Case Study 2

Case Study Description

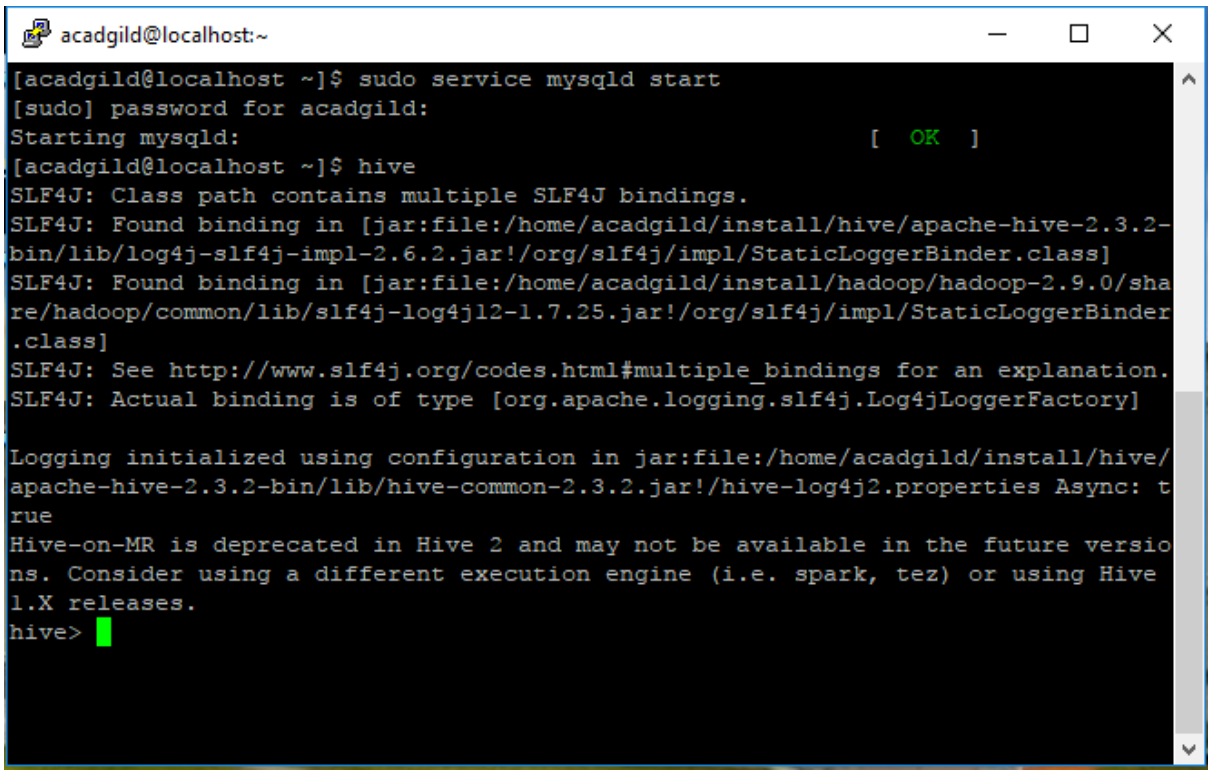
Let us take up the CUSTOMER and TRANSACTIONS table we have created in the Let's Do Together section. Let us solve the following use cases using these tables :-

Step 1 :

Start Hive

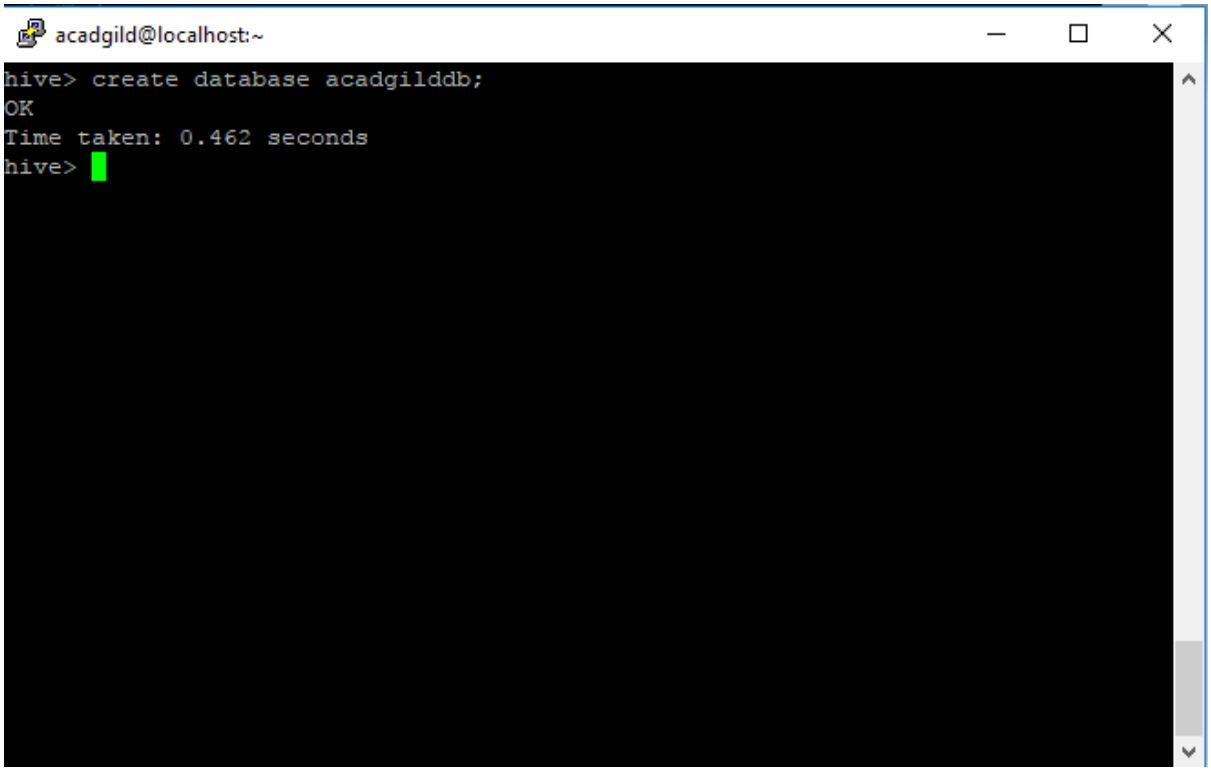
sudo service mysqld start

hive



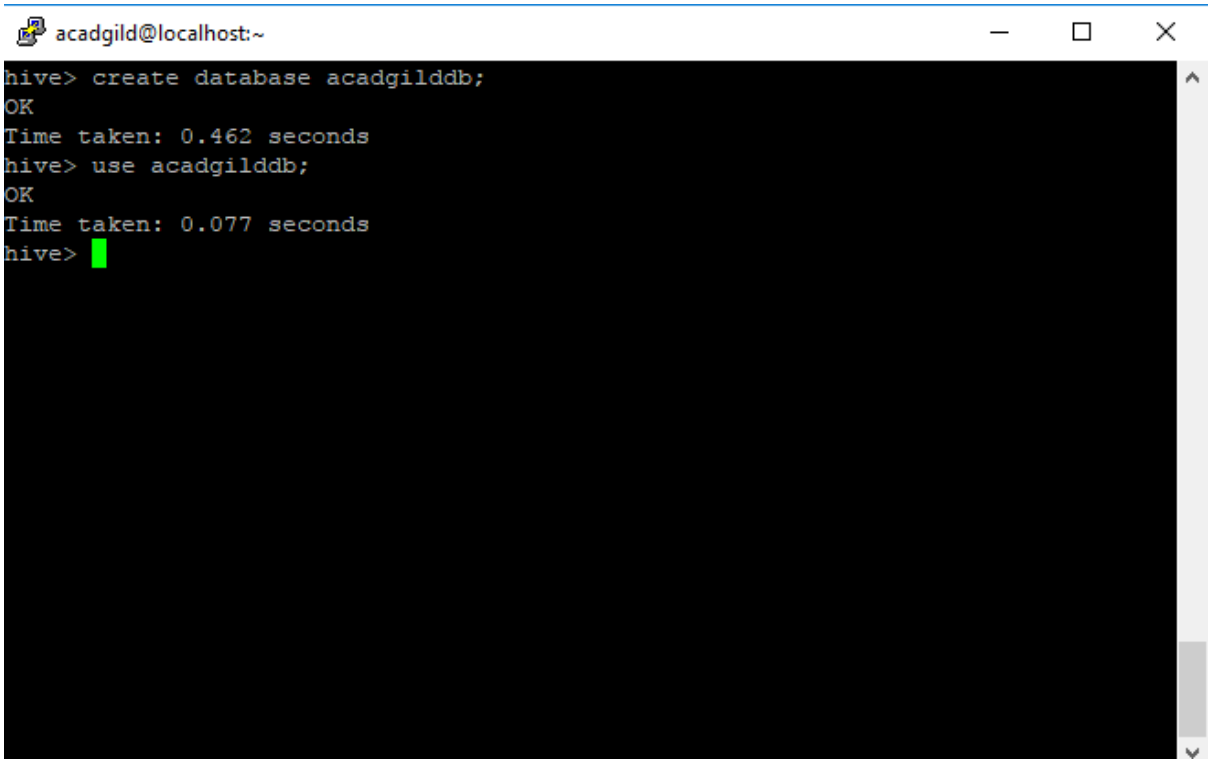
Step2: Create database

create database acadgilddb;



