Session 25:
BIG DATA ECOSYSTEM NTEGRATION
Assignment 1

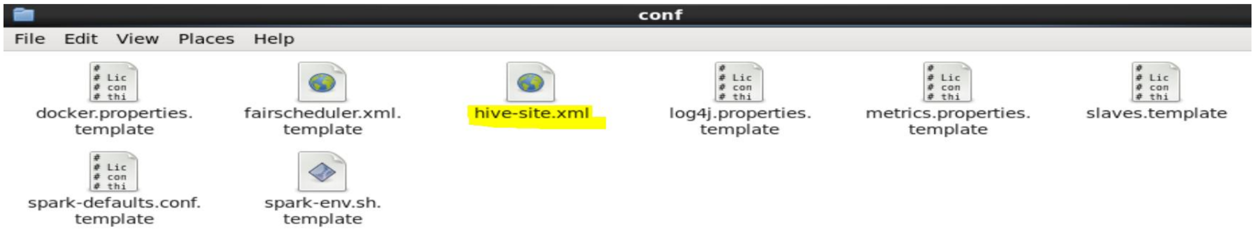
-Prachi Mohite

Task 1
As discussed in class integrate Spark Hive

To demonstrate this we have a program which will list the databases in HIVE

Steps to be followed

1. Copy the hive-site.xml file from $HIVE_HOME/conf to $SPARK_HOME/conf


conf
File Edit View Places Help
docker.properties.template
fairscheduler.xml.template
hive-site.xml
log4j.properties.template
metrics.properties.template
slaves.template
spark-defaults.conf.template
spark-env.sh.template

2. Add the following properties to hive-site.xml on spark side :

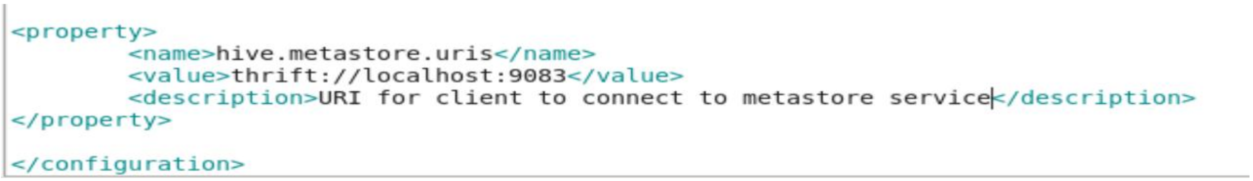
    <property>

    <name>hive.metastore.uris</name>

    <value>thrift://localhost:9083</value>

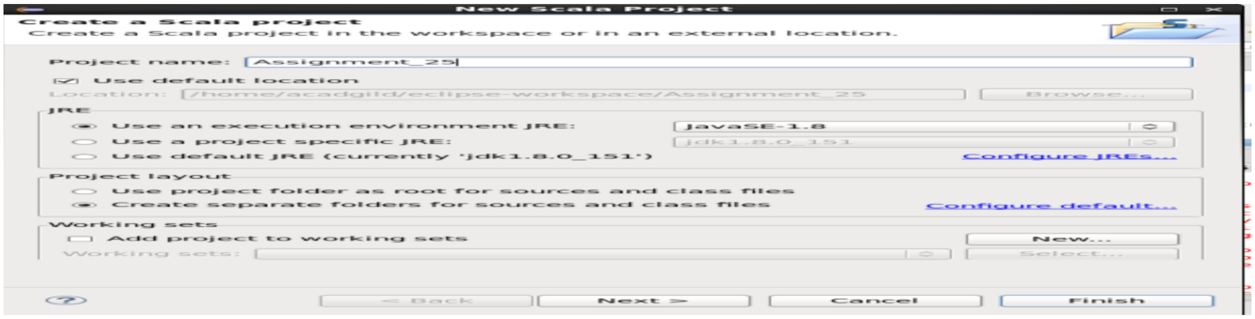
    <description>URI Client to connect to metastore service</description>

    </property>



```
<property>
        <name>hive.metastore.uris</name>
        <value>thrift://localhost:9083</value>
        <description>URI for client to connect to metastore service</description>
</property>

</configuration>
```

3. Downloaded the code from here and added to a project created as Assignment_25 as below



New Scala Project

Create a Scala project

Create a Scala project in the workspace or in an external location.

Project name: Assignment_25

☑ Use default location

Location: /home/acadgild/eclipse-workspace/Assignment_25

JRE

● Use an execution environment JRE: JavaSE-1.8

○ Use a project specific JRE: jdk1.8.0_151

○ Use default JRE (currently 'jdk1.8.0_151')

Project layout

○ Use project folder as root for sources and class files

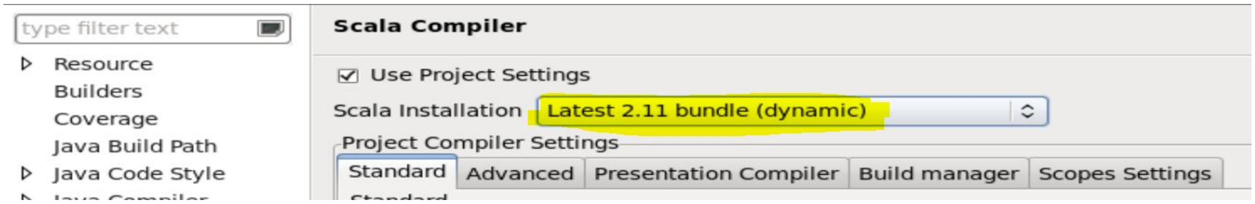
● Create separate folders for sources and class files

Working sets

□ Add project to working sets

Working sets:

4. After adding the file to project change the scala compiler to Scala 2.11 as below

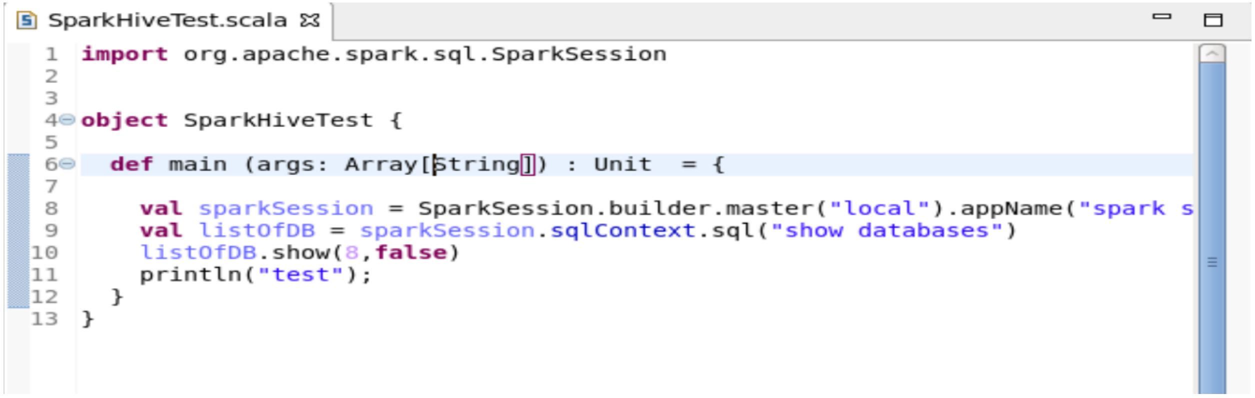


Scala Compiler

☑ Use Project Settings

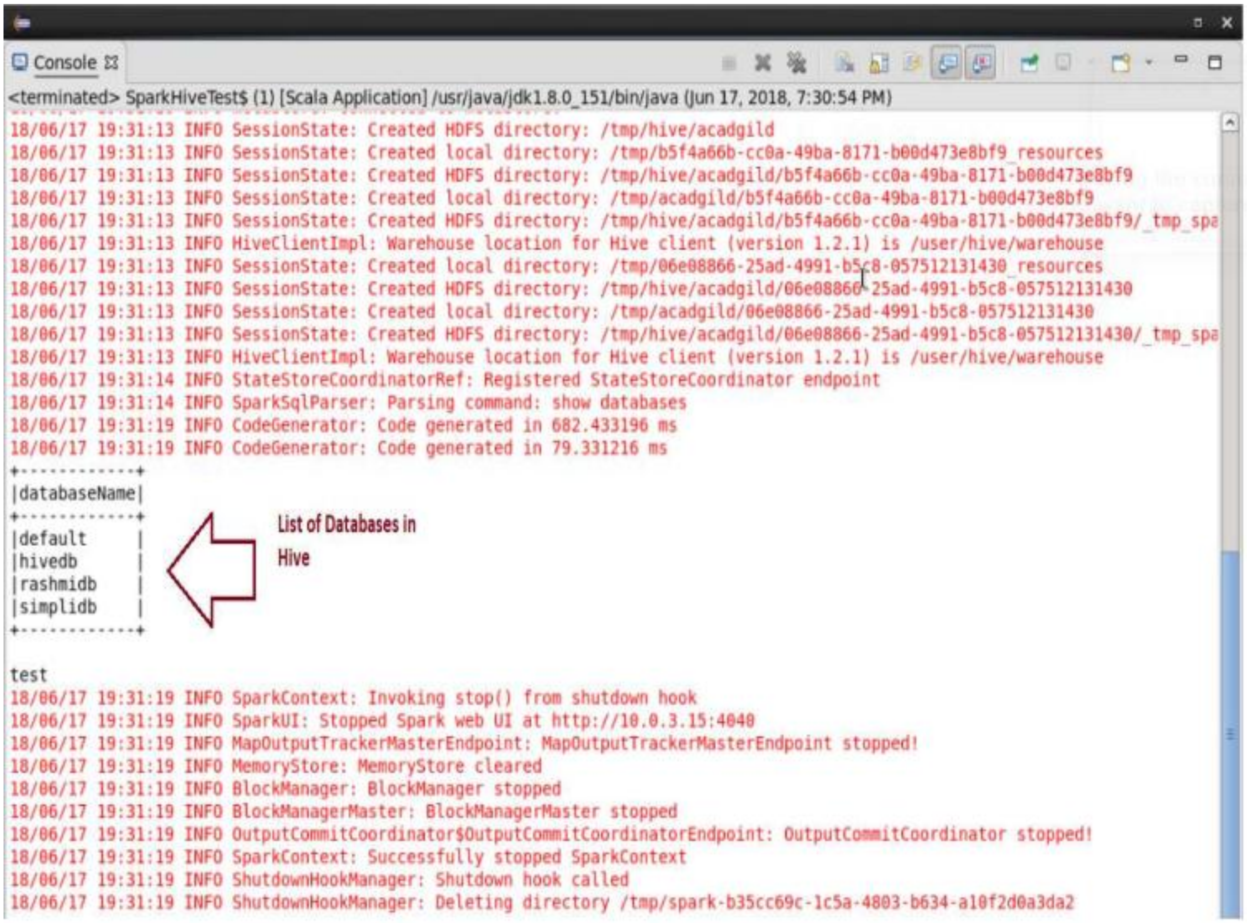
Scala Installation: Latest 2.11 bundle (dynamic)

5. Add Spark jars and hbase jars to the project



```scala
import org.apache.spark.sql.SparkSession


object SparkHiveTest {

  def main (args: Array[String]) : Unit  = {

    val sparkSession = SparkSession.builder.master("local").appName("spark s
    val listOfDB = sparkSession.sqlContext.sql("show databases")
    listOfDB.show(8,false)
    println("test");
  }
}
```

6. Execute the code and verify the list of databases



```
<terminated> SparkHiveTest$ (1) [Scala Application] /usr/java/jdk1.8.0_151/bin/java (Jun 17, 2018, 7:30:54 PM)
18/06/17 19:31:13 INFO SessionState: Created HDFS directory: /tmp/hive/acadgild
18/06/17 19:31:13 INFO SessionState: Created local directory: /tmp/b5f4a66b-cc0a-49ba-8171-b00d473e8bf9_resources
18/06/17 19:31:13 INFO SessionState: Created HDFS directory: /tmp/hive/acadgild/b5f4a66b-cc0a-49ba-8171-b00d473e8bf9
18/06/17 19:31:13 INFO SessionState: Created local directory: /tmp/acadgild/b5f4a66b-cc0a-49ba-8171-b00d473e8bf9
18/06/17 19:31:13 INFO SessionState: Created HDFS directory: /tmp/hive/acadgild/b5f4a66b-cc0a-49ba-8171-b00d473e8bf9/_tmp_spa
18/06/17 19:31:13 INFO HiveClientImpl: Warehouse location for Hive client (version 1.2.1) is /user/hive/warehouse
18/06/17 19:31:13 INFO SessionState: Created local directory: /tmp/06e08866-25ad-4991-b5c8-057512131430_resources
18/06/17 19:31:13 INFO SessionState: Created HDFS directory: /tmp/hive/acadgild/06e08866-25ad-4991-b5c8-057512131430
18/06/17 19:31:13 INFO SessionState: Created local directory: /tmp/acadgild/06e08866-25ad-4991-b5c8-057512131430
18/06/17 19:31:13 INFO SessionState: Created HDFS directory: /tmp/hive/acadgild/06e08866-25ad-4991-b5c8-057512131430/_tmp_spa
18/06/17 19:31:13 INFO HiveClientImpl: Warehouse location for Hive client (version 1.2.1) is /user/hive/warehouse
18/06/17 19:31:14 INFO StateStoreCoordinatorRef: Registered StateStoreCoordinator endpoint
18/06/17 19:31:14 INFO SparkSqlParser: Parsing command: show databases
18/06/17 19:31:19 INFO CodeGenerator: Code generated in 682.433196 ms
18/06/17 19:31:19 INFO CodeGenerator: Code generated in 79.331216 ms
+------------+
|databaseName|
+------------+
|default     |
|hivedb      |
|rashmidb    |
|simplidb    |
+------------+

test
18/06/17 19:31:19 INFO SparkContext: Invoking stop() from shutdown hook
18/06/17 19:31:19 INFO SparkUI: Stopped Spark web UI at http://10.0.3.15:4040
18/06/17 19:31:19 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/06/17 19:31:19 INFO MemoryStore: MemoryStore cleared
18/06/17 19:31:19 INFO BlockManager: BlockManager stopped
18/06/17 19:31:19 INFO BlockManagerMaster: BlockManagerMaster stopped
18/06/17 19:31:19 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
18/06/17 19:31:19 INFO SparkContext: Successfully stopped SparkContext
18/06/17 19:31:19 INFO ShutdownHookManager: Shutdown hook called
18/06/17 19:31:19 INFO ShutdownHookManager: Deleting directory /tmp/spark-b35cc69c-1c5a-4803-b634-a10f2d0a3da2
```

