main):009:0> scan 'login'
ROW
  COLUMN+CELL
row-001          column=misc:hobby, timestamp=1649841152874, value=Playing Table
row-001          column=per:city, timestamp=1649841048264, value=Punjab
row-001          column=per:fname, timestamp=1649841012640, value=Akarsh
row-001          column=per:lname, timestamp=1649841029937, value=Raj
1 row(s) in 0.0740 seconds
hbase(main):010:0>

```

Activate Windows
Go to Settings to activate Windows.

After creation of login table and by putting some values

```

hbase(main):028:0> scan 'login'
ROW                               COLUMN+CELL
row-001                           column=misc:hobby, timestamp=1649841152874, value=Playing
Tabla                            column=per:city, timestamp=1649841048264, value=Punjab
row-001                           column=per:fname, timestamp=1649841012640, value=Akarsh
row-001                           column=per:lname, timestamp=1649841029937, value=Raj
row-001                           column=prof:degree, timestamp=1649841391432, value=Data
Scientist                         column=misc:hobby, timestamp=1649841764749, value=Playing
row-002                           column=per:city, timestamp=1649841428104, value=Bokaro
Football                          column=per:fname, timestamp=1649841307837, value=Ashish
row-002                           column=per:lname, timestamp=1649841330584, value=Mukherjee
row-002                           column=prof:degree, timestamp=1649841566431, value=IAS
row-003                           column=misc:hobby, timestamp=1649841749835, value=Yoga

```

```
row-003          column=per:fname, timestamp=1649841524007, value=Rahul
row-003          column=per:lname, timestamp=1649841534970, value=Singh
row-003          column=prof:degree, timestamp=1649841722889,
value=Engineer
3 row(s) in 0.0640 seconds
```

If we want to delete some data

```
hbase(main):004:0> delete 'login','row-001','misc:hobby'
0 row(s) in 0.0740 seconds
```

```
hbase(main):005:0> scan 'login'
ROW                  COLUMN+CELL
row-001          column=per:city, timestamp=1649841048264, value=Punjab
row-001          column=per:fname, timestamp=1649841012640, value=Akarsh
row-001          column=per:lname, timestamp=1649841029937, value=Raj
row-001          column=prof:degree, timestamp=1649841391432, value=Data
Scientist          column=misc:hobby, timestamp=1649841764749, value=Playing
row-002          column=per:city, timestamp=1649841428104, value=Bokaro
row-002          column=per:fname, timestamp=1649841307837, value=Ashish
row-002          column=per:lname, timestamp=1649841330584, value=Mukherjee
row-002          column=prof:degree, timestamp=1649841566431, value=IAS
row-003          column=misc:hobby, timestamp=1649841749835, value=Yoga
row-003          column=per:fname, timestamp=1649841524007, value=Rahul
row-003          column=per:lname, timestamp=1649841534970, value=Singh
row-003          column=prof:degree, timestamp=1649841722889,
value=Engineer
3 row(s) in 0.1910 seconds
```

```

Microsoft Office Home & Business | Daily Connect INTCDI | Daily Connect INT | MakeMyLabs | VConn | Big-Akarsh-0-4892-12 | Settings
← → C 🔍 remote2.makemylabs.in/guacamole/#/client/QmlnLUFrYXJzaCwLTQ4OTltMTI4ODk4OTgyNwBjAGp3dA==?GUAC_ID=Big-Akarsh-0-4892-12...
Activities Terminal Wed 15:19 ubh01@ubh01: ~
Ubuntu (BigData) [Running] - Oracle VM VirtualBox
File Edit View Search Terminal Help
hbase(nmtn):001:0> scan 'Login'
ROW COLUMN+CELL
row-001 column=misc:hobby, timestamp=1649841152874, value=Playing Table
row-001 column=per:fname, timestamp=1649841012640, value=Akarsh
row-001 column=per:lname, timestamp=1649841029937, value=Raj
row-001 column=prof:degree, timestamp=16498411391432, value=Data Scientist
row-002 column=misc:hobby, timestamp=1649841764749, value=Playing Football
row-002 column=per:city, timestamp=1649841428104, value=Bokaro
row-002 column=per:fname, timestamp=1649841307837, value=Ashish
row-002 column=per:lname, timestamp=1649841330584, value=Mukherjee
row-002 column=prof:degree, timestamp=1649841566431, value=IAS
row-003 column=misc:hobby, timestamp=16498411534970, value=Yoga
row-003 column=per:fname, timestamp=1649841524007, value=Rahul
row-003 column=prof:degree, timestamp=1649841722889, value=Engineer
3 row(s) in 0.2850 seconds
hbase(nmtn):002:0> get 'login','row-001'
COLUMN CELL
misc:hobby timestamp=1649841152874, value=Playing Table
per:city timestamp=1649841048264, value=Punjab
per:fname timestamp=1649841012640, value=Akarsh
per:lname timestamp=1649841029937, value=Raj
prof:degree timestamp=16498411391432, value=Data Scientist
1 row(s) in 0.0320 seconds
hbase(nmtn):003:0> get 'login','row-001','misc:hobby'
COLUMN CELL
misc:hobby timestamp=1649841152874, value=Playing Table
1 row(s) in 0.0270 seconds
hbase(nmtn):004:0> delete 'login','row-001','misc:hobby'
0 row(s) in 0.0740 seconds
hbase(nmtn):005:0> scan 'login'
ROW COLUMN+CELL
row-001 column=per:city, timestamp=1649841048264, value=Punjab
row-001 column=per:fname, timestamp=1649841012640, value=Akarsh
row-001 column=per:lname, timestamp=1649841029937, value=Raj
row-001 column=prof:degree, timestamp=16498411391432, value=Data Scientist
row-002 column=misc:hobby, timestamp=1649841764749, value=Playing Football
row-002 column=per:cityV, timestamp=1649841428104, value=Bokaro
row-002 column=per:fname, timestamp=1649841307837, value=Ashish
row-002 column=per:lname, timestamp=1649841330584, value=Mukherjee
row-002 column=prof:degree, timestamp=1649841566431, value=IAS
row-003 column=misc:hobby, timestamp=16498411534970, value=Yoga
row-003 column=per:fname, timestamp=1649841524007, value=Rahul
row-003 column=prof:degree, timestamp=1649841722889, value=Engineer
3 row(s) in 0.1910 seconds
hbase(nmtn):006:0>

```

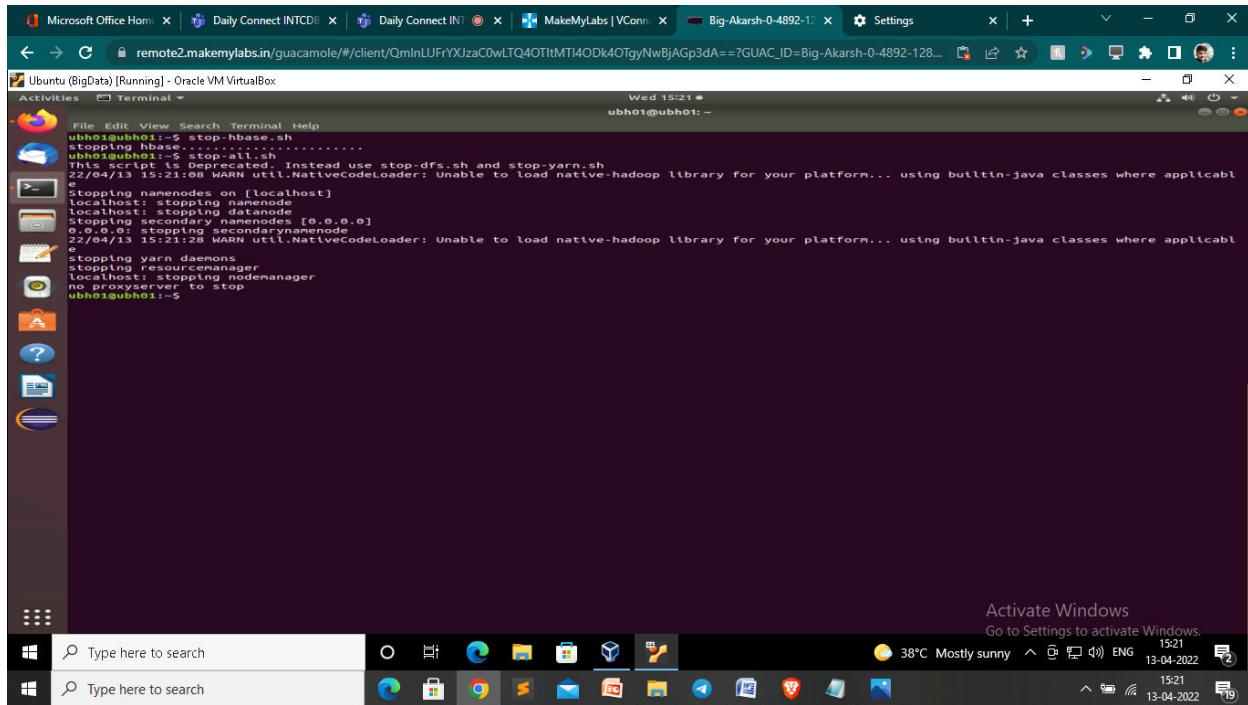
To stop hbase