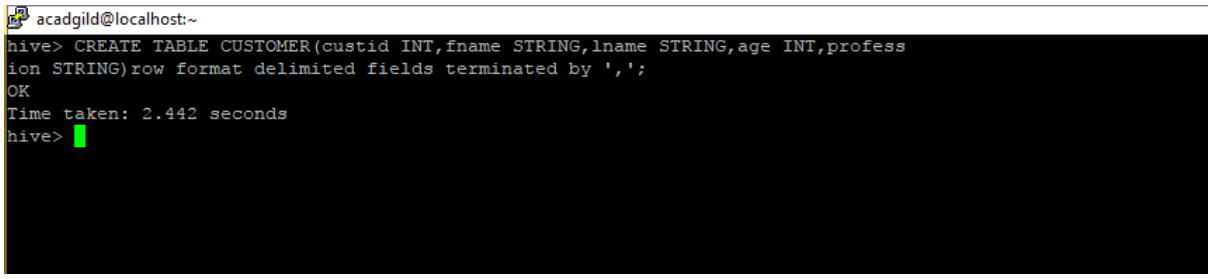
Step 3:

use acadgilddb;

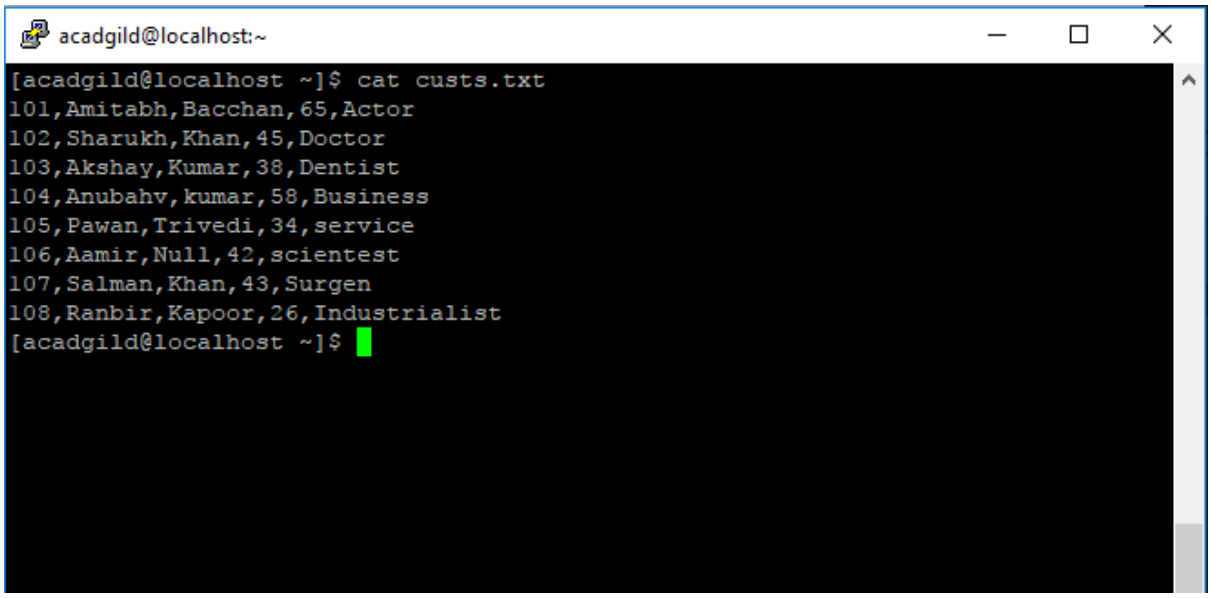


Step 4: Create Customer Table

CREATE TABLE CUSTOMER(custid INT,fname STRING,lname STRING,age INT,profession STRING)row format delimited fields terminated by ',';



Step 5: Create custs.txt file

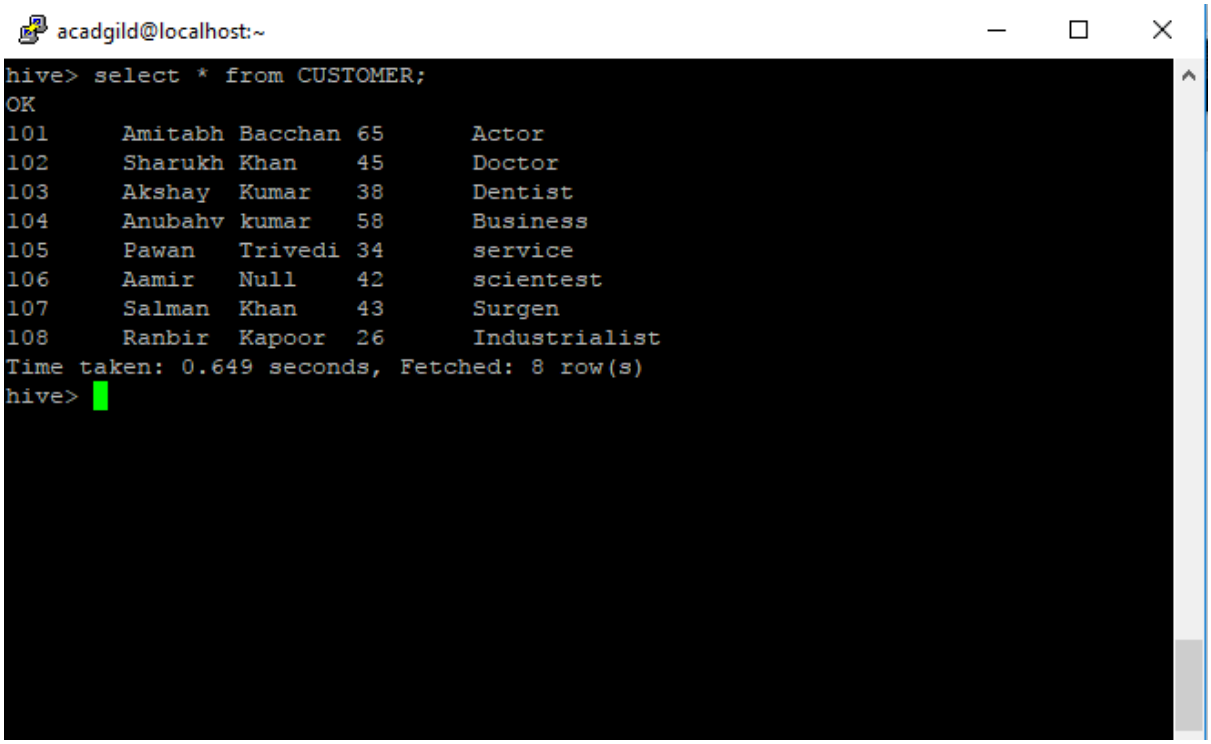


Step 6: Load custs.txt into Customer table

LOAD DATA LOCAL INPATH '/home/acadgild/customer.txt' into table CUSTOMER;

Step 7 :

Select * from CUSTOMER;



Step 8 : Create Transaction Table

CREATE TABLE TRANSACTIONS(txnno INT,txndate STRING,custno INT,amount DOUBLE,category STRING,product STRING,city STRING,state STRING,spendby STRING)row format delimited fields terminated by ',';

```
acadgild@localhost:~
hive> CREATE TABLE TRANSACTIONS(txnno INT,txndate STRING,custno INT,amount DOUBL
E,category STRING,product STRING,city STRING,state STRING,spendby STRING)row for
mat delimited fields terminated by ',';
OK
Time taken: 0.389 seconds
hive>
```

Step 9 : Create txn.txt file

```
acadgild@localhost:~
[acadgild@localhost ~]$ cat txn.txt
97834,05/02/2018,101,965,Entertainment,Movie,Pune,Maharashtra,Daughter
98396,12/01/2018,102,239,Food,Grocery,Patna,Bihar,Self
34908,06/01/2018,101,875,Travel,Air,Bangalore,Karnataka,Spouse
70958,17/02/2018,104,439,Food,Restaurant,Delhi,Delhi,Wife
09874,21/01/2018,105,509,Entertainment,Park,Kolkata,West Bengal
94585,19/01/2018,106,629,Rent,House,Hyderabad,Telangana,Self
45509,20/01/2018,107,953,Travel,Rail,Chennai,Tamil Nadu,Brother
07864,01/02/2018,108,569,Rent,Parking,Goa,Goa,Wife
[acadgild@localhost ~]$
```