List of Databases in Hive

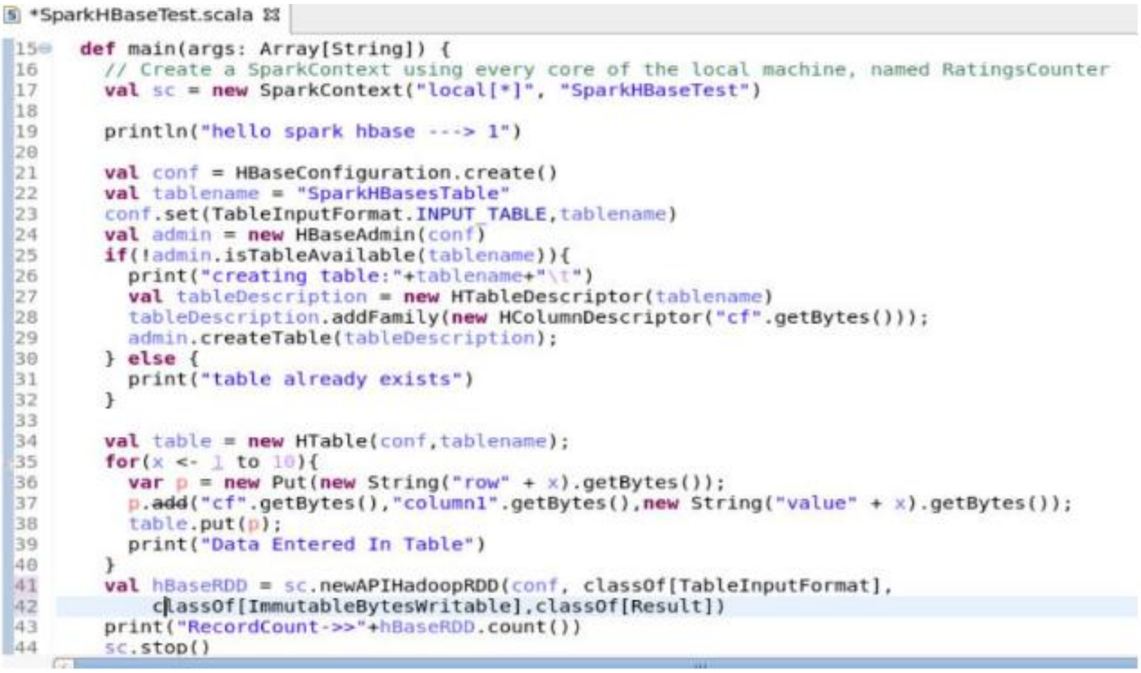
Task 2
As discussed in class integrate Spark HBase

As we have already done some pre-requisite steps. Now download the code and add to the project. This code will integrate with HBase through spark and create a table and insert contents in the same.

Add the required jar files for the HBase.

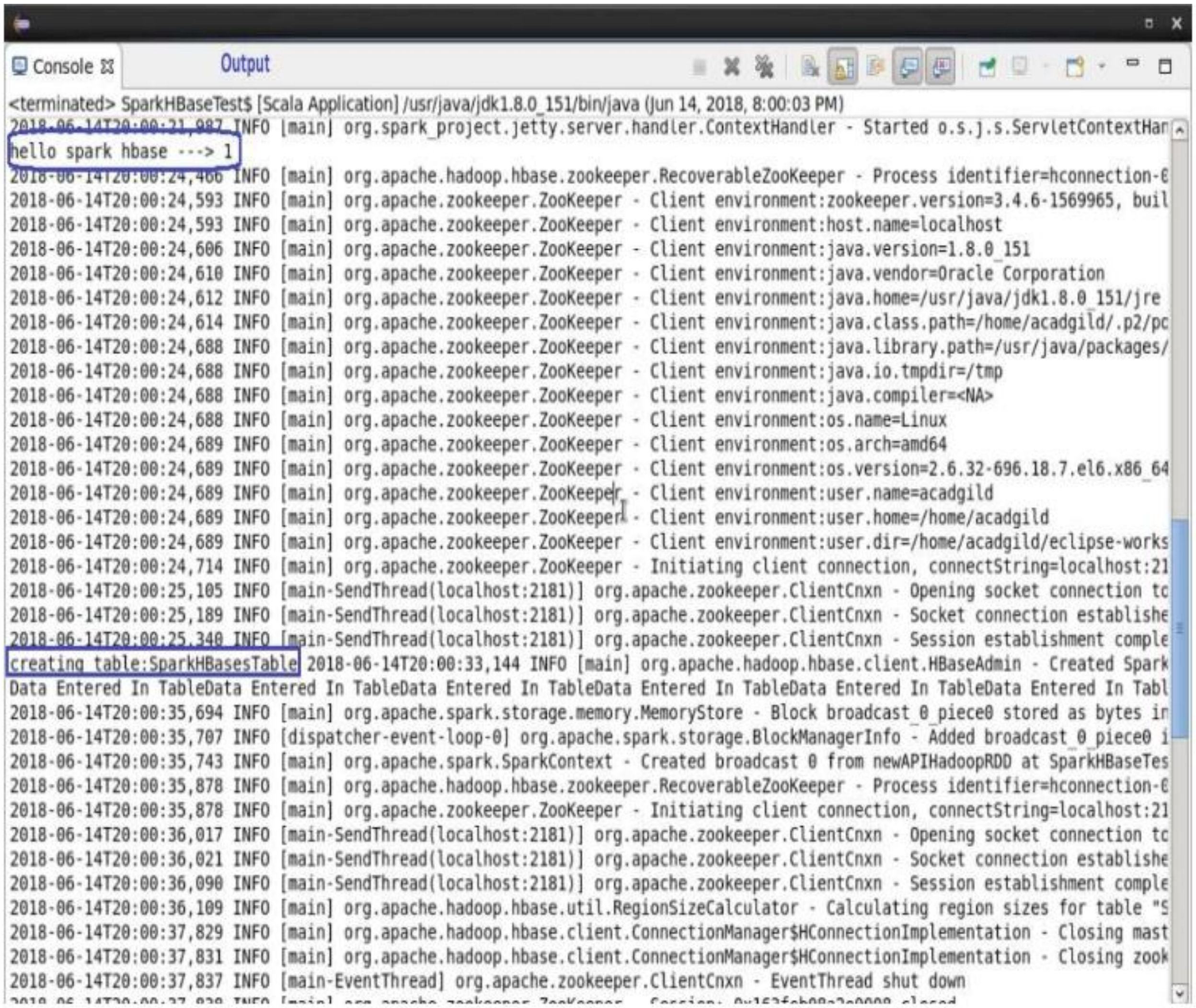
Make sure HBase is started while executing the code.