```

ubh01@ubh01:~$ stop-hbase.sh
stopping hbase.....
ubh01@ubh01:~$ stop-all.sh
This script is Deprecated. Instead use stop-dfs.sh and stop-yarn.sh
22/04/13 15:21:08 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
Stopping namenodes on [localhost]
localhost: stopping namenode
localhost: stopping datanode
Stopping secondary namenodes [0.0.0.0]
0.0.0.0: stopping secondarynamenode

```

```
22/04/13 15:21:28 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
stopping yarn daemons
stopping resourcemanager
localhost: stopping nodemanager
no proxyserver to stop
```



ASSIGNMENT NO.5

TOPIC->Sqoop Program

SUB TOPIC -> USING SQQOP

i) Use one mapper

```
mysql> use akarsh_db;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A
```

```
Database changed
mysql> select * from emp;
+-----+-----+
| empid | ename | salary |
+-----+-----+
| 101 | Akarsh | 3000 |
| 102 | Raj | 4000 |
| 103 | Ram | 5000 |
| 104 | Ramesh | 6000 |
| 105 | Rohan | 7000 |
+-----+-----+
5 rows in set (0.00 sec)
```

```
ubh01@ubh01:~$ sqoop import \
> --connect jdbc:mysql://localhost/akarsh_db \
> --username root \
> --password password \
> --table emp \
> -m 1 \
> --target-dir /sqoop_import_01
Warning: /home/ubh01/sqoop-1.4.7.bin__hadoop-2.6.0/..hcatalog does not exist! HCatalog jobs
will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/ubh01/sqoop-1.4.7.bin__hadoop-2.6.0/..accumulo does not exist! Accumulo
imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /home/ubh01/sqoop-1.4.7.bin__hadoop-2.6.0/..zookeeper does not exist! Accumulo
imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
22/04/18 15:47:28 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
```

```

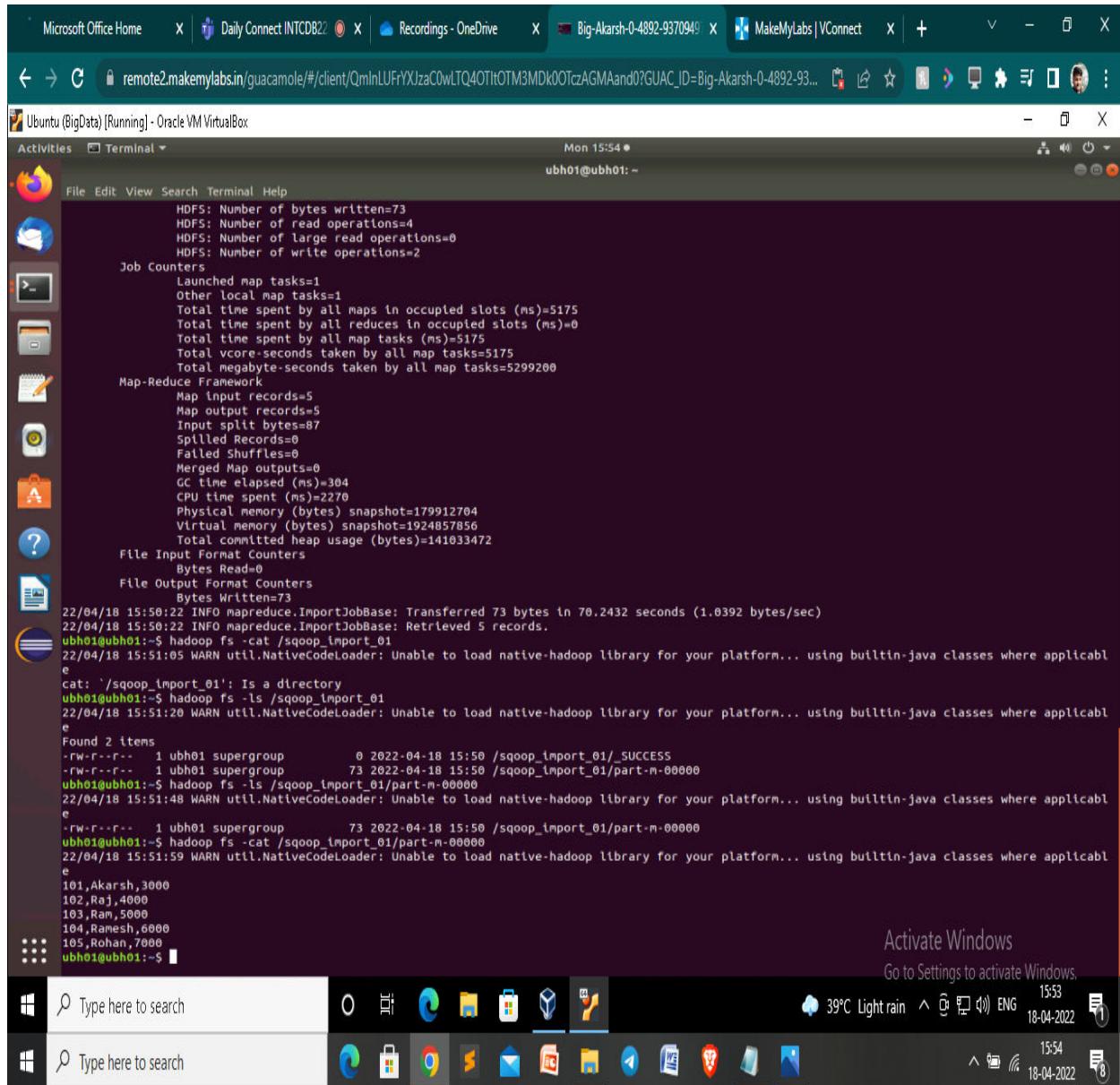
ubh01@ubh01:~$ hadoop fs -cat /sqoop_import_01/part-m-00000
22/04/18 15:51:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
101,Akarsh,3000
102,Raj,4000
103,Ram,5000
104,Ramesh,6000
105,Rohan,7000

```

```

Ubuntu (BigData) [Running] - Oracle VM VirtualBox
Activities Terminal Mon 15:53 ubh01@ubh01: ~
ubh01@ubh01:~$ sqoop import \
> --connect jdbc:mysql://localhost/akarsh_db \
> --username root \
> --password password \
> --table emp \
> -n 1 \
--target-dir /sqoop_import_01
Warning: /home/ubh01/sqoop-1.4.7-bin_hadoop-2.6.0/..hcatalog does not exist! HCatalog jobs will fail.
Please set SPAT_HOME to the root of your HCatalog installation.
Warning: /home/ubh01/sqoop-1.4.7-bin_hadoop-2.6.0/..accumulo does not exist! Accumulo imports will fail.
Please set SACUMULO_HOME to the root of your Accumulo installation.
Warning: /home/ubh01/sqoop-1.4.7-bin_hadoop-2.6.0/..zookeeper does not exist! Accumulo imports will fail.
Please set ZOOKEEPER_HOME to the root of your Zookeeper installation.
22/04/18 15:47:28 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
22/04/18 15:47:28 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/04/18 15:47:29 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/04/18 15:47:29 INFO tool.CodeGenTool: Beginning code generation
Loading class 'com.mysql.jdbc.Driver'. This is deprecated. The new driver class is 'com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and manual loading of the driver class is generally unnecessary.
22/04/18 15:47:29 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `emp` AS t LIMIT 1
22/04/18 15:47:29 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `emp` AS t LIMIT 1
22/04/18 15:47:29 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /home/ubh01/hadoop-2.7.1
Note: /tmp/sqoop-ubh01/compile/B48Bbe35af7e8736875eb551fa4b892e8/emp.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/04/18 15:49:05 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-ubh01/compile/848Bbe35af7e8736875eb551fa4b892e8/emp.jar
22/04/18 15:49:05 WARN manager.MySQLManager: It looks like you are importing from mysql.
22/04/18 15:49:05 WARN manager.MySQLManager: This transfer can be faster! Use the -direct
22/04/18 15:49:05 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
22/04/18 15:49:05 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/ubh01/hadoop-2.7.1/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/ubh01/base-1.1.2/lib/slf4j-log4j12-1.7.5.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.slf4j.impl.Log4jLoggerFactory].
22/04/18 15:49:06 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
22/04/18 15:49:06 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
22/04/18 15:49:11 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
22/04/18 15:49:12 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
22/04/18 15:49:140 INFO db.DBInputFormat: Using read committed transaction isolation
22/04/18 15:49:141 INFO mapreduce.JobSubmitter: number of splits:1
22/04/18 15:49:149 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1650276725951_0001
22/04/18 15:49:56 INFO impl.YarnClientImpl: Submitted application application_1650276725951_0001
22/04/18 15:49:57 INFO mapreduce.Job: The url to track the job: http://ubh01:8088/proxy/application_1650276725951_0001/
22/04/18 15:49:57 INFO mapreduce.Job: Running job: job_1650276725951_0001
22/04/18 15:50:12 INFO mapreduce.Job: Job job_1650276725951_0001 running in uber mode : false
22/04/18 15:50:12 INFO mapreduce.Job: map 0% reduce 0%
22/04/18 15:56:21 INFO mapreduce.Job: map 100% reduce 0%
22/04/18 15:56:21 INFO mapreduce.Job: Job job_1650276725951_0001 completed successfully
22/04/18 15:56:22 INFO mapreduce.Job: Counters: 30
          file system counters
Activate Windows
Go to Settings to activate Windows.
15:53 39°C Light rain ENG 18-04-2022
15:53 18-04-2022

```



ii) Use two Mappers

