ASSIGNMENT 25.1

Task 1 As discussed in class integrate Spark Hive

To perform hive integration with hive first we copy the **hive-site.xml** file form **hive/conf** folder into **spark/conf** folder.

Second, we provide the following property in **hive-site.xml** in **spark/conf** folder as shown below in the screenshot. We need to provide the thrift server address to property to **hive.metastore.uris.**

```
  <property>
    <name>hive.metastore.warehouse.dir</name>
    <value>hdfs://localhost:54310/user/hive/warehouse</value>
    <description>location of default database for the warehouse</description>
  </property>
  <property>
    <name>hive.metastore.uris</name>
    <value>thrift://localhost:9083<value/>
    <description>Thrift URI for the remote metastore. Used by metastore client to connect to r
  </property>
  <property>
    <name>hive.metastore.fastpath</name>
    <value>false</value>
    <description>Used to avoid all of the proxies and object copies in the metastore.  Note, i
(hive.metastore.uris must be empty) otherwise undefined and most likely undesired behavior wil
  </property>
  <property>
```

Third, now we have to start the all Hadoop daemons as shown below:

```
5200 NameNode
5328 DataNode
4802 org.eclipse.equinox.launcher_1.4.0.v20161219-1356.jar
5831 NodeManager
6327 RunJar
6457 RunJar
5707 ResourceManager
5547 SecondaryNameNode
6892 Jps
```

Start the hive service metastore

```
bhaskar@VirtualBox:~$ hive --service metastore
2018-06-08 18:24:17: Starting Hive Metastore Server
```

Start **hive** and check the databases present in warehouse

```
Logging initialized using configuration in jar:file:/usr/local/hive/lib/hive-common-2.3.3
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consi
ases.
hive> show databases;
OK
bank
default
pehla
project
Time taken: 4.681 seconds, Fetched: 4 row(s)
```

Above screen shot shows that there are three databases namely **bank,default,pehla and project.**

<u>Below screen, shot shows the spark application that list the databases of hive and will create a table **spark_new_table** in default database of hive.</u>

SparkHBaseTest.scala | SparkHiveTest.scala

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

object SparkHiveTest {

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

    val sparkSession = SparkSession.builder
                                    .master("local")
                                    .appName("spark session example")
                                    .config("spark.sql.warehouse.dir","/user/hive/warehouse")
                                    .config("hive.metastore.uris", "thrift://localhost:9083")
                                    .enableHiveSupport()
                                    .getOrCreate()

    val listOfDB = sparkSession.sqlContext.sql("show databases")
    listOfDB.show(8,false)
    println("test");

    val createSparkTableinhive = sparkSession
    .sqlContext
    .sql("CREATE TABLE IF NOT EXISTS spark_new_table(name STRING, age INT)" +
          "ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' +
          " LINES TERMINATED BY '\n'")