Added Code as below

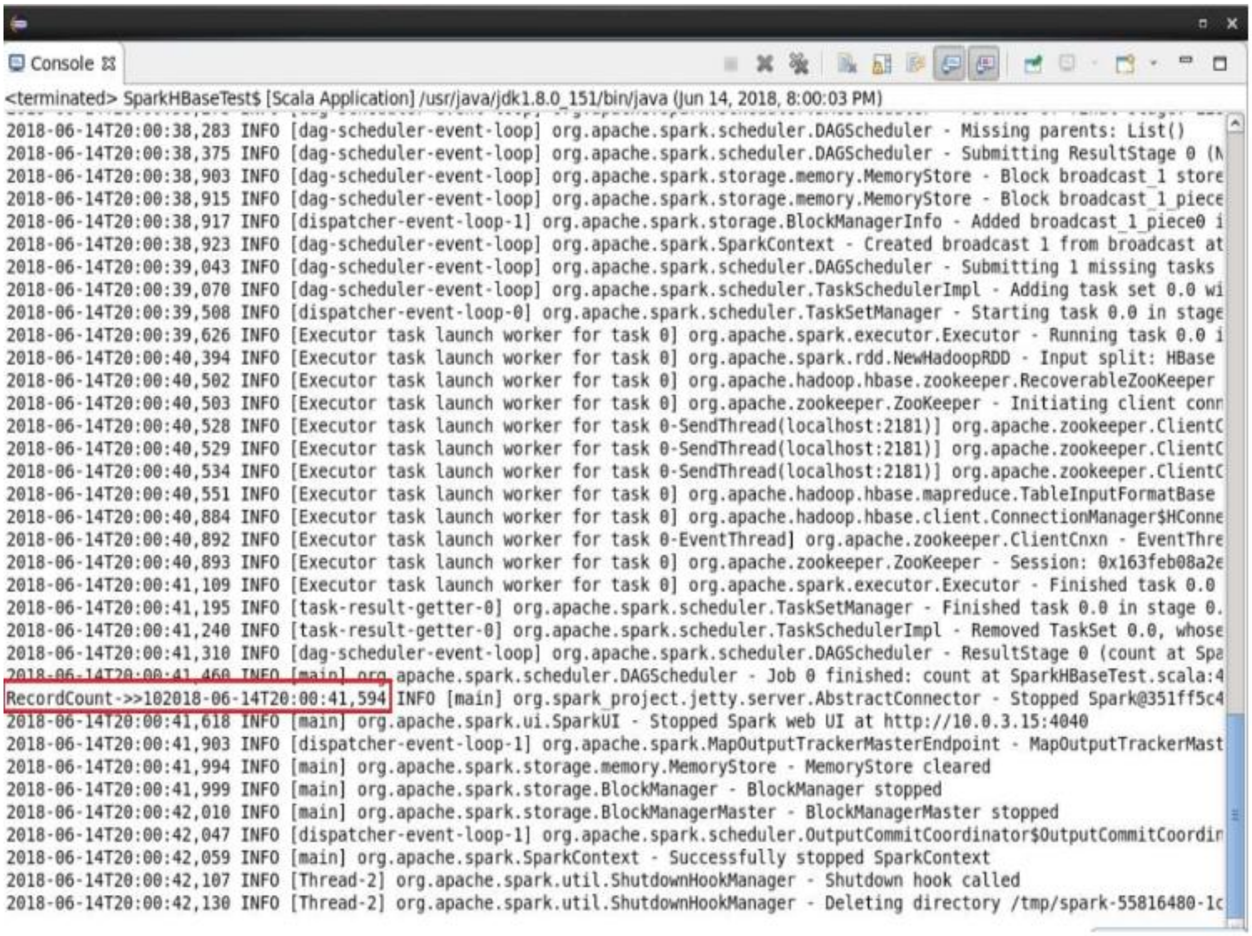


```
def main(args: Array[String]) {
  // Create a SparkContext using every core of the local machine, named RatingsCounter
  val sc = new SparkContext("local[*]", "SparkHBaseTest")

  println("hello spark hbase ---> 1")

  val conf = HBaseConfiguration.create()
  val tablename = "SparkHBasesTable"
  conf.set(TableInputFormat.INPUT_TABLE,tablename)
  val admin = new HBaseAdmin(conf)
  if(!admin.isTableAvailable(tablename)){
    print("creating table:"+tablename+"\t")
    val tableDescription = new HTableDescriptor(tablename)
    tableDescription.addFamily(new HColumnDescriptor("cf".getBytes()));
    admin.createTable(tableDescription);
  } else {
    print("table already exists")
  }

  val table = new HTable(conf,tablename);
  for(x <- 1 to 10){
    var p = new Put(new String("row" + x).getBytes());
    p.add("cf".getBytes(),"column1".getBytes(),new String("value" + x).getBytes());
    table.put(p);
    print("Data Entered In Table")
  }
  val hBaseRDD = sc.newAPIHadoopRDD(conf, classOf[TableInputFormat],
      classOf[ImmutableBytesWritable],classOf[Result])
  print("RecordCount->>"+hBaseRDD.count())
  sc.stop()
```