Step 10: Load txn.txt file into TRANSACTIONS Table

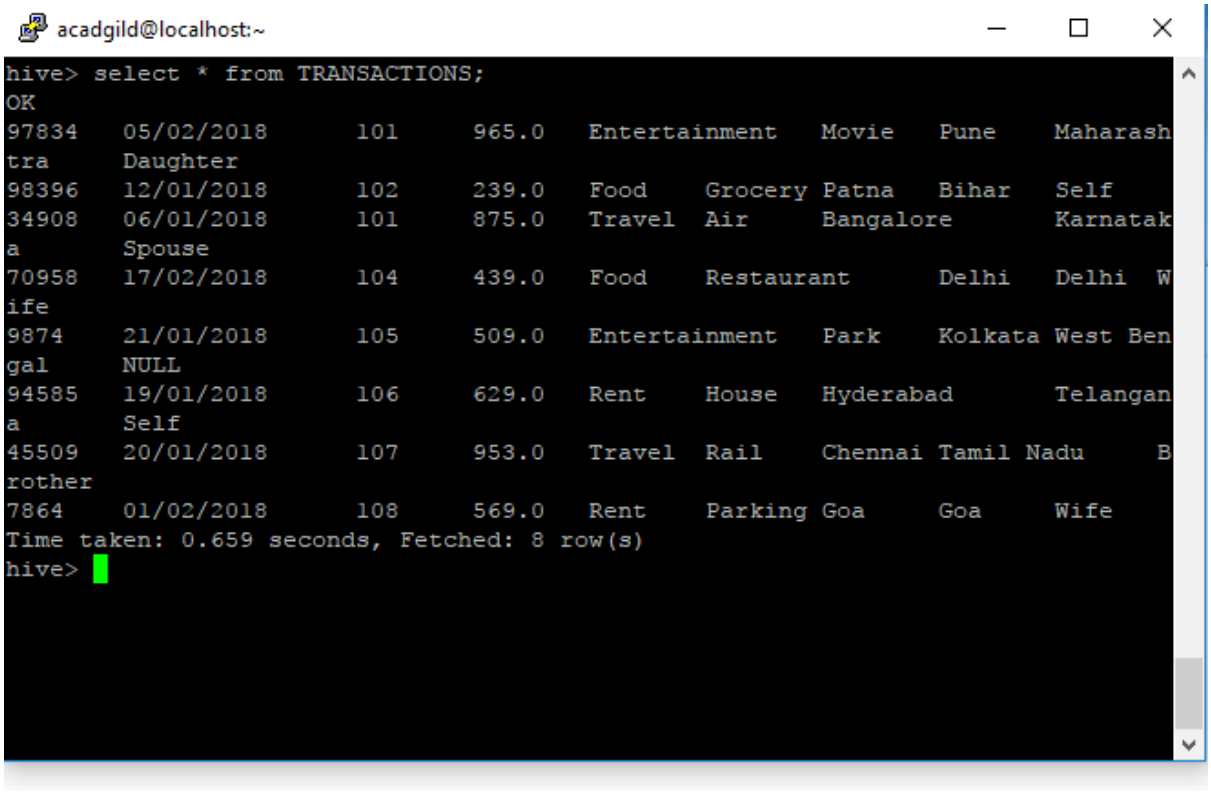
LOAD DATA LOCAL INPATH '/home/acadgild/transactions.txt'

into table TRANSACTIONS;

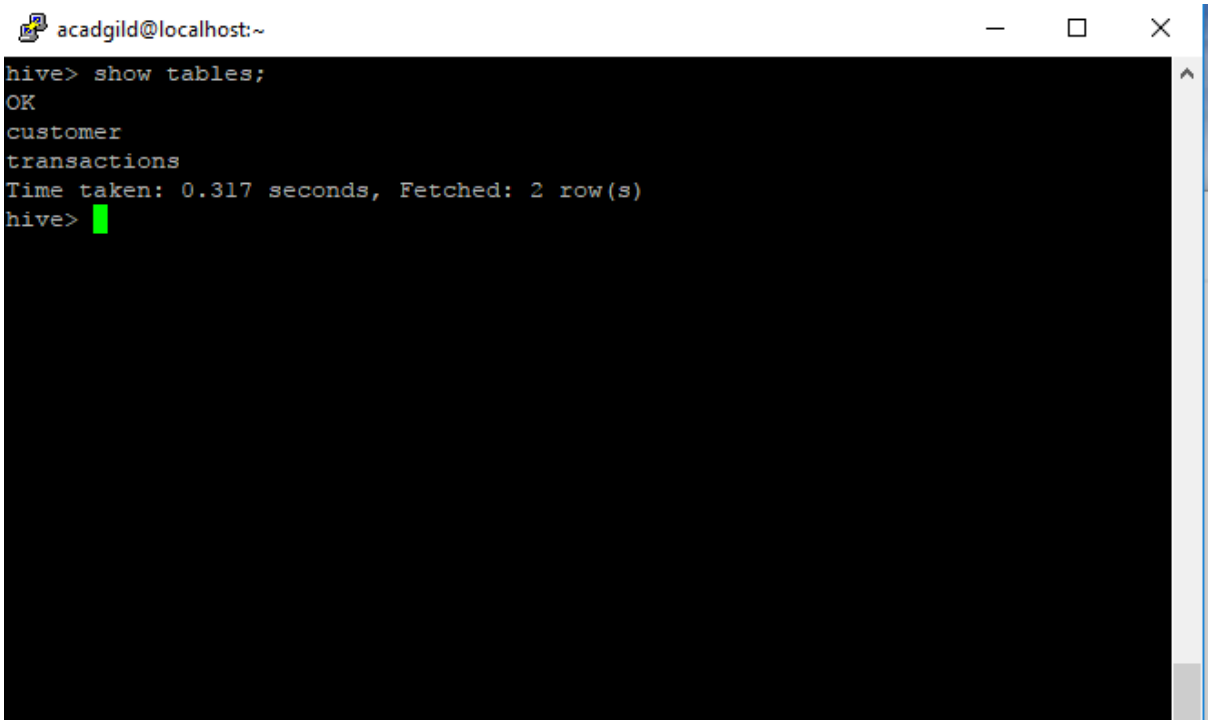
```
acadgild@localhost:~
hive> LOAD DATA LOCAL INPATH '/home/acadgild/txn.txt' into table TRANSACTIONS;
Loading data to table acadgilddb.transactions
OK
Time taken: 1.206 seconds
hive>
```

Step 11:

Select * From TRANSACTIONS;