  }
}
```

Run the above application

```
2018-06-08 19:02:18 WARN  Utils:66 - Your hostname, VirtualBox resolves to a loopback address: 127.0.1.1; using 192.168.0.31 instead (on
2018-06-08 19:02:18 WARN  Utils:66 - Set SPARK_LOCAL_IP if you need to bind to another address
2018-06-08 19:02:19 INFO  SparkContext:54 - Running Spark version 2.3.0
2018-06-08 19:02:19 WARN  NativeCodeLoader:62 - Unable to load native-hadoop library for your platform... using builtin-java classes wher
2018-06-08 19:02:19 INFO  SparkContext:54 - Submitted application: spark session example
2018-06-08 19:02:19 INFO  SecurityManager:54 - Changing view acls to: bhaskar
2018-06-08 19:02:19 INFO  SecurityManager:54 - Changing modify acls to: bhaskar
2018-06-08 19:02:19 INFO  SecurityManager:54 - Changing view acls groups to:
2018-06-08 19:02:19 INFO  SecurityManager:54 - Changing modify acls groups to:
2018-06-08 19:02:19 INFO  SecurityManager:54 - SecurityManager: authentication disabled; ui acls disabled; users  with view permissions:
2018-06-08 19:02:20 INFO  Utils:54 - Successfully started service 'sparkDriver' on port 35837.
2018-06-08 19:02:20 INFO  SparkEnv:54 - Registering MapOutputTracker
2018-06-08 19:02:20 INFO  SparkEnv:54 - Registering BlockManagerMaster
2018-06-08 19:02:20 INFO  BlockManagerMasterEndpoint:54 - Using org.apache.spark.storage.DefaultTopologyMapper for getting topology infor
2018-06-08 19:02:20 INFO  BlockManagerMasterEndpoint:54 - BlockManagerMasterEndpoint up
2018-06-08 19:02:20 INFO  DiskBlockManager:54 - Created local directory at /tmp/blockmgr-c6767bef-6b05-438c-adf4-5e39558e6716
2018-06-08 19:02:21 INFO  MemoryStore:54 - MemoryStore started with capacity 884.7 MB
2018-06-08 19:02:21 INFO  SparkEnv:54 - Registering OutputCommitCoordinator
2018-06-08 19:02:21 INFO  log:192 - Logging initialized @4105ms
2018-06-08 19:02:21 INFO  Server:346 - jetty-9.3.z-SNAPSHOT
2018-06-08 19:02:21 INFO  Server:414 - Started @4460ms
2018-06-08 19:02:21 INFO  AbstractConnector:278 - Started ServerConnector@4bff64c2{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
2018-06-08 19:02:21 INFO  Utils:54 - Successfully started service 'SparkUI' on port 4040.
2018-06-08 19:02:21 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@2e61d218{/jobs,null,AVAILABLE,@Spark}
2018-06-08 19:02:21 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@aafcffa{/jobs/json,null,AVAILABLE,@Spark}
2018-06-08 19:02:21 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@6955cb39{/jobs/job,null,AVAILABLE,@Spark}
2018-06-08 19:02:21 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@2b5f4d54{/jobs/job/json,null,AVAILABLE,@Spark}
2018-06-08 19:02:21 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@5f7b97da{/stages,null,AVAILABLE,@Spark}
2018-06-08 19:02:21 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@18b0930f{/stages/json,null,AVAILABLE,@Spark}
2018-06-08 19:02:21 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@6b7906b3{/stages/stage,null,AVAILABLE,@Spark}
2018-06-08 19:02:21 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@256f8274{/stages/stage/json,null,AVAILABLE,@Spark}
2018-06-08 19:02:21 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@68044f4{/stages/pool,null,AVAILABLE,@Spark}
2018-06-08 19:02:21 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@52d239ba{/stages/pool/json,null,AVAILABLE,@Spark}
```

```
2018-06-08 19:02:22 INFO  BlockManagerMasterEndpoint:54 - Registering block manager 192.168.0.31:34637 with 884.7 MB RAM, BlockManagerId(driver,
2018-06-08 19:02:22 INFO  BlockManagerMaster:54 - Registered BlockManager BlockManagerId(driver, 192.168.0.31, 34637, None)
2018-06-08 19:02:22 INFO  BlockManager:54 - Initialized BlockManager: BlockManagerId(driver, 192.168.0.31, 34637, None)
2018-06-08 19:02:22 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@372ea2bc{/metrics/json,null,AVAILABLE,@Spark}
2018-06-08 19:02:22 INFO  SharedState:54 - Setting hive.metastore.warehouse.dir ('null') to the value of spark.sql.warehouse.dir ('/user/hive/war
2018-06-08 19:02:22 INFO  SharedState:54 - Warehouse path is '/user/hive/warehouse'.
2018-06-08 19:02:23 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@7e094740{/SQL,null,AVAILABLE,@Spark}
2018-06-08 19:02:23 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@7a11c4c7{/SQL/json,null,AVAILABLE,@Spark}
2018-06-08 19:02:23 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@3591009c{/SQL/execution,null,AVAILABLE,@Spark}
2018-06-08 19:02:23 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@5398edd0{/SQL/execution/json,null,AVAILABLE,@Spark}
2018-06-08 19:02:23 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@182f1e9a{/static/sql,null,AVAILABLE,@Spark}
2018-06-08 19:02:24 INFO  StateStoreCoordinatorRef:54 - Registered StateStoreCoordinator endpoint
2018-06-08 19:02:24 INFO  HiveUtils:54 - Initializing HiveMetastoreConnection version 1.2.1 using Spark classes.
2018-06-08 19:02:26 INFO  metastore:376 - Trying to connect to metastore with URI thrift://localhost:9083
2018-06-08 19:02:26 INFO  metastore:472 - Connected to metastore.
2018-06-08 19:02:26 INFO  SessionState:641 - Created local directory: /tmp/aa3d50d5-8050-4022-8027-ec40cec9b8ec_resources
2018-06-08 19:02:26 INFO  SessionState:641 - Created HDFS directory: /tmp/hive/bhaskar/aa3d50d5-8050-4022-8027-ec40cec9b8ec
2018-06-08 19:02:26 INFO  SessionState:641 - Created local directory: /tmp/bhaskar/aa3d50d5-8050-4022-8027-ec40cec9b8ec
2018-06-08 19:02:26 INFO  SessionState:641 - Created HDFS directory: /tmp/hive/bhaskar/aa3d50d5-8050-4022-8027-ec40cec9b8ec/_tmp_space.db
```

Output:

```
2018-06-08 19:02:26 INFO  HiveClientImpl:54 - Warehouse location for Hive client (version 1.2.2) i
2018-06-08 19:02:28 INFO  CodeGenerator:54 - Code generated in 254.131192 ms
2018-06-08 19:02:28 INFO  CodeGenerator:54 - Code generated in 13.744257 ms
+------------+
|databaseName|
+------------+
|bank        |
|default     |
|pehla       |
|project     |
+------------+

test
2018-06-08 19:02:29 INFO  SparkContext:54 - Invoking stop() from shutdown hook
2018-06-08 19:02:29 INFO  AbstractConnector:318 - Stopped Spark@4bff64c2{HTTP/1.1,[http/1.1]}{0.0.
2018-06-08 19:02:29 INFO  SparkUI:54 - Stopped Spark web UI at http://192.168.0.31:4040
```

Above screenshot shows the list of databases, which are present in hive.

Check the table created in hive

```
hive> show tables;
OK
spark_new_table
spark_table
Time taken: 0.043 seconds, Fetched: 2 row(s)
hive> describe spark_new_table;
OK
name                    string
age                     int
Time taken: 0.139 seconds, Fetched: 2 row(s)
hive>
```

Above screen, shot shows the table **spark_new_table** created by spark application in hive

Task2 As discussed in class integrate Spark Hbase.

Start the hbase daemons by **start-hbase.sh**

```
acadgild@localhost:~
[acadgild@localhost ~]$ start-hbase.sh
localhost: starting zookeeper, logging to /home/acadgild/install/hbase/hbase-1.2.6/logs/hbase-acadgild-zookeeper-localhost.localdomain.out
starting master, logging to /home/acadgild/install/hbase/hbase-1.2.6/logs/hbase-acadgild-master-localhost.localdomain.out
starting regionserver, logging to /home/acadgild/install/hbase/hbase-1.2.6/logs/hbase-acadgild-1-regionserver-localhost.localdomain.out
[acadgild@localhost ~]$ jps
```

```
[acadgild@localhost ~]$ jps
28865 SecondaryNameNode
28546 org.eclipse.equinox.launcher_1.4.0.v20161219-1356.jar
29122 NodeManager
28483 NameNode
3812 HQuorumPeer
3909 HMaster
29017 ResourceManager
4426 Jps
29852 RunJar
28684 DataNode
4029 HRegionServer
```

Start hbase shell

acadgild@localhost:~

```
[acadgild@localhost ~]$ hbase shell
2018-05-26 06:35:55,375 WARN  [main] util.NativeCodeLoader: Unable to load native-hadoop library for
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.ja
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 1.2.6, rUnknown, Mon May 29 02:25:32 CDT 2017

hbase(main):001:0> list
TABLE
song-artist-map
station-geo-map
subscribed-users
3 row(s) in 0.3950 seconds

=> ["song-artist-map", "station-geo-map", "subscribed-users"]
hbase(main):002:0>
```

Above screenshot shows the three tables present in hbase.

Below screen shot the spark application, that creates a table **SparkHbaseNewTable**

SparkHBaseTest.scala | SparkHiveTest.scala

```scala
import org.apache.spark.SparkContext

import org.apache.hadoop.hbase.HBaseConfiguration
import org.apache.hadoop.hbase.mapreduce.TableInputFormat
import org.apache.hadoop.hbase.client.HBaseAdmin
import org.apache.hadoop.hbase.{HTableDescriptor,HColumnDescriptor}
import org.apache.hadoop.hbase.util.Bytes
import org.apache.hadoop.hbase.client.{Put,HTable}
import org.apache.log4j._
import org.apache.hadoop.hbase.io.ImmutableBytesWritable
import org.apache.hadoop.hbase.client.Result

object SparkHBaseTest {

  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 = "SparkHBasesNewTable"
    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")
    }
```