```
ubh01@ubh01:~$ sqoop import \
> --connect jdbc:mysql://localhost/akarsh_db \
> --username root \
> --password password \
> --table emp \
> --split-by empid \
> -m 2 \
> --target-dir /hadoop_test_14
Warning: /home/ubh01/sqoop-1.4.7.bin__hadoop-2.6.0/..hcatalog does not exist! HCatalog jobs
will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/ubh01/sqoop-1.4.7.bin__hadoop-2.6.0/..accumulo does not exist! Accumulo
imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /home/ubh01/sqoop-1.4.7.bin__hadoop-2.6.0/..zookeeper does not exist! Accumulo
imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation
```

```
ubh01@ubh01:~$ hadoop fs -ls /hadoop_test_14
22/04/18 15:59:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
Found 3 items
-rw-r--r-- 1 ubh01 supergroup      0 2022-04-18 15:58 /hadoop_test_14/_SUCCESS
-rw-r--r-- 1 ubh01 supergroup    29 2022-04-18 15:58 /hadoop_test_14/part-m-00000
-rw-r--r-- 1 ubh01 supergroup    44 2022-04-18 15:58 /hadoop_test_14/part-m-00001
```

```
ubh01@ubh01:~$ hadoop fs -cat /hadoop_test_14/part-m-00000
22/04/18 16:00:39 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
101,Akarsh,3000
102,Raj,4000
ubh01@ubh01:~$ hadoop fs -cat /hadoop_test_14/part-m-00001
22/04/18 16:00:46 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
103,Ram,5000
104,Ramesh,6000
105,Rohan,7000
```

```
Mon 16:00 •
ubho1@ubho1: ~
File Edit View Search Terminal Help
22/04/18 15:58:28 INFO mapreduce.Job: The url to track the job: http://ubho1:8088/proxy/application_1650276725951_0002
22/04/18 15:58:28 INFO mapreduce.Job: Job: job_1650276725951_0002 running in uber mode : false
22/04/18 15:58:28 INFO mapreduce.Job: map 0% reduce 0%
22/04/18 15:58:49 INFO mapreduce.Job: map 50% reduce 0%
22/04/18 15:58:59 INFO mapreduce.Job: map 100% reduce 0%
22/04/18 15:58:59 INFO mapreduce.Job: job_1650276725951_0002 completed successfully
22/04/18 15:58:59 INFO mapreduce.Job: Counters: 30
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=269948
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=217
    HDFS: Number of bytes written=73
    HDFS: Number of read operations=8
    HDFS: Number of large read operations=4
    HDFS: Number of write operations=4
  Job Counters
    Launched map tasks=2
    Other local map tasks=2
    Total time spent by all maps in occupied slots (ms)=15003
    Total time spent by all reducers in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=15003
    Total vcore-seconds taken by all map tasks=15003
    Total megabyte-seconds taken by all map tasks=15363072
  Map-Reduce Framework
    Map input records=5
    Map output records=5
    Input split bytes=217
    Spilled Records=0
    Failed Shuffles=0
    Merged Map Outputs=0
    GC time elapsed (ms)=894
    CPU time spent (ms)=15003
    Physical memory (bytes) snapshot=353951744
    Virtual memory (bytes) snapshot=3848572928
    Total committed heap usage (bytes)=282066944
  File Input Format Counters
    Bytes Read=0
    Bytes Written=73
22/04/18 15:58:59 INFO mapreduce.ImportJobBase: Transferred 73 bytes in 26.1987 seconds (2.7864 bytes/sec)
22/04/18 15:58:59 INFO mapreduce.ImportJobBase: Retrieved 5 records.
ubho1@ubho1:~$ hadoop fs -ls /hadoop/test/14
22/04/18 15:59:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-Java classes where applicable
e
Found 3 items
-rw-r--r-- 1 ubho1 supergroup          0 2022-04-18 15:58 /hadoop/test/14/_SUCCESS
-rw-r--r-- 1 ubho1 supergroup 29 2022-04-18 15:58 /hadoop/test/14/part-r-00000
-rw-r--r-- 1 ubho1 supergroup 44 2022-04-18 15:58 /hadoop/test/14/part-r-00001
ubho1@ubho1:~$
```

Activate Windows
Go to Settings to activate Windows.

Cloud 39°C Light rain ENG 15:59 18-04-2022

Cloud 16:00 18-04-2022

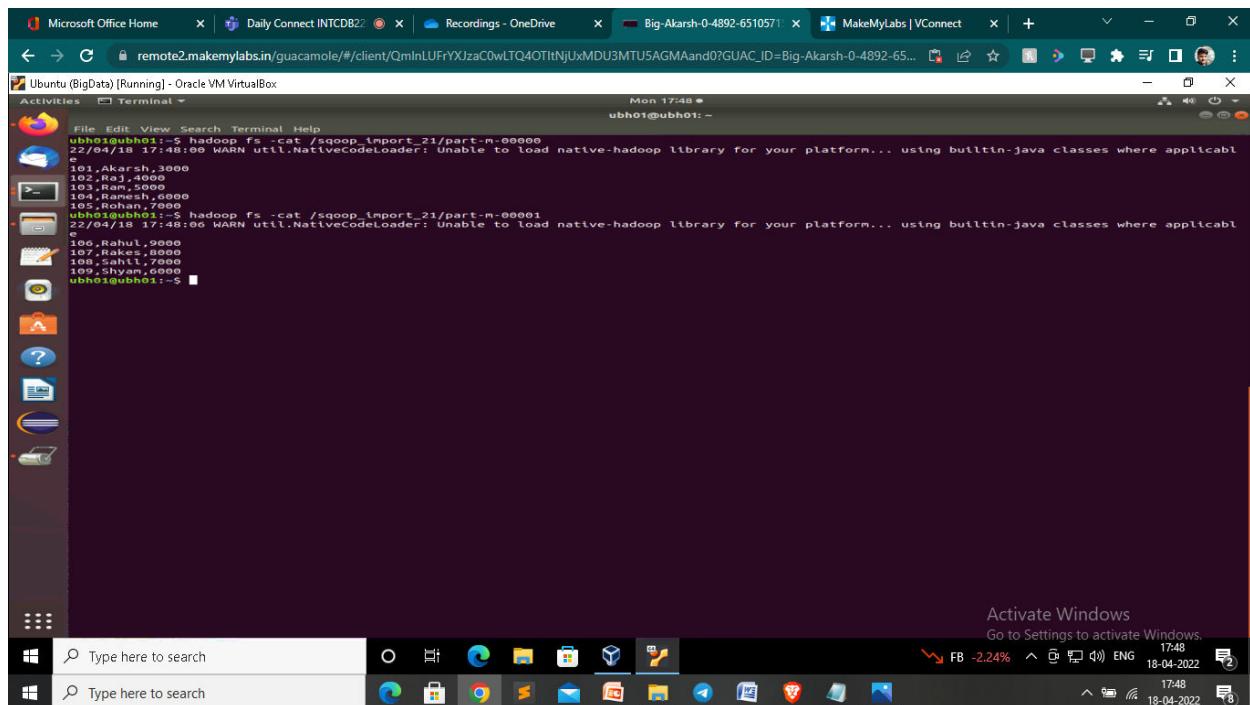
iii) Use incremental-append

```
ubh01@ubh01:~$ sqoop import \
> --connect jdbc:mysql://localhost/akarsh_db \
> --username root \
> --password password \
> --table emp \
> --incremental append \
> --check-column empid \
> -m 1 \
> --target-dir /sqoop_import_21
Warning: /home/ubh01/sqoop-1.4.7.bin__hadoop-2.6.0/..hcatalog does not exist! HCatalog jobs
will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/ubh01/sqoop-1.4.7.bin__hadoop-2.6.0/..accumulo does not exist! Accumulo
imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /home/ubh01/sqoop-1.4.7.bin__hadoop-2.6.0/..zookeeper does not exist! Accumulo
imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
22/04/18 16:45:43 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
22/04/18 16:45:43 WARN tool.BaseSqoopTool: Setting your password on the command-line is
insecure. Consider using -P instead.
22/04/18 16:45:43 INFO manager.MySQLManager: Preparing to use a MySQL streaming
resultset.
22/04/18 16:45:43 INFO tool.CodeGenTool: Beginning code generation
Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is
`com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and manual loading
of the driver class is generally unnecessary.
22/04/18 16:45:44 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM
`emp` AS t LIMIT 1
22/04/18 16:45:44 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM
`emp` AS t LIMIT 1
22/04/18 16:45:44 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is
/home/ubh01/hadoop-2.7.1
```

We have to add some more records

```
mysql> use akarsh_db;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A
```

```
Database changed
mysql> select * from emp;
+-----+-----+
| empid | ename | salary |
+-----+-----+
| 101 | Akarsh | 3000 |
| 102 | Raj | 4000 |
| 103 | Ram | 5000 |
| 104 | Ramesh | 6000 |
| 105 | Rohan | 7000 |
| 106 | Rahul | 9000 |
| 107 | Rakes | 8000 |
| 108 | Sahil | 7000 |
| 109 | Shyam | 6000 |
+-----+-----+
9 rows in set (0.00 sec)
```



iv) Transfer data to hive from sqoop.