Execution of Code

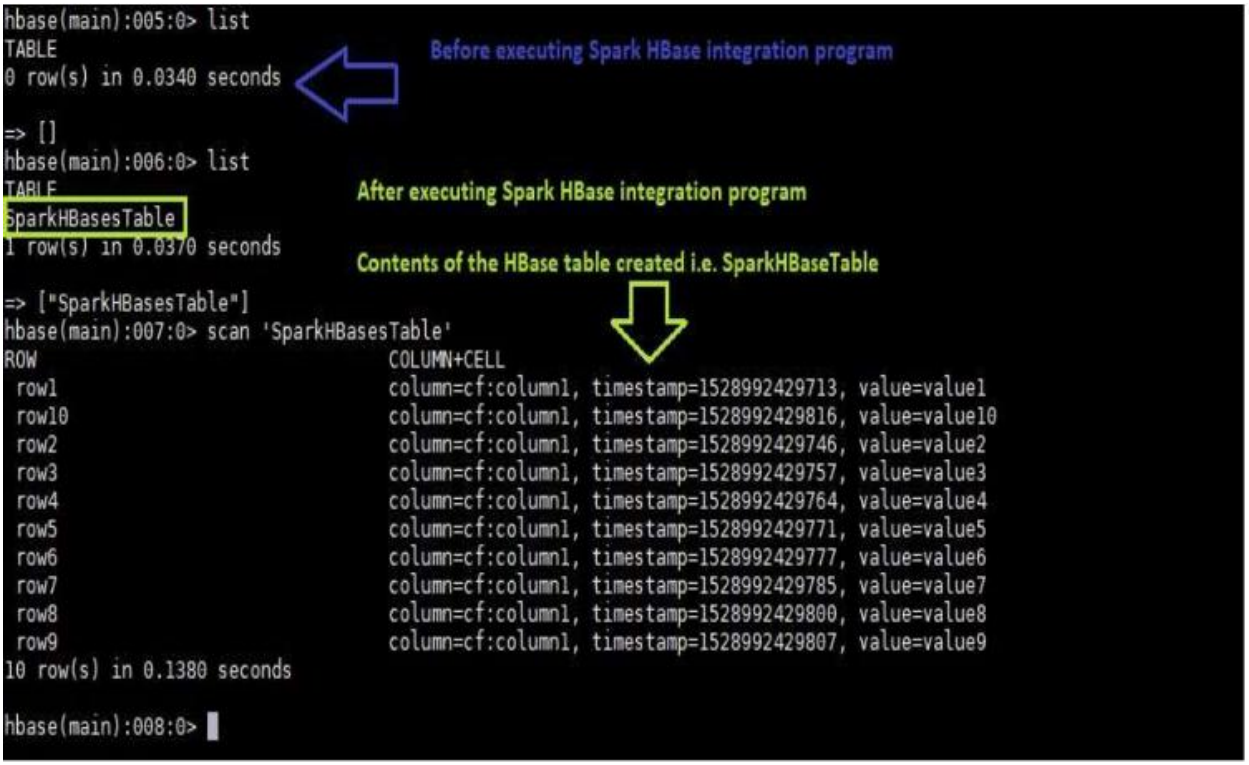


```
<terminated> SparkHBaseTest$ [Scala Application] /usr/java/jdk1.8.0_151/bin/java (Jun 14, 2018, 8:00:03 PM)
2018-06-14T20:00:21,987 INFO [main] org.spark_project.jetty.server.handler.ContextHandler - Started o.s.j.s.ServletContextHan
hello spark hbase ---> 1
2018-06-14T20:00:24,466 INFO [main] org.apache.hadoop.hbase.zookeeper.RecoverableZooKeeper - Process identifier=hconnection-0
2018-06-14T20:00:24,593 INFO [main] org.apache.zookeeper.ZooKeeper - Client environment:zookeeper.version=3.4.6-1569965, buil
2018-06-14T20:00:24,593 INFO [main] org.apache.zookeeper.ZooKeeper - Client environment:host.name=localhost
2018-06-14T20:00:24,606 INFO [main] org.apache.zookeeper.ZooKeeper - Client environment:java.version=1.8.0_151
2018-06-14T20:00:24,610 INFO [main] org.apache.zookeeper.ZooKeeper - Client environment:java.vendor=Oracle Corporation
2018-06-14T20:00:24,612 INFO [main] org.apache.zookeeper.ZooKeeper - Client environment:java.home=/usr/java/jdk1.8.0_151/jre
2018-06-14T20:00:24,614 INFO [main] org.apache.zookeeper.ZooKeeper - Client environment:java.class.path=/home/acadgild/.p2/pc
2018-06-14T20:00:24,688 INFO [main] org.apache.zookeeper.ZooKeeper - Client environment:java.library.path=/usr/java/packages/
2018-06-14T20:00:24,688 INFO [main] org.apache.zookeeper.ZooKeeper - Client environment:java.io.tmpdir=/tmp
2018-06-14T20:00:24,688 INFO [main] org.apache.zookeeper.ZooKeeper - Client environment:java.compiler=<NA>
2018-06-14T20:00:24,688 INFO [main] org.apache.zookeeper.ZooKeeper - Client environment:os.name=Linux
2018-06-14T20:00:24,689 INFO [main] org.apache.zookeeper.ZooKeeper - Client environment:os.arch=amd64
2018-06-14T20:00:24,689 INFO [main] org.apache.zookeeper.ZooKeeper - Client environment:os.version=2.6.32-696.18.7.el6.x86_64
2018-06-14T20:00:24,689 INFO [main] org.apache.zookeeper.ZooKeeper - Client environment:user.name=acadgild
2018-06-14T20:00:24,689 INFO [main] org.apache.zookeeper.ZooKeeper - Client environment:user.home=/home/acadgild
2018-06-14T20:00:24,689 INFO [main] org.apache.zookeeper.ZooKeeper - Client environment:user.dir=/home/acadgild/eclipse-works
2018-06-14T20:00:24,714 INFO [main] org.apache.zookeeper.ZooKeeper - Initiating client connection, connectString=localhost:21
2018-06-14T20:00:25,105 INFO [main-SendThread(localhost:2181)] org.apache.zookeeper.ClientCnxn - Opening socket connection to
2018-06-14T20:00:25,189 INFO [main-SendThread(localhost:2181)] org.apache.zookeeper.ClientCnxn - Socket connection establishe
2018-06-14T20:00:25,340 INFO [main-SendThread(localhost:2181)] org.apache.zookeeper.ClientCnxn - Session establishment comple
creating table:SparkHBasesTable 2018-06-14T20:00:33,144 INFO [main] org.apache.hadoop.hbase.client.HBaseAdmin - Created Spark
Data Entered In TableData Entered In TableData Entered In TableData Entered In TableData Entered In TableData Entered In Tabl
2018-06-14T20:00:35,694 INFO [main] org.apache.spark.storage.memory.MemoryStore - Block broadcast_0_piece0 stored as bytes in
2018-06-14T20:00:35,707 INFO [dispatcher-event-loop-0] org.apache.spark.storage.BlockManagerInfo - Added broadcast_0_piece0 i
2018-06-14T20:00:35,743 INFO [main] org.apache.spark.SparkContext - Created broadcast 0 from newAPIHadoopRDD at SparkHBaseTes
2018-06-14T20:00:35,878 INFO [main] org.apache.hadoop.hbase.zookeeper.RecoverableZooKeeper - Process identifier=hconnection-0
2018-06-14T20:00:35,878 INFO [main] org.apache.zookeeper.ZooKeeper - Initiating client connection, connectString=localhost:21
2018-06-14T20:00:36,017 INFO [main-SendThread(localhost:2181)] org.apache.zookeeper.ClientCnxn - Opening socket connection to
2018-06-14T20:00:36,021 INFO [main-SendThread(localhost:2181)] org.apache.zookeeper.ClientCnxn - Socket connection establishe
2018-06-14T20:00:36,090 INFO [main-SendThread(localhost:2181)] org.apache.zookeeper.ClientCnxn - Session establishment comple
2018-06-14T20:00:36,109 INFO [main] org.apache.hadoop.hbase.util.RegionSizeCalculator - Calculating region sizes for table "S
2018-06-14T20:00:37,829 INFO [main] org.apache.hadoop.hbase.client.ConnectionManager$HConnectionImplementation - Closing mast
2018-06-14T20:00:37,831 INFO [main] org.apache.hadoop.hbase.client.ConnectionManager$HConnectionImplementation - Closing zook
2018-06-14T20:00:37,837 INFO [main-EventThread] org.apache.zookeeper.ClientCnxn - EventThread shut down