```scala
    } 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()
  }
}
```

Run the above spark application

```
Console
<terminated> SparkHBaseTest$ (1) [Scala Application] /usr/java/jdk1.8.0_151/bin/java (May 26, 2018, 6:37:00 AM)
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/spark/spark-2.2.1-bin-hadoop2.7/jars/slf4j
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
18/05/26 06:37:01 INFO SparkContext: Running Spark version 2.2.1
18/05/26 06:37:02 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform...
18/05/26 06:37:02 WARN Utils: Your hostname, localhost.localdomain resolves to a loopback address:
18/05/26 06:37:02 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
18/05/26 06:37:02 INFO SparkContext: Submitted application: SparkHBaseTest
18/05/26 06:37:02 INFO SecurityManager: Changing view acls to: acadgild
18/05/26 06:37:02 INFO SecurityManager: Changing modify acls to: acadgild
18/05/26 06:37:02 INFO SecurityManager: Changing view acls groups to:
18/05/26 06:37:02 INFO SecurityManager: Changing modify acls groups to:
18/05/26 06:37:02 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled;
18/05/26 06:37:02 INFO Utils: Successfully started service 'sparkDriver' on port 38272.
18/05/26 06:37:03 INFO SparkEnv: Registering MapOutputTracker
18/05/26 06:37:03 INFO SparkEnv: Registering BlockManagerMaster
18/05/26 06:37:03 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMa
18/05/26 06:37:03 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
18/05/26 06:37:03 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-2ec6e200-c5a6-43f
18/05/26 06:37:03 INFO MemoryStore: MemoryStore started with capacity 886.8 MB
18/05/26 06:37:03 INFO SparkEnv: Registering OutputCommitCoordinator
18/05/26 06:37:03 INFO Utils: Successfully started service 'SparkUI' on port 4040.
18/05/26 06:37:03 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://192.168.0.43:4040
18/05/26 06:37:04 INFO Executor: Starting executor ID driver on host localhost
18/05/26 06:37:04 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlo
18/05/26 06:37:04 INFO NettyBlockTransferService: Server created on 192.168.0.43:40183
```

```
Console
<terminated> SparkHBaseTest$ (1) [Scala Application] /usr/java/jdk1.8.0_151/bin/java (May 26, 2018, 6:37:00 AM)
18/05/26 06:37:04 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replicat
18/05/26 06:37:04 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, 192.168.0.43, 40183, Nor
18/05/26 06:37:04 INFO BlockManagerMasterEndpoint: Registering block manager 192.168.0.43:40183 with 886.8 MB RAM,
18/05/26 06:37:04 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, 192.168.0.43, 40183, None
18/05/26 06:37:04 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, 192.168.0.43, 40183, None)
hello spark hbase ---> 1
18/05/26 06:37:04 INFO RecoverableZooKeeper: Process identifier=hconnection-0x1b9ea3e3 connecting to ZooKeeper ense
18/05/26 06:37:04 INFO ZooKeeper: Client environment:zookeeper.version=3.4.6-1569965, built on 02/20/2014 09:09 GMT
18/05/26 06:37:04 INFO ZooKeeper: Client environment:host.name=localhost
18/05/26 06:37:04 INFO ZooKeeper: Client environment:java.version=1.8.0_151
18/05/26 06:37:04 INFO ZooKeeper: Client environment:java.vendor=Oracle Corporation
18/05/26 06:37:04 INFO ZooKeeper: Client environment:java.home=/usr/java/jdk1.8.0_151/jre
18/05/26 06:37:04 INFO ZooKeeper: Client environment:java.class.path=/home/acadgild/.p2/pool/plugins/org.scala-ide.
18/05/26 06:37:04 INFO ZooKeeper: Client environment:java.library.path=/usr/java/packages/lib/amd64:/usr/lib64:/lib
18/05/26 06:37:04 INFO ZooKeeper: Client environment:java.io.tmpdir=/tmp
18/05/26 06:37:04 INFO ZooKeeper: Client environment:java.compiler=<NA>
18/05/26 06:37:04 INFO ZooKeeper: Client environment:os.name=Linux
18/05/26 06:37:04 INFO ZooKeeper: Client environment:os.arch=amd64
18/05/26 06:37:04 INFO ZooKeeper: Client environment:os.version=2.6.32-696.18.7.el6.x86_64
18/05/26 06:37:04 INFO ZooKeeper: Client environment:user.name=acadgild
18/05/26 06:37:04 INFO ZooKeeper: Client environment:user.home=/home/acadgild
18/05/26 06:37:04 INFO ZooKeeper: Client environment:user.dir=/home/acadgild/eclipse-workspace/Session_25
```