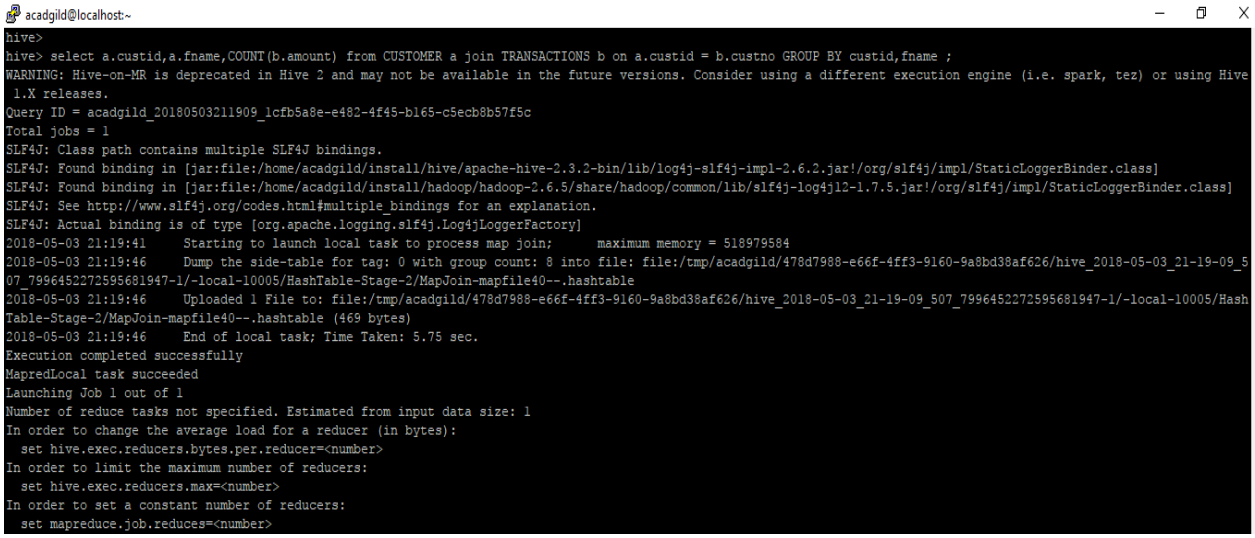


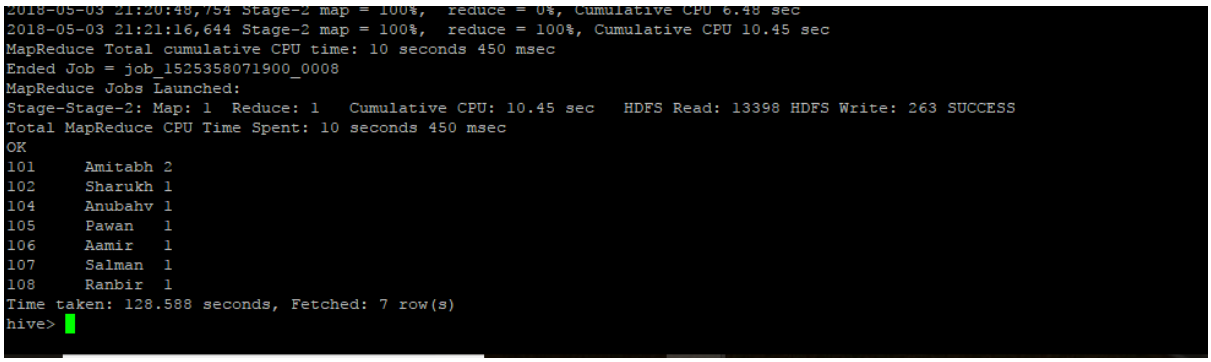
Step 12 : Show tables



1. **Find out the number of transaction done by each customer (These should be take up in module 8 itself**

   select a.custid,a.fname,COUNT(b.amount) from CUSTOMER a join TRANSACTIONS b on a.custid = b.custno GROUP BY custid,fname ;





2. **Create a new table called TRANSACTIONS_COUNT. This table should have 3 fields - custid, fname and count. (Again to be done in module 8)**

   CREATE TABLE TRANSACTIONS_COUNT(custid INT,fname STRING,count INT)row format delimited fields terminated by ',';

```
acadgild@localhost:~
hive> CREATE TABLE TRANSACTIONS_COUNT(custid INT,fname STRING,count INT)row format delimited fields terminated by ',';
OK
Time taken: 0.453 seconds
hive>
```

3. **Now write a hive query in such a way that the query populates the data obtained in Step 1 above and populate the table in step 2 above. (This has to be done in module 9).**

insert into table TRANSACTIONS_COUNT select a.custid,a.fname,COUNT(b.amount) from CUSTOMER a join TRANSACTIONS b on a.custid = b.custno GROUP BY custid,fname;

```
acadgild@localhost:~
hive> insert into table TRANSACTIONS_COUNT select a.custid,a.fname,COUNT(b.amount) from CUSTOMER a join TRANSACTIONS b on a.custid = b.custno GROUP BY custid,fname;
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 = acadgild_20180503212310_6dcea29c-5d10-4f21-9c68-9afff4d5bdde
Total jobs = 1
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hive/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/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/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.apache.logging.slf4j.Log4jLoggerFactory]
2018-05-03 21:23:33     Starting to launch local task to process map join;      maximum memory = 518979584
2018-05-03 21:23:42     Dump the side-table for tag: 0 with group count: 8 into file: file:/tmp/acadgild/478d7988-e66f-4ff3-9160-9a8bd38af626/hive_2018-05-03_21-23-10_8
28_2157845336841837900-1/-local-10003/HashTable-Stage-2/MapJoin-mapfile50--.hashtable
2018-05-03 21:23:43     Uploaded 1 File to: file:/tmp/acadgild/478d7988-e66f-4ff3-9160-9a8bd38af626/hive_2018-05-03_21-23-10_828_2157845336841837900-1/-local-10003/Hash
Table-Stage-2/MapJoin-mapfile50--.hashtable (469 bytes)
2018-05-03 21:23:43     End of local task; Time Taken: 9.511 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
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_1525358071900_0009, Tracking URL = http://localhost:8088/proxy/application_1525358071900_0009/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1525358071900_0009
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2018-05-03 21:24:10,625 Stage-2 map = 0%,  reduce = 0%
2018-05-03 21:24:35,182 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 5.12 sec
2018-05-03 21:25:01,565 Stage-2 map = 100%,  reduce = 67%, Cumulative CPU 9.45 sec
2018-05-03 21:25:05,118 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 11.32 sec
MapReduce Total cumulative CPU time: 11 seconds 320 msec
Ended Job = job_1525358071900_0009
Loading data to table acadgilddb.transactions_count
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 11.32 sec   HDFS Read: 14132 HDFS Write: 177 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 320 msec
OK
Time taken: 118.721 seconds
hive>
```

select * from TRANSACTIONS_COUNT;

```
acadgild@localhost:~
hive> select * from TRANSACTIONS_COUNT;
OK
101     Amitabh 2
102     Sharukh 1
104     Anubahv 1
105     Pawan   1
106     Aamir   1
107     Salman  1
108     Ranbir  1
Time taken: 0.677 seconds, Fetched: 7 row(s)
hive>
```

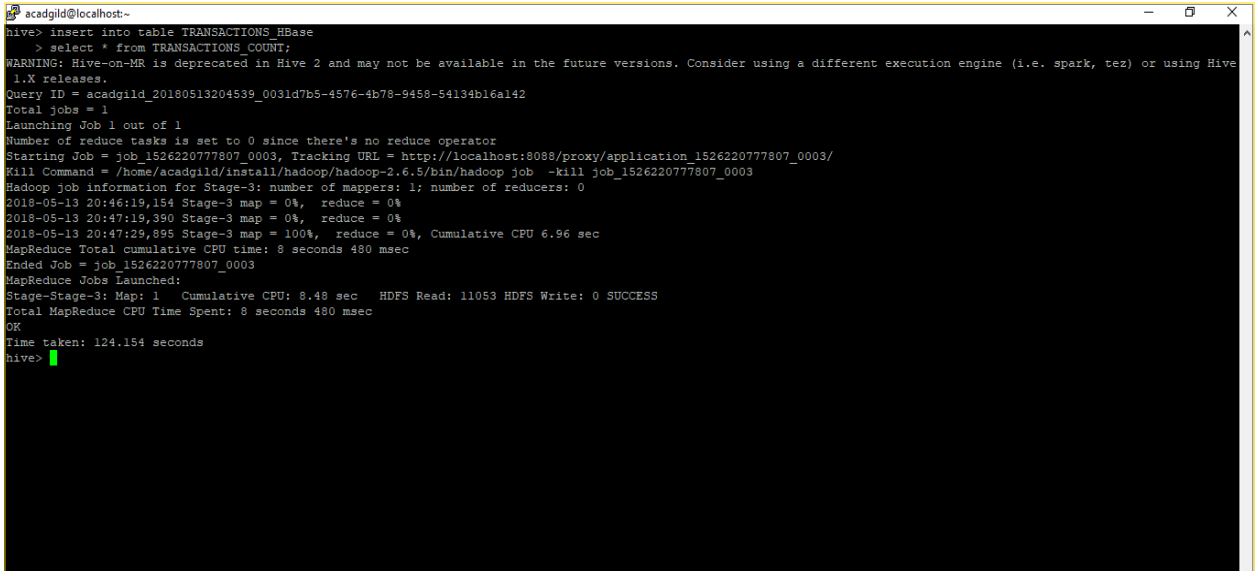
4. **Now lets make the TRANSACTIONS_COUNT table Hbase complaint. In the sence, use Ser Des And Storate handler features of hive to change the TRANSACTIONS_COUNT table to be able to create a TRANSACTIONS table in Hbase. (This has to be done in module 10)**

CREATE TABLE TRANSACTIONS_HBase(ID INT,username STRING, count INT) STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'WITH SERDEPROPERTIES ('hbase.columns.mapping'':key,cf1:username,cf2:count');