```
ubh01@ubh01:~$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/ubh01/apache-hive-2.3.2-bin/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/ubh01/hadoop-2.7.1/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]
```

Logging initialized using configuration in jar:file:/home/ubh01/apache-hive-2.3.2-bin/lib/hive-common-2.3.2.jar!/hive-log4j2.properties Async: true

Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.

```
hive> show databases;
```

Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is `com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and manual loading of the driver class is generally unnecessary.

```
OK
```

```
db
```

```
default
```

```
pr
```

```
sumitdb
```

```
Time taken: 30.997 seconds, Fetched: 4 row(s)
```

```
hive> create database retail;
```

```
OK
```

```
Time taken: 0.641 seconds
```

```
hive> exit;
```

```
ubh01@ubh01:~$ sqoop import --connect jdbc:mysql://localhost/akarsh_db --username root --password password --hive-database retail --table emp --hive-import -m 1
```

Warning: /home/ubh01/sqoop-1.4.7.bin__hadoop-2.6.0/..hcatalog does not exist! HCatalog jobs will fail.

Please set \$HCAT_HOME to the root of your HCatalog installation.

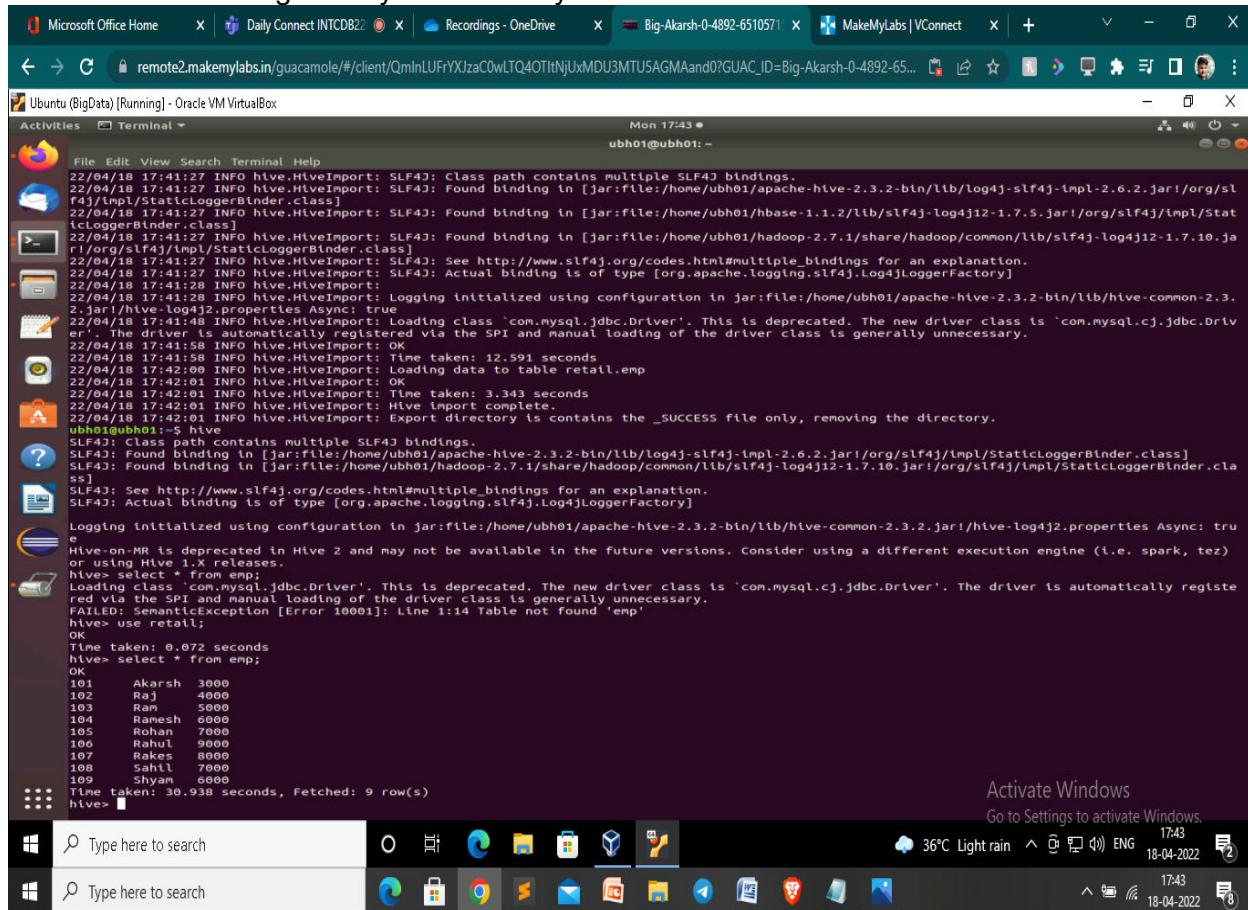
Warning: /home/ubh01/sqoop-1.4.7.bin__hadoop-2.6.0/..accumulo does not exist! Accumulo imports will fail.

Please set \$ACCUMULO_HOME to the root of your Accumulo installation.

Warning: /home/ubh01/sqoop-1.4.7.bin__hadoop-2.6.0/..zookeeper does not exist! Accumulo imports will fail.

Please set \$ZOOKEEPER_HOME to the root of your Zookeeper installation.

22/04/18 17:40:56 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
 22/04/18 17:40:56 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
 22/04/18 17:40:56 INFO tool.BaseSqoopTool: Using Hive-specific delimiters for output. You can override
 22/04/18 17:40:56 INFO tool.BaseSqoopTool: delimiters with --fields-terminated-by, etc.
 22/04/18 17:40:56 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
 22/04/18 17:40:56 INFO tool.CodeGenTool: Beginning code generation
 Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is `com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and manual loading of the driver class is generally unnecessary.



```

Microsoft Office Home      Daily Connect INTCDB2...      Recordings - OneDrive      Big-Akash-0-4892-6510571      MakeMyLabs | VConnect
← → C remote2.makemylabs.in/gucamole/#/client/QmlnLUFrYXJzaCwLTQ4OTltNjUxMDU3MTU5AGMAand0?GUAC_ID=Big-Akash-0-4892-65...
Ubuntu (BigData) [Running] - Oracle VM VirtualBox
Activities Terminal Mon 17:43 ubh01@ubh01: ~
ubh01@ubh01: ~
File Edit View Search Terminal Help
22/04/18 17:41:27 INFO hive.HiveImport: SLF4J: Class path contains multiple SLF4J bindings.
22/04/18 17:41:27 INFO hive.HiveImport: SLF4J: Found binding in [jar:file:/home/ubh01/apache-hive-2.3.2-bin/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
22/04/18 17:41:27 INFO hive.HiveImport: SLF4J: Found binding in [jar:file:/home/ubh01/hbase-1.1.2/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
22/04/18 17:41:27 INFO hive.HiveImport: SLF4J: Found binding in [jar:file:/home/ubh01/hadoop-2.7.1/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
22/04/18 17:41:27 INFO hive.HiveImport: SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
22/04/18 17:41:27 INFO hive.HiveImport: SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
22/04/18 17:41:28 INFO hive.HiveImport: Logging initialized using configuration in jar:file:/home/ubh01/apache-hive-2.3.2-bin/lib/hive-common-2.3.2.jar!/hive-log4j2.properties Async: true
22/04/18 17:41:48 INFO hive.HiveImport: Loading class 'com.mysql.jdbc.Driver'. This is deprecated. The new driver class is 'com.mysql.cj.jdbc.Driver'.
22/04/18 17:41:58 INFO hive.HiveImport: OK
22/04/18 17:41:58 INFO hive.HiveImport: Time taken: 12.591 seconds
22/04/18 17:42:00 INFO hive.HiveImport: Loading data to table retail.emp
22/04/18 17:42:00 INFO hive.HiveImport: OK
22/04/18 17:42:01 INFO hive.HiveImport: Time taken: 3.343 seconds
22/04/18 17:42:01 INFO hive.HiveImport: Hive import complete.
22/04/18 17:42:01 INFO hive.HiveImport: Export directory contains the _SUCCESS file only, removing the directory.
ubh01@ubh01: ~
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/ubh01/apache-hive-2.3.2-bin/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/ubh01/hadoop-2.7.1/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]
Logging initialized using configuration in jar:file:/home/ubh01/apache-hive-2.3.2-bin/lib/hive-common-2.3.2.jar!/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez)
Hive> select * from emp;
Loading class 'com.mysql.jdbc.Driver'. This is deprecated. The new driver class is 'com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and manual loading of the driver class is generally unnecessary.
FAILED: SemanticException [Error 10001]: Line 1:14 Table not found 'emp'
hive> use retail;
OK
Time taken: 0.072 seconds
hive> select * from emp;
OK
101 Akash 3000
102 Raj 4000
103 Ram 5000
104 Ramesh 6000
105 Rohan 7000
106 Rahul 9000
107 Rakesh 8000
108 Sahlil 7000
109 Shyam 6000
Time taken: 36.938 seconds, Fetched: 9 row(s)
hive>

```

ubh01@ubh01:~\$ hive
 SLF4J: Class path contains multiple SLF4J bindings.
 SLF4J: Found binding in [jar:file:/home/ubh01/apache-hive-2.3.2-bin/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]