```
Console
<terminated> SparkHBaseTest$ (1) [Scala Application] /usr/java/jdk1.8.0_151/bin/java (May 26, 2018, 6:37:00 AM)
18/05/26 06:37:04 INFO ClientCnxn: Session establishment complete on server localhost/0:0:0:0:0:0:0:1:2181, session
creating table:SparkHBasesNewTable      18/05/26 06:37:07 INFO HBaseAdmin: Created SparkHBasesNewTable
Data Entered In TableData Entered In TableData Entered In TableData Entered In TableData Entered In TableData Ente
18/05/26 06:37:09 INFO MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 28.3 KB, fre
18/05/26 06:37:09 INFO BlockManagerInfo: Added broadcast_0_piece0 in memory on 192.168.0.43:40183 (size: 28.3 KB, f
18/05/26 06:37:09 INFO SparkContext: Created broadcast 0 from newAPIHadoopRDD at SparkHBaseTest.scala:42
18/05/26 06:37:09 INFO RecoverableZooKeeper: Process identifier=hconnection-0x665522c2 connecting to ZooKeeper ense
18/05/26 06:37:09 INFO ZooKeeper: Initiating client connection, connectString=localhost:2181 sessionTimeout=90000 w
18/05/26 06:37:09 INFO ClientCnxn: Opening socket connection to server localhost/0:0:0:0:0:0:0:1:2181. Will not att
18/05/26 06:37:09 INFO ClientCnxn: Socket connection established to localhost/0:0:0:0:0:0:0:1:2181, initiating sess
18/05/26 06:37:09 INFO ClientCnxn: Session establishment complete on server localhost/0:0:0:0:0:0:0:1:2181, session
18/05/26 06:37:09 INFO RegionSizeCalculator: Calculating region sizes for table "SparkHBasesNewTable".
18/05/26 06:37:09 INFO ConnectionManager$HConnectionImplementation: Closing master protocol: MasterService
18/05/26 06:37:09 INFO ConnectionManager$HConnectionImplementation: Closing zookeeper sessionid=0x16399f5e5b30009
18/05/26 06:37:09 INFO ClientCnxn: EventThread shut down
18/05/26 06:37:09 INFO ZooKeeper: Session: 0x16399f5e5b30009 closed
18/05/26 06:37:09 INFO SparkContext: Starting job: count at SparkHBaseTest.scala:43
18/05/26 06:37:09 INFO DAGScheduler: Got job 0 (count at SparkHBaseTest.scala:43) with 1 output partitions
18/05/26 06:37:09 INFO DAGScheduler: Final stage: ResultStage 0 (count at SparkHBaseTest.scala:43)
18/05/26 06:37:09 INFO DAGScheduler: Parents of final stage: List()
18/05/26 06:37:09 INFO DAGScheduler: Missing parents: List()
18/05/26 06:37:09 INFO DAGScheduler: Submitting ResultStage 0 (NewHadoopRDD[0] at newAPIHadoopRDD at SparkHBaseTest
18/05/26 06:37:09 INFO MemoryStore: Block broadcast_1 stored as values in memory (estimated size 2040.0 B, free 886
18/05/26 06:37:09 INFO MemoryStore: Block broadcast_1_piece0 stored as bytes in memory (estimated size 1278.0 B, fr
18/05/26 06:37:09 INFO BlockManagerInfo: Added broadcast_1_piece0 in memory on 192.168.0.43:40183 (size: 1278.0 B,
18/05/26 06:37:09 INFO SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:1006
```