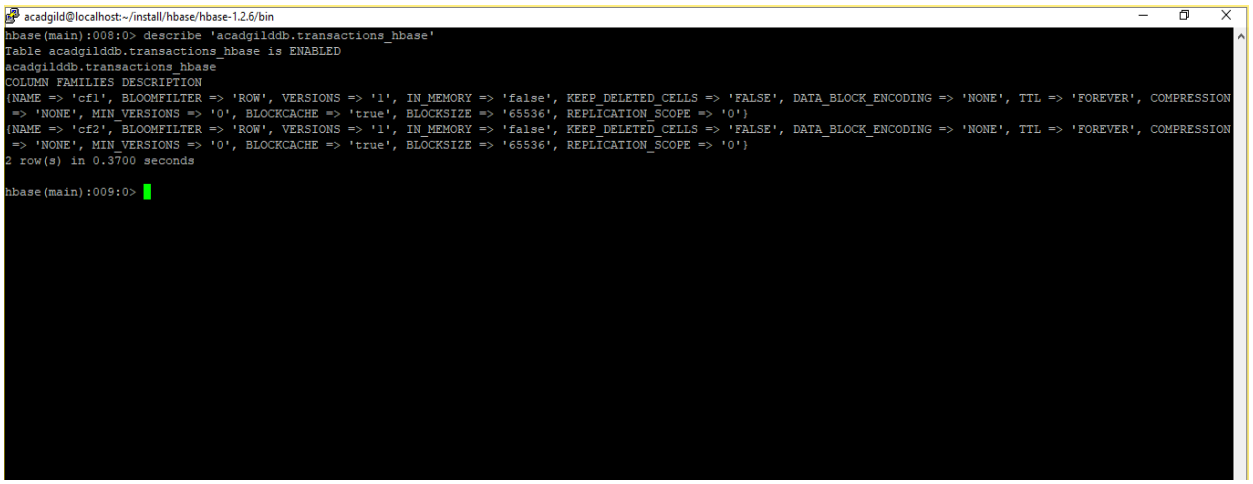
```
acadgild@localhost:~
hive> CREATE TABLE TRANSACTIONS_HBase(ID INT,username STRING, count INT)
    > STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
    > WITH SERDEPROPERTIES ('hbase.columns.mapping' = ':key,cf1:username,cf2:cou
nt');
OK
Time taken: 2.904 seconds
hive>
```

5. **Now insert the data in TRANSACTIONS_COUNT table using the query in step 3 again, this should populate the Hbase TRANSACTIONS table automatically (This has to be done in module 10)**

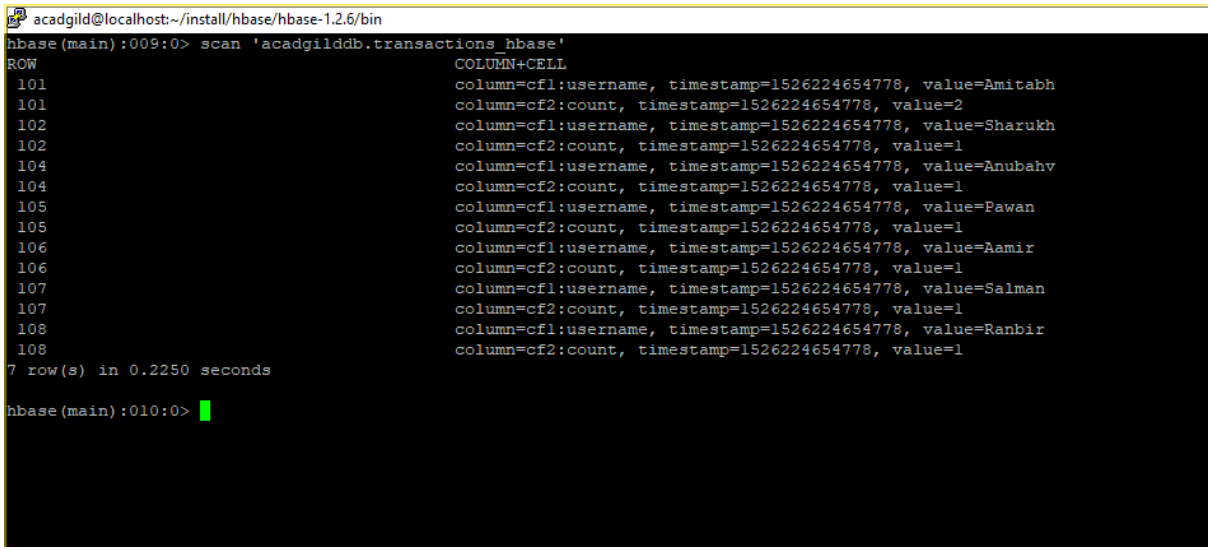
insert into table TRANSACTIONS_HBase
select * from TRANSACTIONS_COUNT;



In HBASE:
describe 'acadgilddb.transactions_hbase'



scan 'acadgilddb.transactions_hbase'



6. **Now from the Hbase level, write the Hbase java API code to access and scan the TRANSACTIONS table data from java level.**

Step 1:

```
import org.apache.spark.{SparkConf, SparkContext}
import org.apache.spark.rdd.RDD
import org.apache.spark.streaming.{Seconds, StreamingContext, Time}
import org.apache.spark.sql.SparkSession
import org.apache.log4j.{Level,Logger}


object SqlNetworkWordCount {


  def main(args: Array[String]): Unit = {
    println("hey Spark SQL Streaming")
   val conf = new
SparkConf().setMaster("local[2]").setAppName("SparkSteamingExample")
    val sc = new SparkContext(conf)
   val rootLogger =Logger.getRootLogger()
```

```
rootLogger.setLevel(Level.ERROR)
 println("hey Spark Streaming ---> 1")
 //val sparkConf = new SparkConf().setAppName("NetworkWordCount")
 println("hey Spark Streaming ---> 2")
 val ssc = new StreamingContext(sc, Seconds(10))
 val lines = ssc.socketTextStream("localhost", 9999)
 println("hey Spark Streaming ---> 3")
 val words = lines.flatMap(_.split(" "))

 // Convert RDDs of the words DStream to DataFrame and run SQL query
 words.foreachRDD { (rdd: RDD[String], time: Time) =>
   val spark = SparkSessionSingleton.getInstance(rdd.sparkContext.getConf)
   import spark.implicits._

// Convert RDD[String] to RDD[case class] to DataFrame
   val wordsDataFrame = rdd.map(w => Record(w)).toDF()

 // Creates a temporary view using the DataFrame
   wordsDataFrame.createOrReplaceTempView("words")
// Do word count on table using SQL and print it
   val wordCountsDataFrame =
     spark.sql("select word, count(*) as total from words group by word")
   println(s"========= $time =========")
   wordCountsDataFrame.show()
 }
 ssc.start()
```

```
  ssc.awaitTermination()
 }
 /** Case class for converting RDD to DataFrame */
 case class Record(word: String)
 /** Lazily instantiated singleton instance of SparkSession */
 object SparkSessionSingleton {
@transient private var instance: SparkSession = _
  def getInstance(sparkConf: SparkConf): SparkSession = {
   if (instance == null) {
    instance = SparkSession
     .builder
     .config(sparkConf)
     .getOrCreate()
   }
   instance
  } } }
```

SqlNetworkWordCount.scala

```scala

import org.apache.spark.{SparkConf, SparkContext}

object SqlNetworkWordCount {

  def main(args: Array[String]): Unit = {
    println("hey Spark SQL Streaming");

    val conf = new SparkConf().setMaster("local[2]").setAp
    val sc = new SparkContext(conf);
   val rootLogger =Logger.getRootLogger();
   rootLogger.setLevel(Level.ERROR);


    println("hey Spark Streaming ---> 1");
    //val sparkConf = new SparkConf().setAppName("NetworkW
    println("hey Spark Streaming ---> 2");
    val ssc = new StreamingContext(sc, Seconds(10));
    val lines = ssc.socketTextStream("localhost", 9999);
    println("hey Spark Streaming ---> 3");
```

SqlNetworkWordCount.scala

```scala
    val words = lines.flatMap(_.split(" "));

    // Convert RDDs of the words DStream to DataFrame and
    words.foreachRDD { (rdd: RDD[String], time: Time) =>
      val spark = SparkSessionSingleton.getInstance(rdd.sp
      import spark.implicits._;

      // Convert RDD[String] to RDD[case class] to DataFra
      val wordsDataFrame = rdd.map(w => Record(w)).toDF();

      // Creates a temporary view using the DataFrame
      wordsDataFrame.createOrReplaceTempView("words");

      // Do word count on table using SQL and print it
      val wordCountsDataFrame =
        spark.sql("select word, count(*) as total from wor
      println(s"========= $time =========")
      wordCountsDataFrame.show()
    }
```