```
SLF4J: Found binding in [jar:file:/home/ubh01/hadoop-2.7.1/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]
```

```
Logging initialized using configuration in jar:file:/home/ubh01/apache-hive-2.3.2-bin/lib/hive-common-2.3.2.jar!/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
```

```
hive> select * from emp;
```

```
Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is
`com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and manual loading
of the driver class is generally unnecessary.
```

```
FAILED: SemanticException [Error 10001]: Line 1:14 Table not found 'emp'
```

```
hive> use retail;
```

```
OK
```

```
Time taken: 0.072 seconds
```

```
hive> select * from emp;
```

```
OK
```

```
101 Akarsh 3000
```

```
102 Raj 4000
```

```
103 Ram 5000
```

```
104 Ramesh 6000
```

```
105 Rohan 7000
```

```
106 Rahul 9000
```

```
107 Rakes 8000
```

```
108 Sahil 7000
```

```
109 Shyam 6000
```

```
Time taken: 30.938 seconds, Fetched: 9 row(s)
```

V) Transfer data to hbase from sqoop.

```
ubh01@ubh01:~$ start-hbase.sh
starting master, logging to /home/ubh01/hbase-1.1.2/logs/hbase-ubh01-master-ubh01.out
OpenJDK 64-Bit Server VM warning: ignoring option PermSize=128m; support was removed in
8.0
OpenJDK 64-Bit Server VM warning: ignoring option MaxPermSize=128m; support was
removed in 8.0
ubh01@ubh01:~$ hbase shell
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/ubh01/hbase-1.1.2/lib/slf4j-log4j12-
1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/ubh01/hadoop-2.7.1/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.slf4j.impl.Log4jLoggerFactory]
2022-04-18 17:50:59,479 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 1.1.2, rcc2b70cf03e3378800661ec5cab11eb43fafe0fc, Wed Aug 26 20:11:27 PDT
2015
```

```
hbase(main):001:0> create 'ht3','cf3'
0 row(s) in 1.6320 seconds
```

```
=> Hbase::Table - ht3
hbase(main):002:0> list
TABLE
```

```
ht3
```

```
login
```

```
2 row(s) in 0.0240 seconds
```

```
=> ["ht3", "login"]
```

```
ubh01@ubh01:~$ sqoop import --connect jdbc:mysql://localhost/akarsh_db \
> --username root \
> --password password \
> --table emp \
> --hbase-table ht3 \
> --column-family cf3 \
> --hbase-row-key empid \
> --hbase-create-table -m 1
Warning: /home/ubh01/sqoop-1.4.7.bin__hadoop-2.6.0/../hcatalog does not exist! HCatalog jobs
will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/ubh01/sqoop-1.4.7.bin__hadoop-2.6.0/../accumulo does not exist! Accumulo
imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /home/ubh01/sqoop-1.4.7.bin__hadoop-2.6.0/../zookeeper does not exist! Accumulo
imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
22/04/18 17:56:29 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
22/04/18 17:56:29 WARN tool.BaseSqoopTool: Setting your password on the command-line is
insecure. Consider using -P instead.
22/04/18 17:56:30 INFO manager.MySQLManager: Preparing to use a MySQL streaming
resultset.
22/04/18 17:56:30 INFO tool.CodeGenTool: Beginning code generation
Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is
`com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and manual loading
of the driver class is generally unnecessary.
22/04/18 17:56:30 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM
`emp` AS t LIMIT 1
22/04/18 17:56:30 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM
`emp` AS t LIMIT 1
22/04/18 17:56:30 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is
/home/ubh01/hadoop-2.7.1
Note: /tmp/sqoop-ubh01/compile/c18be4c0c713dd7bae16652a036e304a/emp.java uses or
overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/04/18 17:56:32 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-
ubh01/compile/c18be4c0c713dd7bae16652a036e304a/emp.jar
22/04/18 17:56:32 WARN manager.MySQLManager: It looks like you are importing from mysql.
22/04/18 17:56:32 WARN manager.MySQLManager: This transfer can be faster! Use the --
direct
22/04/18 17:56:32 WARN manager.MySQLManager: option to exercise a MySQL-specific fast
path.
22/04/18 17:56:32 INFO manager.MySQLManager: Setting zero DATETIME behavior to
convertToNull (mysql)
```

The screenshot shows a Microsoft Windows desktop environment with several open windows. At the top, there are browser tabs for Microsoft Office Home, Daily Connect INTCD82, Recordings - OneDrive, Big-Akarsh-0-4892-651057, and MakeMyLabs | VConnect. Below the tabs, a taskbar displays icons for File Explorer, Control Panel, and other system utilities. The main focus is a terminal window titled "Ubuntu (BigData) [Running] - Oracle VM VirtualBox". The terminal session is as follows:

```
ubh01@ubh01:~$ start-hbase.sh
starting master, logging to /home/ubh01/hbase-1.1.2/logs/hbase-ubh01-master-ubh01.out
OpenJDK 64-Bit Server VM warning: ignoring option PermSize=128m; support was removed in 8.0
OpenJDK 64-Bit Server VM warning: ignoring option MaxPermSize=128m; support was removed in 8.0
ubh01@ubh01:~$ hbase shell
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [/home/ubh01/Hadoop-2.7.1/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [/jar:/file:/home/ubh01/Hadoop-2.7.1/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.slf4j.impl.Log4jLoggerFactory]
2015-08-20 20:11:59,479 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes w
2015-08-20 20:11:59,479 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes w
HBase Shell; enter 'help' for list of supported commands.
Type 'exit+RETURN' to leave the HBase Shell
Version 1.1.2, build c2b7ecfcf03e33780000dec5cabileb43fafc, Wed Aug 20 20:11:27 PDT 2015

hbase(main):001:0> create 'ht3','cf3'
0 row(s) in 1.6326 seconds
==> hbase:001:0> list
hbase(main):002:0> list
TABLE
ht3
2 row(s) in 0.0240 seconds
==> ["ht3", "logIn"]
hbase(main):003:0>
```

Ubuntu (BigData) [Running] - Oracle VM VirtualBox