2018-06-14T20:00:37,838 INFO [main] org.apache.zookeeper.ZooKeeper - Session: 0x163fcb08a3e0008 closed
```



```
<terminated> SparkHBaseTest$ [Scala Application] /usr/java/jdk1.8.0_151/bin/java (Jun 14, 2018, 8:00:03 PM)
2018-06-14T20:00:38,283 INFO [dag-scheduler-event-loop] org.apache.spark.scheduler.DAGScheduler - Missing parents: List()
2018-06-14T20:00:38,375 INFO [dag-scheduler-event-loop] org.apache.spark.scheduler.DAGScheduler - Submitting ResultStage 0 (N
2018-06-14T20:00:38,903 INFO [dag-scheduler-event-loop] org.apache.spark.storage.memory.MemoryStore - Block broadcast_1 store
2018-06-14T20:00:38,915 INFO [dag-scheduler-event-loop] org.apache.spark.storage.memory.MemoryStore - Block broadcast_1_piece
2018-06-14T20:00:38,917 INFO [dispatcher-event-loop-1] org.apache.spark.storage.BlockManagerInfo - Added broadcast_1_piece0 i
2018-06-14T20:00:38,923 INFO [dag-scheduler-event-loop] org.apache.spark.SparkContext - Created broadcast 1 from broadcast at
2018-06-14T20:00:39,043 INFO [dag-scheduler-event-loop] org.apache.spark.scheduler.DAGScheduler - Submitting 1 missing tasks
2018-06-14T20:00:39,070 INFO [dag-scheduler-event-loop] org.apache.spark.scheduler.TaskSchedulerImpl - Adding task set 0.0 wi
2018-06-14T20:00:39,508 INFO [dispatcher-event-loop-0] org.apache.spark.scheduler.TaskSetManager - Starting task 0.0 in stage
2018-06-14T20:00:39,626 INFO [Executor task launch worker for task 0] org.apache.spark.executor.Executor - Running task 0.0 i
2018-06-14T20:00:40,394 INFO [Executor task launch worker for task 0] org.apache.spark.rdd.NewHadoopRDD - Input split: HBase
2018-06-14T20:00:40,502 INFO [Executor task launch worker for task 0] org.apache.hadoop.hbase.zookeeper.RecoverableZooKeeper
2018-06-14T20:00:40,503 INFO [Executor task launch worker for task 0] org.apache.zookeeper.ZooKeeper - Initiating client conn
2018-06-14T20:00:40,528 INFO [Executor task launch worker for task 0-SendThread(localhost:2181)] org.apache.zookeeper.ClientC
2018-06-14T20:00:40,529 INFO [Executor task launch worker for task 0-SendThread(localhost:2181)] org.apache.zookeeper.ClientC
2018-06-14T20:00:40,534 INFO [Executor task launch worker for task 0-SendThread(localhost:2181)] org.apache.zookeeper.ClientC
2018-06-14T20:00:40,551 INFO [Executor task launch worker for task 0] org.apache.hadoop.hbase.mapreduce.TableInputFormatBase
2018-06-14T20:00:40,884 INFO [Executor task launch worker for task 0] org.apache.hadoop.hbase.client.ConnectionManager$HConne
2018-06-14T20:00:40,892 INFO [Executor task launch worker for task 0-EventThread] org.apache.zookeeper.ClientCnxn - EventThre
2018-06-14T20:00:40,893 INFO [Executor task launch worker for task 0] org.apache.zookeeper.ZooKeeper - Session: 0x163feb08a2e
2018-06-14T20:00:41,109 INFO [Executor task launch worker for task 0] org.apache.spark.executor.Executor - Finished task 0.0
2018-06-14T20:00:41,195 INFO [task-result-getter-0] org.apache.spark.scheduler.TaskSetManager - Finished task 0.0 in stage 0.
2018-06-14T20:00:41,240 INFO [task-result-getter-0] org.apache.spark.scheduler.TaskSchedulerImpl - Removed TaskSet 0.0, whose
2018-06-14T20:00:41,310 INFO [dag-scheduler-event-loop] org.apache.spark.scheduler.DAGScheduler - ResultStage 0 (count at Spa
2018-06-14T20:00:41,460 INFO [main] org.apache.spark.scheduler.DAGScheduler - Job 0 finished: count at SparkHBaseTest.scala:4
RecordCount->>102018-06-14T20:00:41,594 INFO [main] org.spark_project.jetty.server.AbstractConnector - Stopped Spark@351ff5c4
2018-06-14T20:00:41,618 INFO [main] org.apache.spark.ui.SparkUI - Stopped Spark web UI at http://10.0.3.15:4040
2018-06-14T20:00:41,903 INFO [dispatcher-event-loop-1] org.apache.spark.MapOutputTrackerMasterEndpoint - MapOutputTrackerMast
2018-06-14T20:00:41,994 INFO [main] org.apache.spark.storage.memory.MemoryStore - MemoryStore cleared
2018-06-14T20:00:41,999 INFO [main] org.apache.spark.storage.BlockManager - BlockManager stopped
2018-06-14T20:00:42,010 INFO [main] org.apache.spark.storage.BlockManagerMaster - BlockManagerMaster stopped
2018-06-14T20:00:42,047 INFO [dispatcher-event-loop-1] org.apache.spark.scheduler.OutputCommitCoordinator$OutputCommitCoordin
2018-06-14T20:00:42,059 INFO [main] org.apache.spark.SparkContext - Successfully stopped SparkContext
2018-06-14T20:00:42,107 INFO [Thread-2] org.apache.spark.util.ShutdownHookManager - Shutdown hook called
2018-06-14T20:00:42,130 INFO [Thread-2] org.apache.spark.util.ShutdownHookManager - Deleting directory /tmp/spark-55816480-1c
```