SqlNetworkWordCount.scala

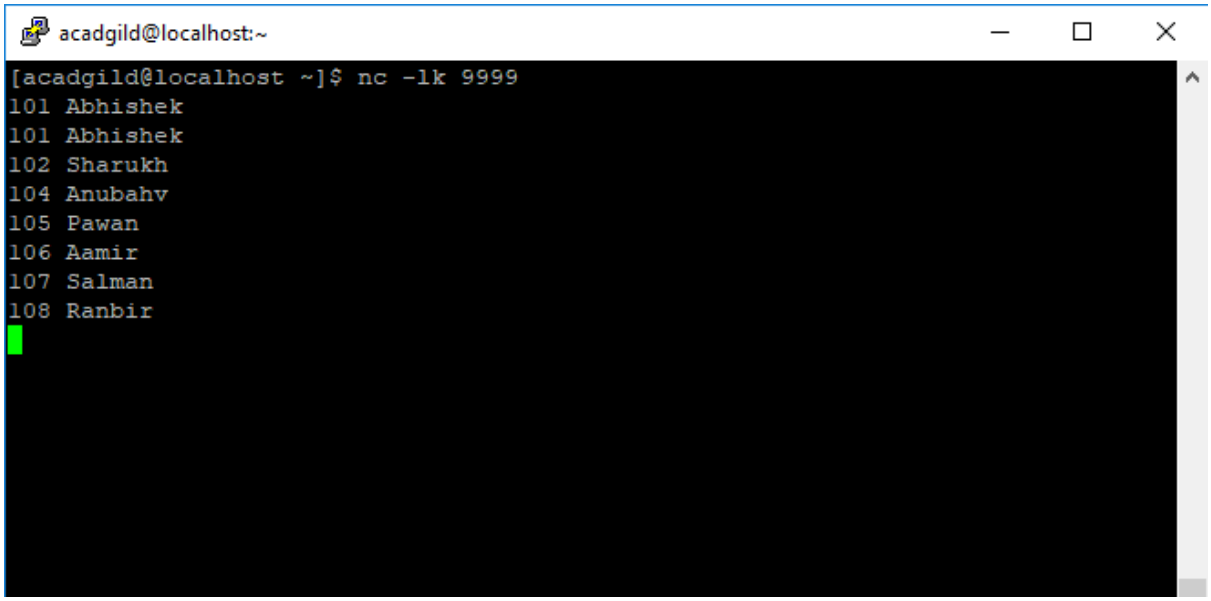
```
46
47      ssc.start()
48      ssc.awaitTermination()
49
50
51    }
52
53    /** Case class for converting RDD to DataFrame */
54    case class Record(word: String)
55
56    /** Lazily instantiated singleton instance of SparkSessi
57    object SparkSessionSingleton {
58
59      @transient private var instance: SparkSession = _
60
61      def getInstance(sparkConf: SparkConf): SparkSession =
62        if (instance == null) {
63          instance = SparkSession
64            .builder
65            .config(sparkConf)
```

SqlNetworkWordCount.scala

```
55
56    /** Lazily instantiated singleton instance of SparkSessi
57    object SparkSessionSingleton {
58
59      @transient private var instance: SparkSession = _
60
61      def getInstance(sparkConf: SparkConf): SparkSession =
62        if (instance == null) {
63          instance = SparkSession
64            .builder
65            .config(sparkConf)
66            .getOrCreate()
67        }
68        instance
69      }
70    }
71
72  }
73
74
```

Step 2: Start netcat from a terminal

nc -lk 9999



Step 3: Display Results:

101 Abhishek 2

102 Sharukh 1

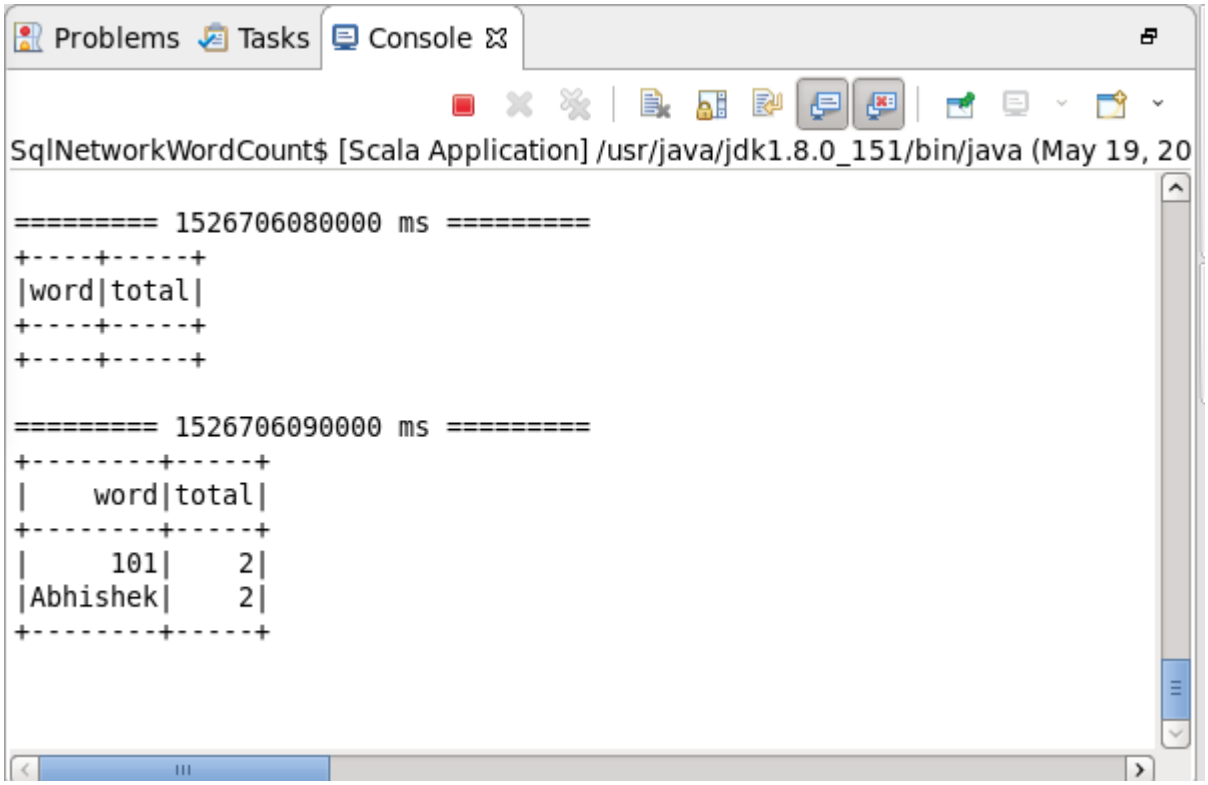
104 Anubahv 1

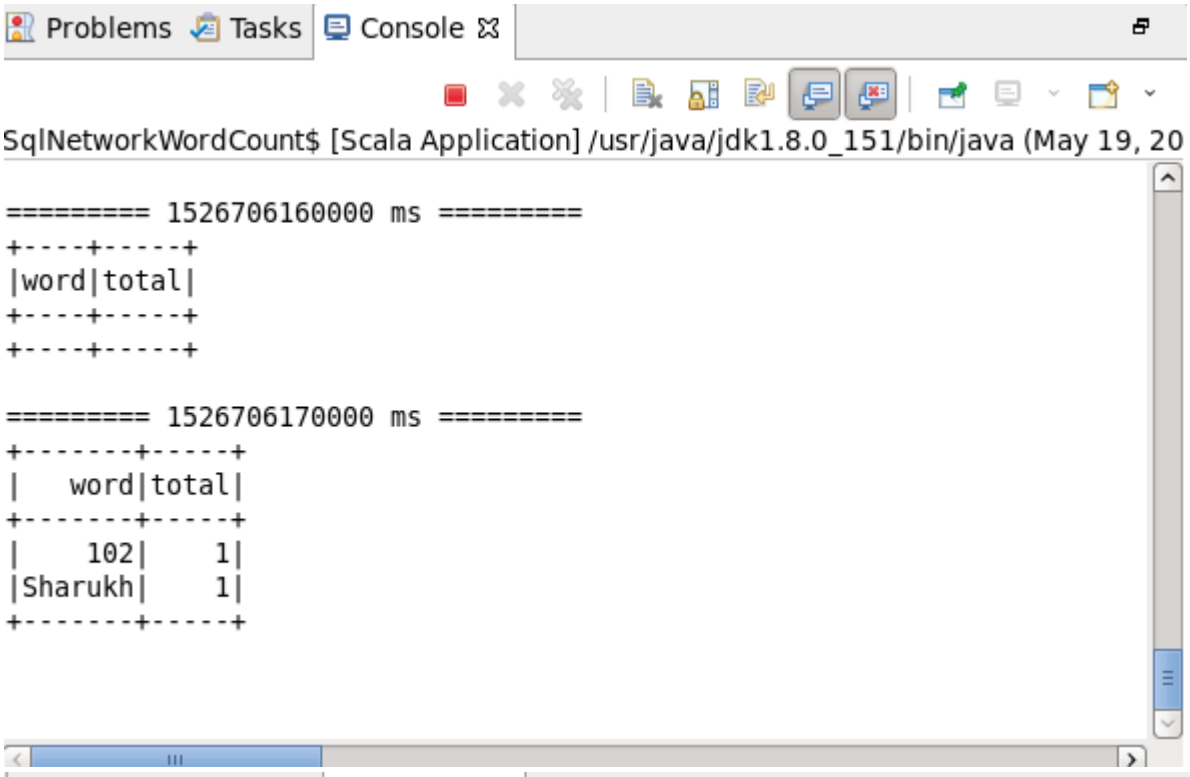
105 Pawan 1

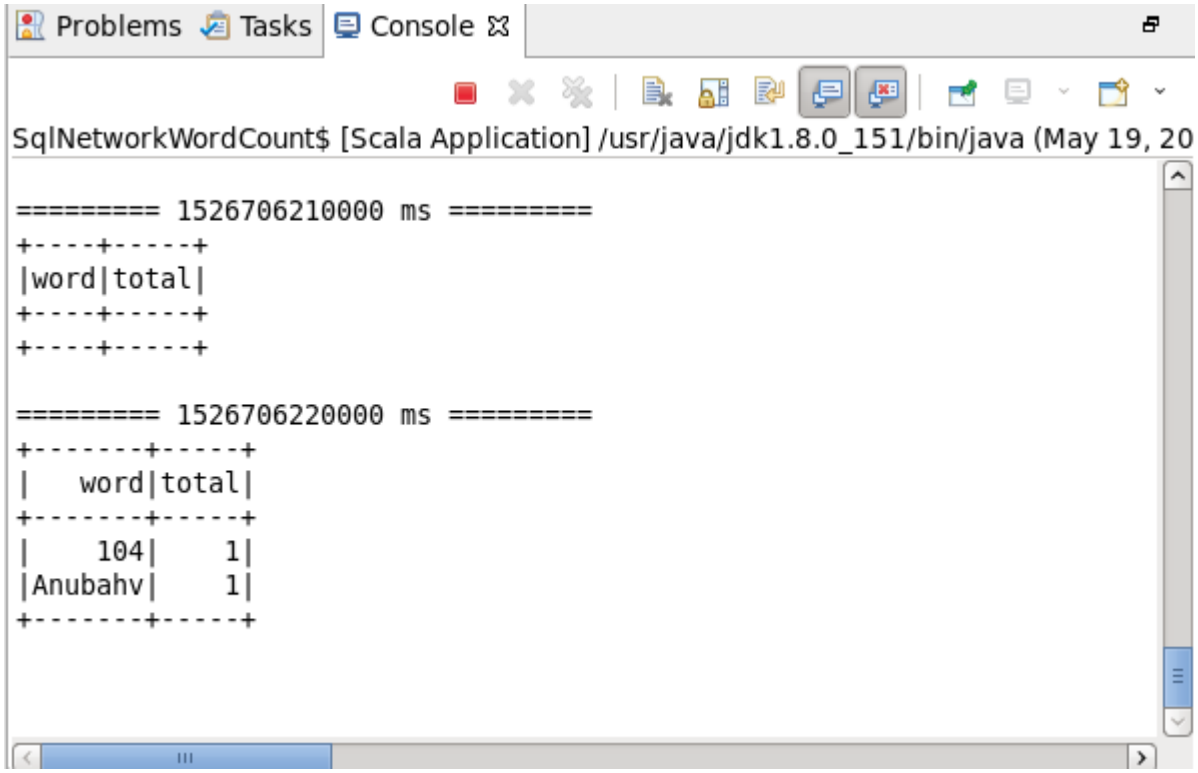
106 Aamir 1

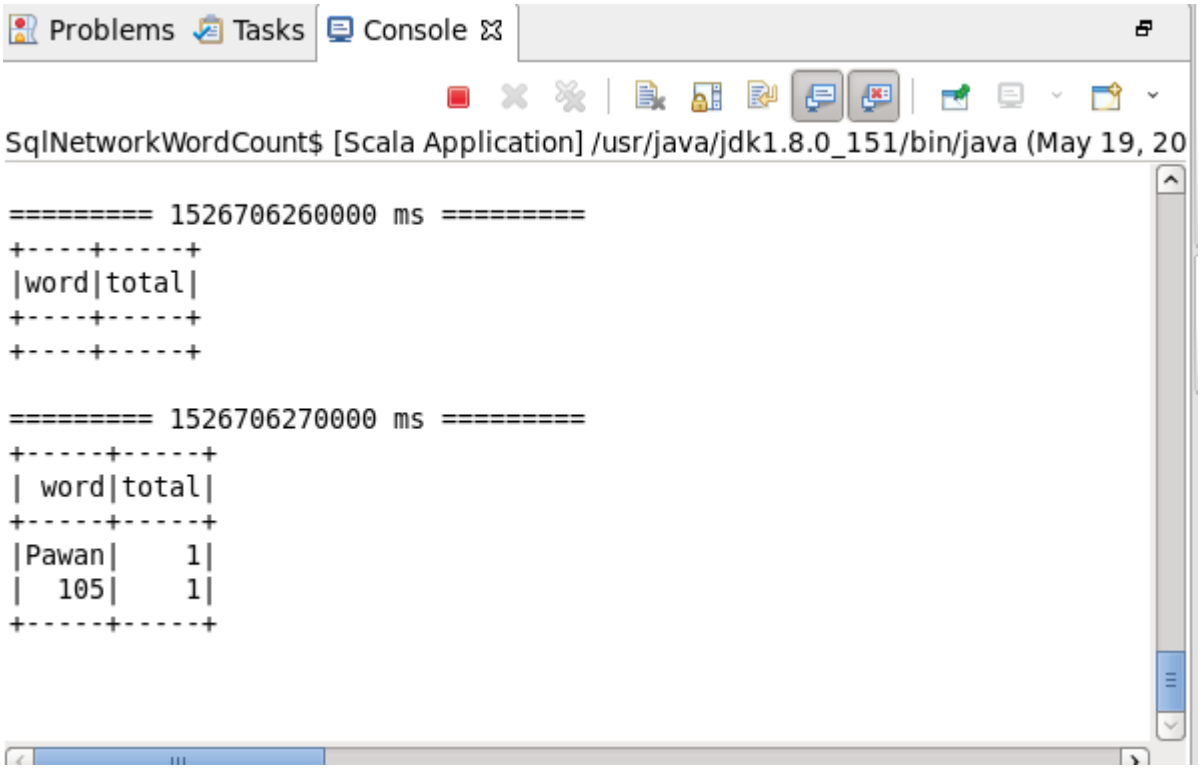
107 Salman 1

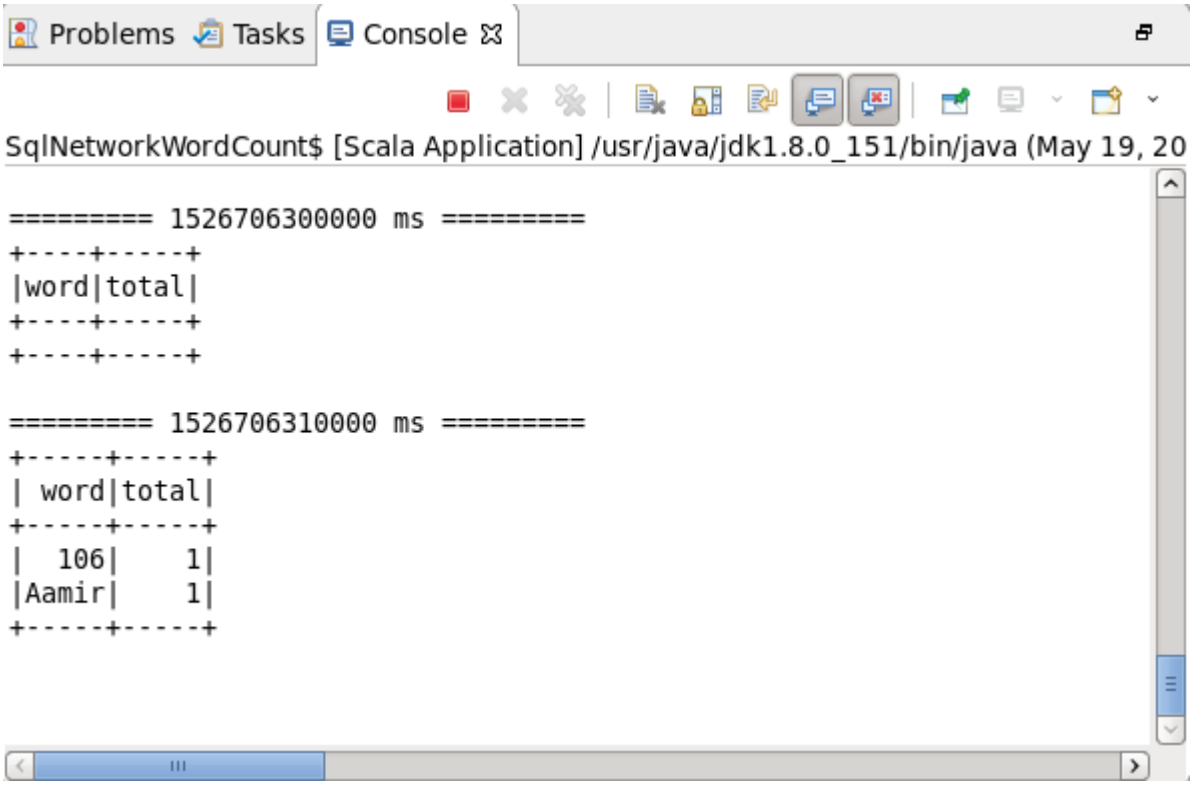
108 Ranbir 1

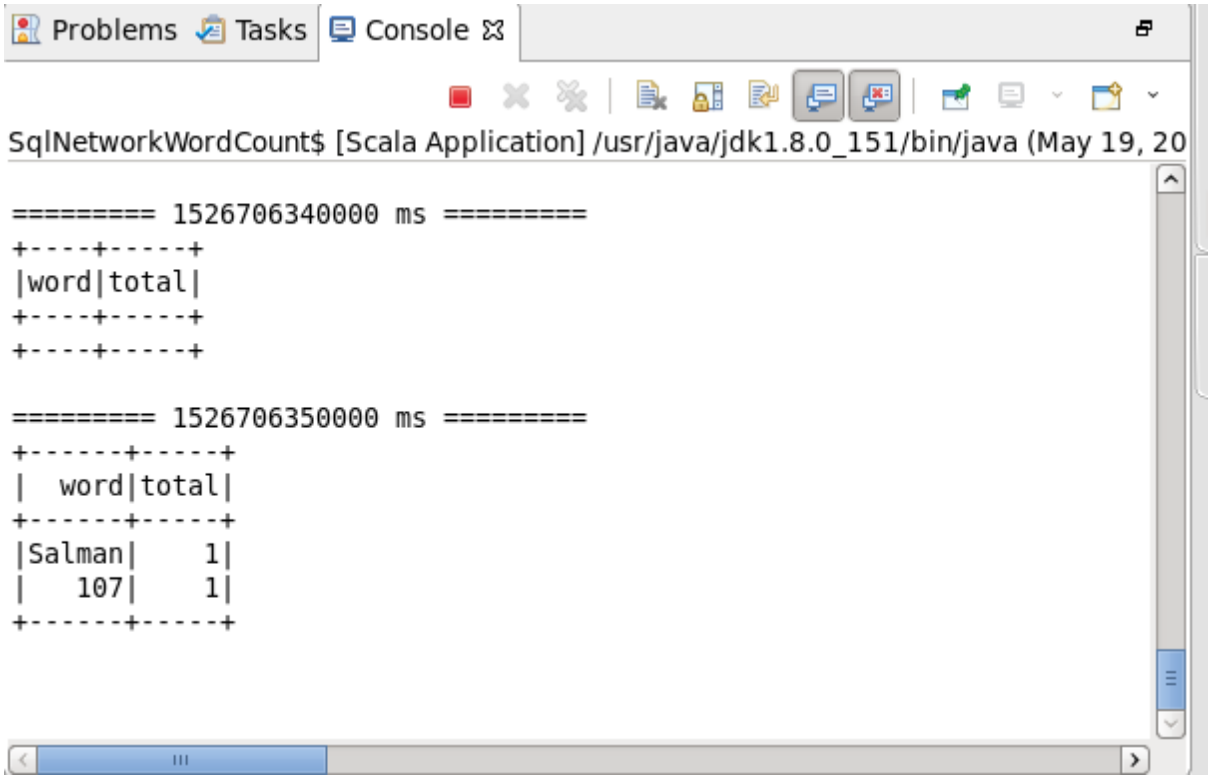

Problems
Tasks
Console
SqlNetworkWordCount$ [Scala Application] /usr/java/jdk1.8.0_151/bin/java (May 19, 20
========= 1526706080000 ms =========
+----+-----+
|word|total|
+----+-----+
+----+-----+
========= 1526706090000 ms =========
+--------+-----+
| word|total|
+--------+-----+
| 101| 2|
|Abhishek| 2|
+--------+-----+

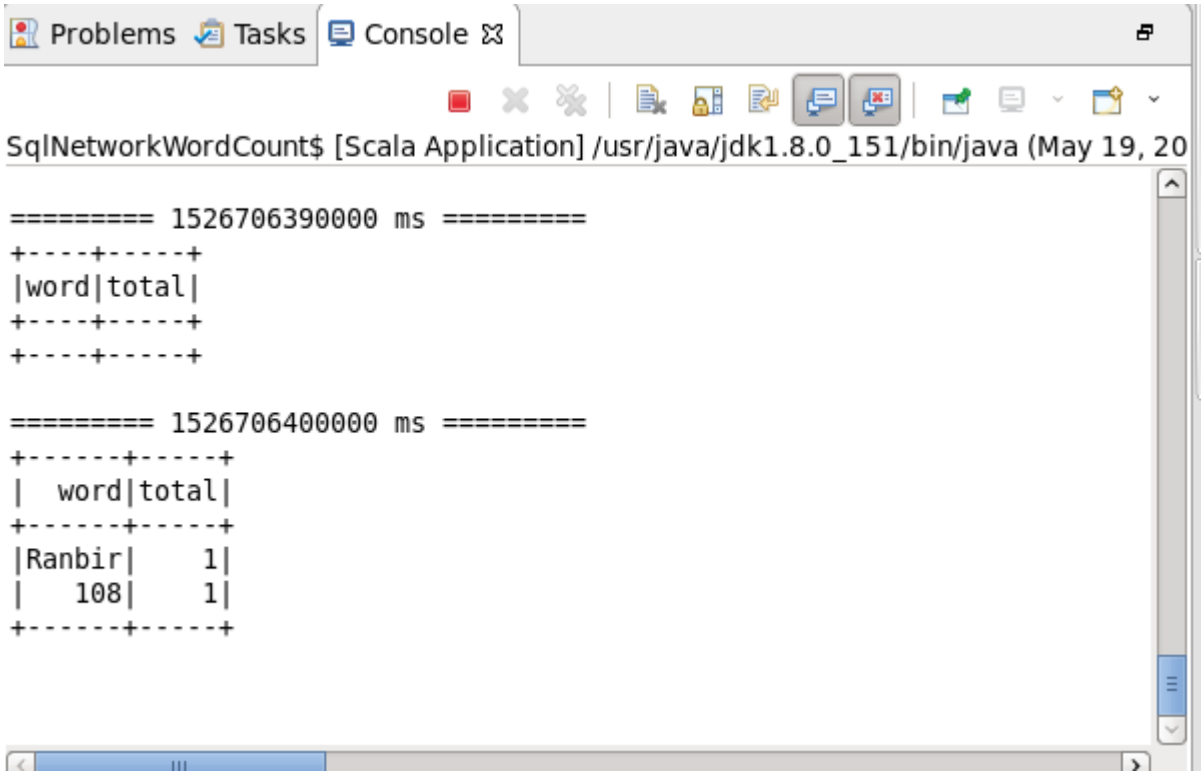

Problems
Tasks
Console
SqlNetworkWordCount$ [Scala Application] /usr/java/jdk1.8.0_151/bin/java (May 19, 20
========= 1526706160000 ms =========
+----+-----+
|word|total|
+----+-----+
+----+-----+
========= 1526706170000 ms =========
+-------+-----+
| word|total|
+-------+-----+
| 102| 1|
|Sharukh| 1|
+-------+-----+


Problems
Tasks
Console
SqlNetworkWordCount$ [Scala Application] /usr/java/jdk1.8.0_151/bin/java (May 19, 20
========= 1526706210000 ms =========
+----+-----+
|word|total|
+----+-----+
+----+-----+
========= 1526706220000 ms =========
+-------+-----+
| word|total|