```
Mon 17:59 ● ubh01@ubh01: ~
File Edit View Search Terminal Help
ubh01@ubh01: ~$ start-hbase.sh
starting master, logging to /home/ubh01/hbase-1.1.2/logs/hbase-ubh01-master.ubh01.out
OpenJDK 64-Bit Server VM warning: ignoring option MaxPermSize=128M; support was removed in 8.0
OpenJDK 64-Bit Server VM warning: ignoring option MaxPermSize=128M; support was removed in 8.0
ubh01@ubh01: ~$ hbase shell
hbase:001:0> create 'ht3', 'cf3'
SLF4J: Found binding in [jar:file:/home/ubh01/hbase-1.1.2/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/ubh01/hadoop-2.1.0/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
ZooKeeper@ubh01:~$:59479 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes w
here applicable
HBase Shell; enter 'help' or 'RETURN?' for list of supported commands.
Type 'exit()' or 'quit()' to leave the HBase Shell.
Version 1.1.2, rcc2b70cf03e37980000001ec5cab1eb43fafe0fc, Wed Aug 26 20:11:27 PDT 2015
hbase(main):001:0> create 'ht3','cf3'
9 row(s) in 0.0326 seconds
hbase(main):002:0> scan 'ht3'
Table 'ht3' has 2 row(s)
2 row(s) in 0.0240 seconds
hbase(main):003:0> list
TABLE ht3
Login
2 row(s) in 0.0240 seconds
hbase(main):004:0> scan 'ht3'
[{"row": "R01", "cf": "cf3", "qualifier": "ename", "value": "Akash", "timestamp": 1650284821977}, {"row": "R01", "cf": "cf3", "qualifier": "salary", "value": 3000, "timestamp": 1650284821977}, {"row": "R02", "cf": "cf3", "qualifier": "ename", "value": "Raj", "timestamp": 1650284821977}, {"row": "R02", "cf": "cf3", "qualifier": "salary", "value": 4000, "timestamp": 1650284821977}, {"row": "R03", "cf": "cf3", "qualifier": "ename", "value": "Ran", "timestamp": 1650284821977}, {"row": "R03", "cf": "cf3", "qualifier": "salary", "value": 5000, "timestamp": 1650284821977}, {"row": "R04", "cf": "cf3", "qualifier": "ename", "value": "Roshni", "timestamp": 1650284821977}, {"row": "R04", "cf": "cf3", "qualifier": "salary", "value": 6000, "timestamp": 1650284821977}, {"row": "R05", "cf": "cf3", "qualifier": "ename", "value": "Rohan", "timestamp": 1650284821977}, {"row": "R05", "cf": "cf3", "qualifier": "salary", "value": 7000, "timestamp": 1650284821977}, {"row": "R06", "cf": "cf3", "qualifier": "ename", "value": "Rahul", "timestamp": 1650284821977}, {"row": "R06", "cf": "cf3", "qualifier": "salary", "value": 9000, "timestamp": 1650284821977}, {"row": "R07", "cf": "cf3", "qualifier": "ename", "value": "Ritesh", "timestamp": 1650284821977}, {"row": "R07", "cf": "cf3", "qualifier": "salary", "value": 8000, "timestamp": 1650284821977}, {"row": "R08", "cf": "cf3", "qualifier": "ename", "value": "Sahil", "timestamp": 1650284821977}, {"row": "R08", "cf": "cf3", "qualifier": "salary", "value": 10000, "timestamp": 1650284821977}, {"row": "R09", "cf": "cf3", "qualifier": "ename", "value": "Shyan", "timestamp": 1650284821977}, {"row": "R09", "cf": "cf3", "qualifier": "salary", "value": 6000, "timestamp": 1650284821977}
9 row(s) in 0.2020 seconds
hbase(main):004:0>
Activate Windows
Go to Settings to activate Windows.
```

ASSIGNMENT NO.6

TOPIC -> Scala Program

SUB TOPIC -> USING SCALA

1) Define one empty ArrayBuffer

```
scala> var a6=ArrayBuffer[Int]()
a6: scala.collection.mutable.ArrayBuffer[Int] = ArrayBuffer()
```

2) Add 10 into ArrayBuffer

```
scala> a6+=10
res27: scala.collection.mutable.ArrayBuffer[Int] = ArrayBuffer(10)
```

3) Add 20,30,40 into ArrayBuffer

```
scala> a6+=(20,30,40)
res28: scala.collection.mutable.ArrayBuffer[Int] = ArrayBuffer(10, 20, 30, 40)
```

4) Define arr={1,2,3,4,5}

```
scala> var array=Array(1,2,3,4,5)
array: Array[Int] = Array(1, 2, 3, 4, 5)
```

5) Append array into ArrayBuffer

```
scala> a6.appendAll(array)
scala> a6
res32: scala.collection.mutable.ArrayBuffer[Int] = ArrayBuffer(10, 20, 30, 40, 1, 2, 3, 4, 5)
```

6) Remove 30

```
scala> a6.remove(2)
res35: Int = 30

scala> a6
res36: scala.collection.mutable.ArrayBuffer[Int] = ArrayBuffer(10, 20, 40, 1, 2, 3, 4, 5)
```

7)Add 35 in between 20 and 40.

```
scala> a6.insert(2,35)
scala> a6
res38: scala.collection.mutable.ArrayBuffer[Int] = ArrayBuffer(10, 20, 35, 40, 1, 2, 3, 4, 5)
```

8)Check 35 is there in ArrayBuffer

```
scala> a6.contains(35)
res40: Boolean = true
```

9)Check ArrayBuffer is empty or not.

```
scala> a6.isEmpty
res39: Boolean = false
```

10)Remove last 3 elements

```
scala> a6.trimEnd(3)
scala> a6
res42: scala.collection.mutable.ArrayBuffer[Int] = ArrayBuffer(10, 20, 35, 40, 1, 2)
```

11)Add 10 in each element

```
scala> a6.map(a6=>(a6+10))
res43: scala.collection.mutable.ArrayBuffer[Int] = ArrayBuffer(20, 30, 45, 50, 11, 12)
```

12)Get element of ArrayBuffer where element >10.

```
scala> a6.filter(a6=>(a6>10))
res44: scala.collection.mutable.ArrayBuffer[Int] = ArrayBuffer(20, 35, 40)
```

ASSIGNMENT NO.7

TOPIC -> SCALA PROGRAM AND WORD COUNT

PROGRAM

SUB TOPIC ->USING SCALA

1)CREATED ONE EMPTY MAP AND ADDED ONE KEY-VALUE IN IT.

```
scala> import scala.collection.mutable.Map
import scala.collection.mutable.Map

scala> val mp=Map[String,Float]()
mp: scala.collection.mutable.Map[String,Float] = Map()

scala> mp+="c1"->121.1f
res11: mp.type = Map(c1 -> 121.1)

scala> mp
res12: scala.collection.mutable.Map[String,Float] = Map(c1 -> 121.1)
```

i)To get key of the map

```
scala> mp.keys
res13: Iterable[String] = Set(c1)
```

ii)To get values of the map

```
scala> mp.values
res14: Iterable[Float] = HashMap(121.1)
```

iii)To check whether key is present or not.

```
scala> mp.contains("c1")
res15: Boolean = true
```

iv)To get the value of given key.

```
scala> mp("c1")
res16: Float = 121.1
```

v)If value is not there return 0.0

```
scala> mp.getOrElse("c3",0.0)
res17: AnyVal = 0.0
```

vi)If the value is not there,then add 0.0 to c3.

```
scala> mp.update("c3",0.0f)
scala> mp
res19: scala.collection.mutable.Map[String,Float] = Map(c1 -> 121.1, c3 -> 0.0)
```

2) Word Count Program in Scala

```
import scala.collection.mutable.Map
val str = "car car river car river"
var mp = Map[String,Integer]()
for(word<-str.split(" ")){
    if(!mp.contains(word))
        mp+=(word->1)
    else
    {
        mp.get(word) match {
            case Some(mp_val) => mp+=(word->(mp_val+1))
            case None => print("Not valid")
        }
    }
}
print(mp)
```

```
HashMap(car -> 3, river -> 2)
```

Assignment No. 8

Topic- Read data from custs.txt and then sort the data by age(in ascending order and descending order)

Sub Topic->Using Scala

Data -> 'Custs.txt'

Code->

The screenshot shows the IntelliJ IDEA interface with the following details:

- Project Structure:** The project is named "ScalaP". It contains a "src" directory with "main" and "scala" sub-directories. "main" contains "com" and "examples" packages. "examples" contains "student.scala", "emp.scala", "practise.scala", "math.scala", "cust.scala", and "top.scala". "scala" contains "com.examples" which has "student", "emp", "practise", "math", "cust", and "top" sub-packages.
- Code Editor:** The file "top.scala" is open. The code defines a case class "top" and an object "testtop". The "testtop" object contains a main method that reads lines from "custs.txt", splits them into fields, and sorts them by age.
- Run Tab:** The "Run" tab shows a successful build with 1 warning in 15 sec, 315 ms (4 minutes ago).
- System Tray:** The system tray shows the date and time as 22-04-2022 18:25.

```
1 package com.examples
2 import scala.io.Source
3 import scala.collection.mutable.ArrayBuffer
4 case class top(cid:String,fname:String,lname:String,age:Int,desig:String) {
5   override def toString: String = s"$cid $fname $lname $age $desig"
6 }
7 object testtop{
8   def main(args: Array[String]): Unit = {
9     try {
10       val k = Source.fromFile("C:\\Users\\HP\\Desktop\\custs.txt").getLines()
11       val arrBuff = ArrayBuffer[top]()
12       for(records<-k){
13         val rec = records.split( regex= "\\s" )
14         val cid:String = rec(0)
15         val fname:String = rec(1)
16         val lname:String = rec(2)
17         val age:Int = rec(3).toInt
18         val desig:String = if (rec.length!=5) "NA" else rec(4)
19         arrBuff+=top(cid,fname,lname,age,desig)
20       }
21       val new_ab=arrBuff.sortBy(ob=>ob.age)
22       for(i<-new_ab)
23         println(i)
24     }}}
```

Output->(In ascending order)

The screenshot shows the IntelliJ IDEA interface with a Scala project named 'ScalaP'. The 'Run' tool window is open, displaying the output of a run named 'testtop'. The output consists of a list of employee records, each containing their ID (cid), first name (fname), last name (lname), age, and job title (desig). The records are sorted by ID in ascending order. A message at the bottom of the Run window indicates a successful build with one warning.