Hre before execution of the Code, we cannot see the table is created. However after execution of the code we could see table is created and records are inserted into it as below.



```
hbase(main):005:0> list
TABLE
0 row(s) in 0.0340 seconds

=> []
hbase(main):006:0> list
TABLE
SparkHBasesTable
1 row(s) in 0.0370 seconds

=> ["SparkHBasesTable"]
hbase(main):007:0> scan 'SparkHBasesTable'
ROW                                  COLUMN+CELL
 row1                                column=cf:column1, timestamp=1528992429713, value=value1
 row10                               column=cf:column1, timestamp=1528992429816, value=value10
 row2                                column=cf:column1, timestamp=1528992429746, value=value2
 row3                                column=cf:column1, timestamp=1528992429757, value=value3
 row4                                column=cf:column1, timestamp=1528992429764, value=value4
 row5                                column=cf:column1, timestamp=1528992429771, value=value5
 row6                                column=cf:column1, timestamp=1528992429777, value=value6
 row7                                column=cf:column1, timestamp=1528992429785, value=value7
 row8                                column=cf:column1, timestamp=1528992429800, value=value8
 row9                                column=cf:column1, timestamp=1528992429807, value=value9
10 row(s) in 0.1380 seconds

hbase(main):008:0> 
```

Before executing Spark HBase integration program

After executing Spark HBase integration program

Contents of the HBase table created i.e. SparkHBaseTable

Task 3
As discussed in class integrate Spark HBase

Pre-requisite

Start the zookeeper server in Kafka by navigating into $KAFKA_HOME with the command given below:

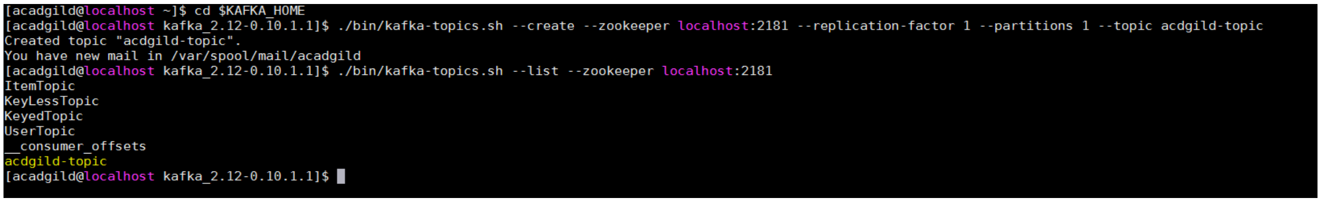
```
./bin/zookeeper-server-start.sh config/zookeeper.properties
```

Keep the terminal running, open one new terminal, and start the Kafka broker using the following command:

```
./bin/kafka-server-start.sh config/server.properties
```

After starting, leave both the terminals running, open a new terminal, and create a Kafka topic with the following command:

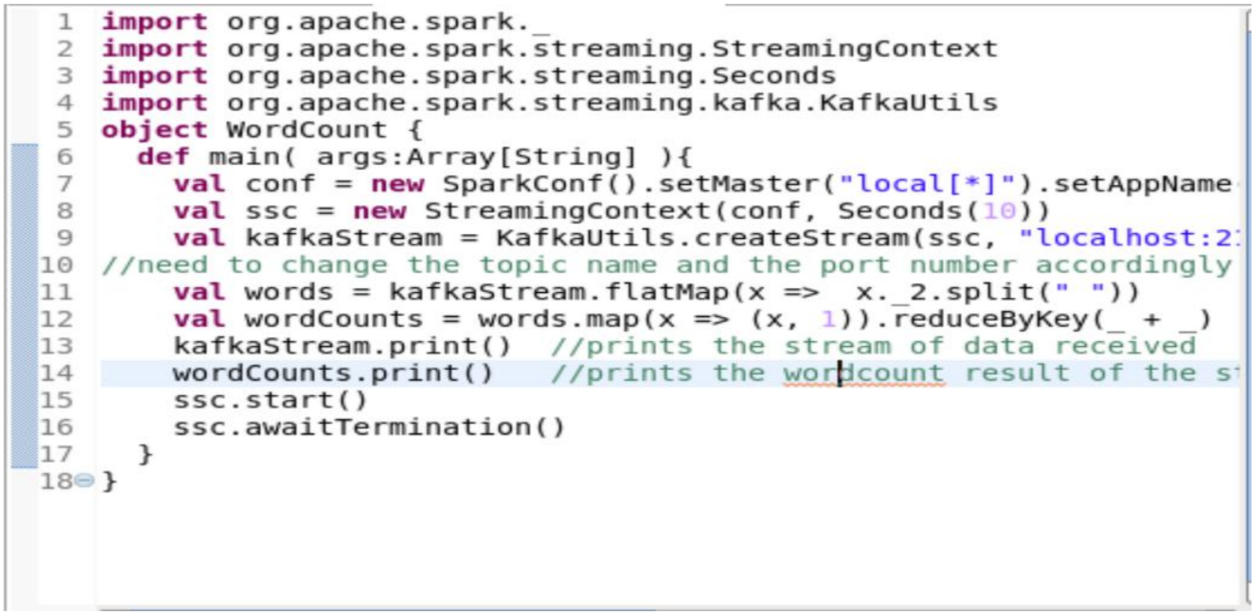
```
./bin/kafka-topics.sh --create --zookeeper localhost:2181 --replication-
factor 1 --partitions 1 --topic acdgild-topic
```



```
[acadgild@localhost ~]$ cd $KAFKA_HOME
[acadgild@localhost kafka_2.12-0.10.1.1]$ ./bin/kafka-topics.sh --create --zookeeper localhost:2181 --replication-factor 1 --partitions 1 --topic acdgild-topic
Created topic "acdgild-topic".
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost kafka_2.12-0.10.1.1]$ ./bin/kafka-topics.sh --list --zookeeper localhost:2181
ItemTopic
KeyLessTopic
KeyedTopic
UserTopic
__consumer_offsets
acdgild-topic
[acadgild@localhost kafka_2.12-0.10.1.1]$
```