+-------+-----+
| 104| 1|
|Anubahv| 1|
+-------+-----+


Problems
Tasks
Console
SqlNetworkWordCount$ [Scala Application] /usr/java/jdk1.8.0_151/bin/java (May 19, 20
========= 1526706260000 ms =========
+----+-----+
|word|total|
+----+-----+
+----+-----+
========= 1526706270000 ms =========
+-----+-----+
| word|total|
+-----+-----+
|Pawan| 1|
| 105| 1|
+-----+-----+


Problems
Tasks
Console
SqlNetworkWordCount$ [Scala Application] /usr/java/jdk1.8.0_151/bin/java (May 19, 20
========= 1526706300000 ms =========
+----+-----+
|word|total|
+----+-----+
+----+-----+
========= 1526706310000 ms =========
+-----+-----+
| word|total|
+-----+-----+
| 106| 1|
|Aamir| 1|
+-----+-----+


Problems
Tasks
Console
SqlNetworkWordCount$ [Scala Application] /usr/java/jdk1.8.0_151/bin/java (May 19, 20
========= 1526706340000 ms =========
+----+-----+
|word|total|
+----+-----+
+----+-----+
========= 1526706350000 ms =========
+------+-----+
| word|total|
+------+-----+
|Salman| 1|
| 107| 1|
+------+-----+


Problems
Tasks
Console
SqlNetworkWordCount$ [Scala Application] /usr/java/jdk1.8.0_151/bin/java (May 19, 20
========= 1526706390000 ms =========
+----+-----+
|word|total|
+----+-----+
+----+-----+
========= 1526706400000 ms =========
+------+-----+
| word|total|
+------+-----+
|Ranbir| 1|
| 108| 1|
+------+-----+