```
cid = 4001541 fname = Nina lname = Humphrey age = 21 desig = Human resources assistant
cid = 4001563 fname = Derek lname = Foster age = 21 desig = Civil engineer
cid = 4001589 fname = Penny lname = Lu age = 21 desig = Judge
cid = 4001732 fname = Vanessa lname = Franklin age = 21 desig = Teacher
cid = 4001843 fname = Carole lname = Rogers age = 21 desig = Teacher
cid = 4001869 fname = Kristine lname = Brown age = 21 desig = Financial analyst
cid = 4001893 fname = Gilbert lname = Ball age = 21 desig = Farmer
cid = 4001965 fname = Charlene lname = Snow age = 21 desig = Statistician
cid = 4002440 fname = Stephanie lname = Levine age = 21 desig = Nurse
cid = 4002543 fname = Edwin lname = Lim age = 21 desig = Artist
cid = 4002668 fname = Erin lname = Marsh age = 21 desig = Financial analyst
cid = 4002693 fname = Barry lname = Bowles age = 21 desig = Firefighter
cid = 4002715 fname = Russell lname = Huang age = 21 desig = Writer
cid = 4002769 fname = Valerie lname = Smith age = 21 desig = Loan officer
cid = 4002901 fname = Paula lname = Jiang age = 21 desig = Pilot
cid = 4003207 fname = Stacey lname = Stroud age = 21 desig = Coach
cid = 4003265 fname = Greg lname = Michael age = 21 desig = Engineering technician
cid = 4003412 fname = Bruce lname = Werner age = 21 desig = Accountant
cid = 4003486 fname = Yvonne lname = Andrews age = 21 desig = Statistician
cid = 4003519 fname = Stanley lname = Frost age = 21 desig = Automotive mechanic
cid = 4003523 fname = Derek lname = Dudley age = 21 desig = Psychologist
cid = 4003580 fname = Henry lname = Bond age = 21 desig = Photographer
cid = 4003596 fname = Carol lname = Weaver age = 21 desig = Coach
cid = 4004032 fname = Tommy lname = Parrott age = 21 desig = Lawyer
```

```
cid = 4000052 fname = Shirley lname = Merritt age = 21 desig = Reporter
cid = 4000176 fname = Kristin lname = Alexander age = 21 desig = Coach
cid = 4000290 fname = Gail lname = Whitehead age = 21 desig = Nurse
cid = 4000292 fname = Donna lname = Rice age = 21 desig = Social worker
cid = 4000349 fname = Russell lname = Dalton age = 21 desig = Civil engineer
cid = 4000451 fname = Russell lname = Hess age = 21 desig = Architect
cid = 4000465 fname = Peter lname = Guthrie age = 21 desig = Nurse
cid = 4000765 fname = Janice lname = Bowden age = 21 desig = Politician
cid = 4000779 fname = Jason lname = Bowling age = 21 desig = Coach
cid = 4001043 fname = Ian lname = Hodges age = 21 desig = Coach
cid = 4001101 fname = Toni lname = Byers age = 21 desig = Economist
cid = 4001109 fname = Edwin lname = Ball age = 21 desig = Carpenter
```

```

cid = 4001213 fname = Brandon lname = Nance age = 21 desig = Reporter

cid = 4001430 fname = Lindsay lname = Reynolds age = 21 desig = Computer support specialist

cid = 4001453 fname = Paige lname = Wolfe age = 21 desig = Financial analyst

cid = 4001541 fname = Nina lname = Humphrey age = 21 desig = Human resources assistant

cid = 4001563 fname = Derek lname = Foster age = 21 desig = Civil engineer

cid = 4001589 fname = Penny lname = Lu age = 21 desig = Judge

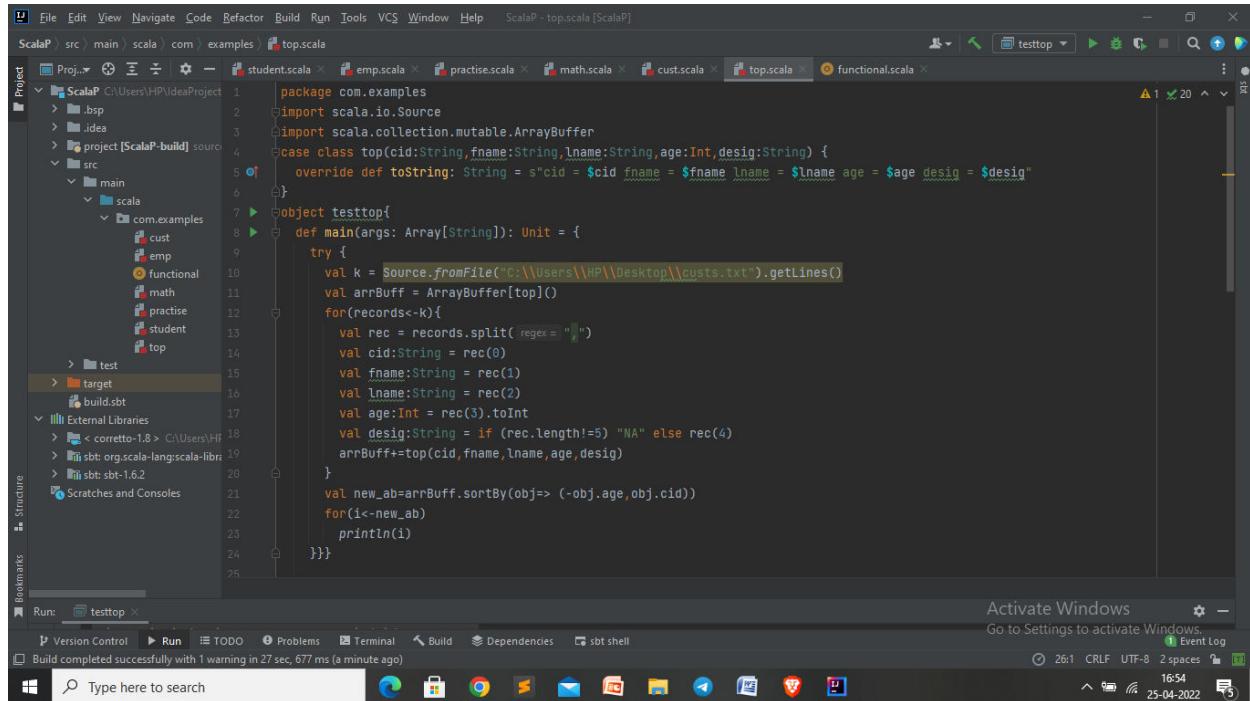
cid = 4001732 fname = Vanessa lname = Franklin age = 21 desig = Teacher

cid = 4001843 fname = Carole lname = Rogers age = 21 desig = Teacher

.....10000 records

```

In descending order



```

File Edit View Navigate Code Refactor Build Run Tools VCS Window Help ScalaP - top.scala [ScalaP]
ScalaP > main > scala > com > examples > top.scala
Project学生.scala < emp.scala < practise.scala < math.scala < cust.scala < top.scala < functional.scala
src > .bsp > .idea > [Scala-build] source > main > scala > com.examples > student > top > test > target > build.sbt > External Libraries < corretto-1.8 > C:\Users\HF> sbt org.scala-lang:scalalib > sbt sbt-1.6.2 > Scratches and Consoles
Structure Bookmarks
Activate Windows Go to Settings to activate Windows.
Run: testtop > Version Control Run TODO Problems Terminal Build Dependencies sbt shell
Build completed successfully with 1 warning in 27 sec, 677 ms (a minute ago)
26.1 CRLF UTF-8 2 spaces Event Log
25-04-2022 16:54
Type here to search

```

```

package com.examples
import scala.io.Source
import scala.collection.mutable.ArrayBuffer
case class top(cid:String, fname:String, lname:String, age:Int, desig:String) {
  override def toString: String = s"cid = $cid fname = $fname lname = $lname age = $age desig = $desig"
}
object testtop {
  def main(args: Array[String]): Unit = {
    try {
      val k = Source.fromFile("C:\\Users\\HF\\Desktop\\custs.txt").getLines()
      val arrBuff = ArrayBuffer[top]()
      for(records<-k){
        val rec = records.split( regex = "\\s" )
        val cid:String = rec(0)
        val fname:String = rec(1)
        val lname:String = rec(2)
        val age:Int = rec(3).toInt
        val desig:String = if (rec.length!=5) "NA" else rec(4)
        arrBuff+=top(cid,fname,lname,age,desig)
      }
      val new_ab=arrBuff.sortBy(obj=> (-obj.age,obj.cid))
      for(i<-new_ab)
        println(i)
    }}}

```