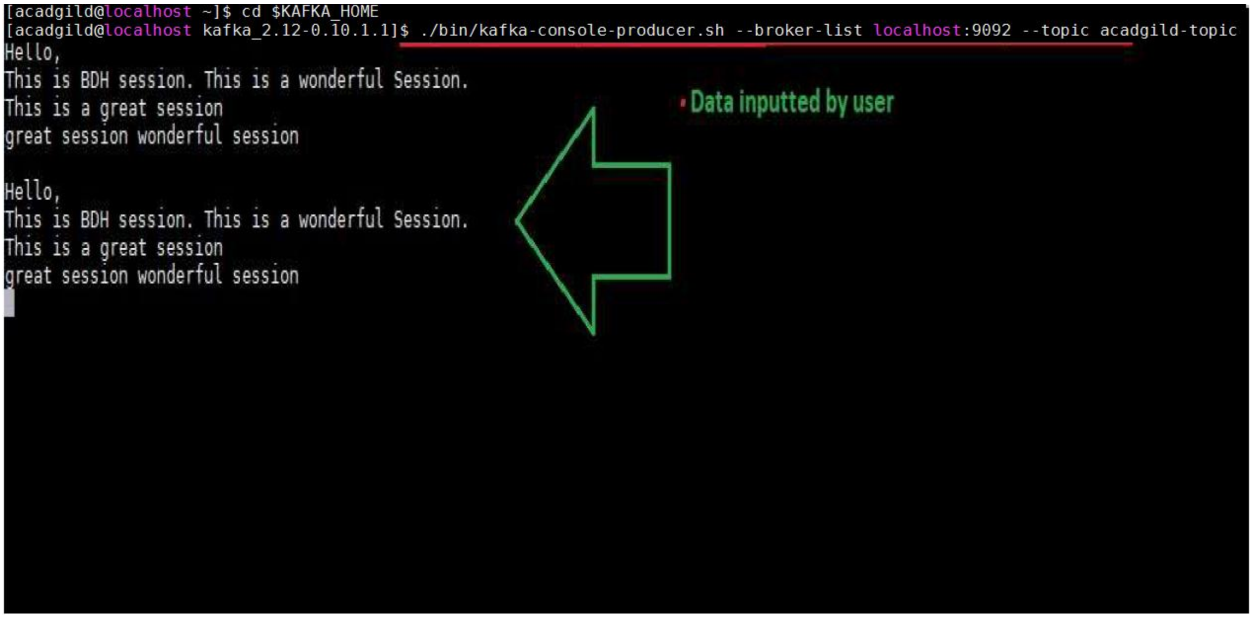
Program which runs the word count program by reading the contents from kafka and run in spark.

Code



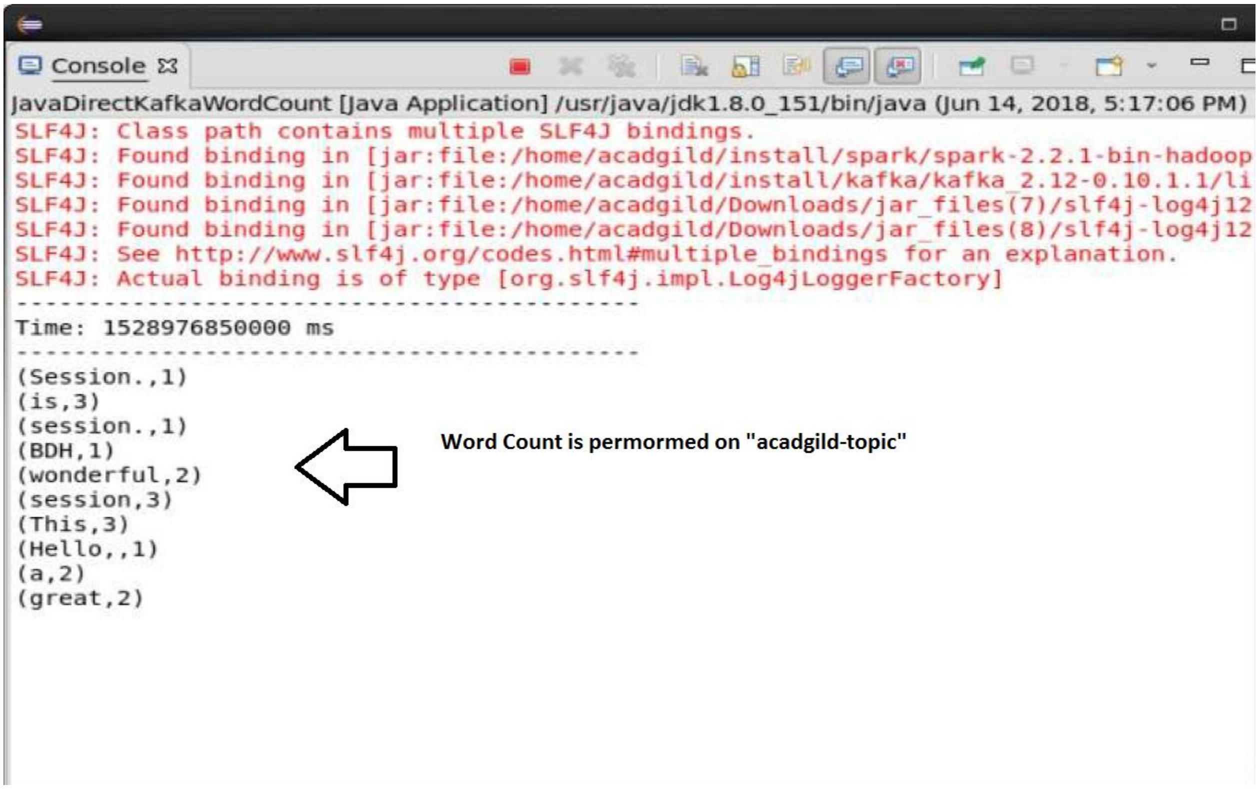
```
import org.apache.spark._
import org.apache.spark.streaming.StreamingContext
import org.apache.spark.streaming.Seconds
import org.apache.spark.streaming.kafka.KafkaUtils
object WordCount {
  def main( args:Array[String] ){
    val conf = new SparkConf().setMaster("local[*]").setAppName
    val ssc = new StreamingContext(conf, Seconds(10))
    val kafkaStream = KafkaUtils.createStream(ssc, "localhost:2
//need to change the topic name and the port number accordingly
    val words = kafkaStream.flatMap(x =>  x._2.split(" "))
    val wordCounts = words.map(x => (x, 1)).reduceByKey(_ + _)
    kafkaStream.print()  //prints the stream of data received
    wordCounts.print()   //prints the wordcount result of the s
    ssc.start()
    ssc.awaitTermination()
  }
}
```

Output



```
[acadgild@localhost ~]$ cd $KAFKA_HOME
[acadgild@localhost kafka_2.12-0.10.1.1]$ ./bin/kafka-console-producer.sh --broker-list localhost:9092 --topic acadgild-topic
Hello,
This is BDH session. This is a wonderful Session.
This is a great session
great session wonderful session

Hello,
This is BDH session. This is a wonderful Session.
This is a great session
great session wonderful session
```

• Data inputted by user



JavaDirectKafkaWordCount [Java Application] /usr/java/jdk1.8.0_151/bin/java (Jun 14, 2018, 5:17:06 PM)

```
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/spark/spark-2.2.1-bin-hadoop
SLF4J: Found binding in [jar:file:/home/acadgild/install/kafka/kafka_2.12-0.10.1.1/li
SLF4J: Found binding in [jar:file:/home/acadgild/Downloads/jar_files(7)/slf4j-log4j12
SLF4J: Found binding in [jar:file:/home/acadgild/Downloads/jar_files(8)/slf4j-log4j12
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
-------------------------------------------
Time: 1528976850000 ms
-------------------------------------------
(Session.,1)
(is,3)
(session.,1)
(BDH,1)
(wonderful,2)
(session,3)
(This,3)
(Hello,,1)
(a,2)
(great,2)
```

Word Count is permormed on "acadgild-topic"