```
D:\Users\HP\.jdks\corretto-1.8.0_322\bin\java.exe ...
cid = 4000089 fname = Sherri lname = Sutton age = 75 desig = Social worker
cid = 4000325 fname = Ruth lname = Harmon age = 75 desig = Artist
cid = 4000494 fname = Stephen lname = Waller age = 75 desig = NA
cid = 4000740 fname = Franklin lname = Corbett age = 75 desig = Firefighter
cid = 4000780 fname = Neal lname = Robinson age = 75 desig = Artist
cid = 4000984 fname = Ann lname = Branch age = 75 desig = Computer hardware engineer
cid = 4001068 fname = Pamela lname = Fink age = 75 desig = Statistician
cid = 4001089 fname = Gretchen lname = Frost age = 75 desig = Accountant
cid = 4001097 fname = Jesse lname = Faulkner age = 75 desig = Recreation and fitness worker
cid = 4001216 fname = Mike lname = Robertson age = 75 desig = Pharmacist
cid = 4001322 fname = Russell lname = Frye age = 75 desig = Coach
cid = 4001362 fname = Martin lname = Kearney age = 75 desig = Architect
cid = 4001364 fname = Nancy lname = Hunter age = 75 desig = Judge
cid = 4001426 fname = Matthew lname = Kay age = 75 desig = Judge
cid = 4001436 fname = Melanie lname = Middleton age = 75 desig = Librarian
cid = 4001500 fname = Bill lname = Boyette age = 75 desig = Accountant
cid = 4001608 fname = Marguerite lname = Montgomery age = 75 desig = Physicist
cid = 4001777 fname = Karl lname = Siegel age = 75 desig = Lawyer
cid = 4001837 fname = Teresa lname = Honeycutt age = 75 desig = Computer hardware engineer
cid = 4001937 fname = Audrey lname = Kelly age = 75 desig = Teacher
cid = 4001945 fname = Curtis lname = Gentry age = 75 desig = Pharmacist
cid = 4001981 fname = Marilyn lname = Hensley age = 75 desig = Recreation and fitness worker
cid = 4002046 fname = Gene lname = Bynum age = 75 desig = Chemist
cid = 4002207 fname = Tony lname = Silverman age = 75 desig = Veterinarian
cid = 4002346 fname = Phyllis lname = Abbott age = 75 desig = Pharmacist
```

```
cid = 4000089 fname = Sherri lname = Sutton age = 75 desig = Social worker
cid = 4000325 fname = Ruth lname = Harmon age = 75 desig = Artist
cid = 4000494 fname = Stephen lname = Waller age = 75 desig = NA
cid = 4000740 fname = Franklin lname = Corbett age = 75 desig = Firefighter
cid = 4000780 fname = Neal lname = Robinson age = 75 desig = Artist
cid = 4000984 fname = Ann lname = Branch age = 75 desig = Computer hardware engineer
cid = 4001068 fname = Pamela lname = Fink age = 75 desig = Statistician
cid = 4001089 fname = Gretchen lname = Frost age = 75 desig = Accountant
cid = 4001097 fname = Jesse lname = Faulkner age = 75 desig = Recreation and fitness worker
cid = 4001216 fname = Mike lname = Robertson age = 75 desig = Pharmacist
cid = 4001322 fname = Russell lname = Frye age = 75 desig = Coach
cid = 4001362 fname = Martin lname = Kearney age = 75 desig = Architect
cid = 4001364 fname = Nancy lname = Hunter age = 75 desig = Judge
cid = 4001426 fname = Matthew lname = Kay age = 75 desig = Judge
cid = 4001436 fname = Melanie lname = Middleton age = 75 desig = Librarian
cid = 4001500 fname = Bill lname = Boyette age = 75 desig = Accountant
cid = 4001608 fname = Marguerite lname = Montgomery age = 75 desig = Physicist
cid = 4001777 fname = Karl lname = Siegel age = 75 desig = Lawyer
cid = 4001837 fname = Teresa lname = Honeycutt age = 75 desig = Computer hardware engineer
cid = 4001937 fname = Audrey lname = Kelly age = 75 desig = Teacher
cid = 4001945 fname = Curtis lname = Gentry age = 75 desig = Pharmacist
cid = 4001981 fname = Marilyn lname = Hensley age = 75 desig = Recreation and fitness worker
cid = 4002046 fname = Gene lname = Bynum age = 75 desig = Chemist
cid = 4002207 fname = Tony lname = Silverman age = 75 desig = Veterinarian
cid = 4002346 fname = Phyllis lname = Abbott age = 75 desig = Pharmacist
```

```
cid = 4001608 fname = Marguerite lname = Montgomery age = 75 desig = Physicist  
cid = 4001777 fname = Karl lname = Siegel age = 75 desig = Lawyer  
.....10000 records
```

Assignment No. 9

Topic:->Spark

Sub topic:-> using spark doing joining,getting top result

i) read data from txns1.txt and show the first data

```
scala> val trans=sc.textFile("C:\\\\Users\\\\HP\\\\Desktop\\\\txns1.txt").map(_.split(","))
trans: org.apache.spark.rdd.RDD[Array[String]] = MapPartitionsRDD[29] at map at <console>:24

scala> trans.first
res10: Array[String] = Array(0000000, 06-26-2011, 4007024, 040.33, Exercise & Fitness, Cardio Machine, Long Beach, California, credit)
```

ii)Take only id and amount and change the amount
into float

```
scala> val trans=sc.textFile("C:\\Users\\HP\\Desktop\\txns1.txt").map(_.split(",")).map(arr=>(arr(2),arr(3).toFloat))
trans: org.apache.spark.rdd.RDD[(String, Float)] = MapPartitionsRDD[33] at map at <console>:24

scala> trans.first
res11: (String, Float) = (4007024,40.33)
```

iii)Customer wise sum of sale from California and
amt > 100 and apply reduceByKey at last

```
scala> val cust=sc.textFile("C:\\Users\\HP\\Desktop\\txns1.txt").map(_.split(",")).filter(arr=>(arr(7)=="California")&&(arr(3).toFloat>100)).map(arr=>(arr(2),arr(3).toFloat))
cust: org.apache.spark.rdd.RDD[(String, Float)] = ShuffledRDD[18] at reduceByKey at <console>:24

scala> cust.first
res3: (String, Float) = (4001485,195.35)
```

Activate Windows
Go to Settings to activate Windows.

iv)Top 10 customer in California

```
scala> val sort_out=cust.sortBy(tu=> (-tu._2))
```

```
scala> val top10desig=sc.parallelize(sort_out.take(10))
top10desig: org.apache.spark.rdd.RDD[(String, Float)] = ParallelCollectionRDD[25] at parallelize at <console>:26

scala> top10desig.collect.foreach(println)
(4001779,669.99)
(4000749,628.33)
(4009406,579.24)
(4006874,559.27)
(4003228,549.6)
(4006616,532.74)
(4001691,531.32)
(4002619,531.03)
(4005483,529.43)
(4007989,521.62)
```

v)Get id,fn,ln,amt

```
scala> val txns=sc.parallelize(sc.textFile("C:\\Users\\\\HP\\\\Desktop\\\\txns1.txt").map(_.split(",")).filter(arr=>(arr(7)==="California")&&(arr(3).toFloat>100)).map(arr=>(arr(2),arr(3).toFloat)).reduceByKey(_+_).sortBy(-arr._2).take(10),1)
txns: org.apache.spark.rdd.RDD[(String, Float)] = ParallelCollectionRDD[297] at parallelize at <console>:24

scala> txns.first
res6: (String, Float) = (4001779,669.99)

scala> val custNamePair=sc.textFile("C:\\Users\\\\HP\\\\Desktop\\\\custs.txt").map(_.split(",")).map(arr=> (arr(0),(arr(1),arr(2))))
custNamePair: org.apache.spark.rdd.RDD[(String, (String, String))] = MapPartitionsRDD[301] at map at <console>:24

scala> custNamePair.first
res7: (String, (String, String)) = (4000001,(Kristina,Chung))

scala> val jo=txns.join(custNamePair).map(obj=>(obj._1,obj._2._1,obj._2._2,obj._2._2,obj._2._1)).sortBy(obj=>-obj._4)
jo: org.apache.spark.rdd.RDD[(String, String, String, Float)] = MapPartitionsRDD[310] at sortBy at <console>:27

scala> jo.collect.foreach(println)
(4001779,William,Cross,669.99)
(4000749,Wesley,Buckley,628.33)
(4009406,Henry,Joyner,579.24)
(4006874,Teresa,Peters,559.27)
(4003228,Elsie,Newton,549.6)
(4006616,Regina,Jiang,532.74)
(4001691,Laura,Berger,531.32)
(4002619,James,Kang,531.03)
(4005483,Clyde,Michael,529.43)
(4007989,Earl,Osborne,521.62)
```

Activate Windows
Go to Settings to activate Windows.