```
18/05/26 06:37:10 INFO ClientCnxn: Session establishment complete on server localhost/127.0.0.1:2181, sessionid = 0
18/05/26 06:37:10 INFO TableInputFormatBase: Input split length: 0 bytes.
18/05/26 06:37:10 INFO ConnectionManager$HConnectionImplementation: Closing zookeeper sessionid=0x16399f5e5b3000a
18/05/26 06:37:10 INFO ZooKeeper: Session: 0x16399f5e5b3000a closed
18/05/26 06:37:10 INFO ClientCnxn: EventThread shut down
18/05/26 06:37:10 INFO Executor: Finished task 0.0 in stage 0.0 (TID 0). 875 bytes result sent to driver
18/05/26 06:37:10 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 412 ms on localhost (executor driv
18/05/26 06:37:10 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
18/05/26 06:37:10 INFO DAGScheduler: ResultStage 0 (count at SparkHBaseTest.scala:43) finished in 0.449 s
RecordCount->>1018/05/26 06:37:10 INFO DAGScheduler: Job 0 finished: count at SparkHBaseTest.scala:43, took 0.7217
18/05/26 06:37:10 INFO SparkUI: Stopped Spark web UI at http://192.168.0.43:4040
18/05/26 06:37:10 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/05/26 06:37:10 INFO MemoryStore: MemoryStore cleared
18/05/26 06:37:10 INFO BlockManager: BlockManager stopped
18/05/26 06:37:10 INFO BlockManagerMaster: BlockManagerMaster stopped
18/05/26 06:37:10 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
18/05/26 06:37:10 INFO SparkContext: Successfully stopped SparkContext
18/05/26 06:37:10 INFO ShutdownHookManager: Shutdown hook called
18/05/26 06:37:10 INFO ShutdownHookManager: Deleting directory /tmp/spark-e83d109d-83bc-43ff-956d-75d96ba49669
```

Now check the list of tables in hbase again

acadgild@localhost:~

```
hbase(main):002:0> list
TABLE
SparkHBasesNewTable
song-artist-map
station-geo-map
subscribed-users
4 row(s) in 0.0170 seconds

=> ["SparkHBasesNewTable", "song-artist-map", "station-geo-map", "subscribed-users"]
hbase(main):003:0>
```

Above screen shot shows the table SparkHbaseNewTable present in hbase

```
hbase(main):003:0> scan "SparkHBasesNewTable"
ROW                                    COLUMN+CELL
 row1                                  column=cf:column1, timestamp=1527296828160, value=value1
 row10                                 column=cf:column1, timestamp=1527296828229, value=value10
 row2                                  column=cf:column1, timestamp=1527296828188, value=value2
 row3                                  column=cf:column1, timestamp=1527296828193, value=value3
 row4                                  column=cf:column1, timestamp=1527296828197, value=value4
 row5                                  column=cf:column1, timestamp=1527296828202, value=value5
 row6                                  column=cf:column1, timestamp=1527296828207, value=value6
 row7                                  column=cf:column1, timestamp=1527296828213, value=value7
 row8                                  column=cf:column1, timestamp=1527296828217, value=value8
 row9                                  column=cf:column1, timestamp=1527296828225, value=value9
10 row(s) in 0.1810 seconds

hbase(main):004:0>
```

<u>Task 3</u> Spark Kafka Integeration

To integrate Kafka with spark First, we need to start the daemon.

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

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

```
acadgild@localhost:~/install/kafka/kafka_2.12-0.10.1.1
login as: acadgild
acadgild@192.168.0.43's password:
Last login: Sun May 27 01:55:20 2018 from 192.168.0.18
[acadgild@localhost ~]$
[acadgild@localhost ~]$ cd $KAFKA_HOME
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost kafka_2.12-0.10.1.1]$ ./bin/zookeeper-server-start.sh ./config/zookeeper.properties
[2018-05-27 07:54:20,338] INFO Reading configuration from: ./config/zookeeper.properties (org.apache.zookeeper.serve
[2018-05-27 07:54:20,341] INFO autopurge.snapRetainCount set to 3 (org.apache.zookeeper.server.DatadirCleanupManager
[2018-05-27 07:54:20,342] INFO autopurge.purgeInterval set to 0 (org.apache.zookeeper.server.DatadirCleanupManager)
[2018-05-27 07:54:20,342] INFO Purge task is not scheduled. (org.apache.zookeeper.server.DatadirCleanupManager)
[2018-05-27 07:54:20,342] WARN Either no config or no quorum defined in config, running  in standalone mode (org.ap
[2018-05-27 07:54:20,370] INFO Reading configuration from: ./config/zookeeper.properties (org.apache.zookeeper.serve
[2018-05-27 07:54:20,370] INFO Starting server (org.apache.zookeeper.server.ZooKeeperServerMain)
[2018-05-27 07:54:20,385] INFO Server environment:zookeeper.version=3.4.8--1, built on 02/06/2016 03:18 GMT (org.ap
[2018-05-27 07:54:20,385] INFO Server environment:host.name=localhost (org.apache.zookeeper.server.ZooKeeperServer)
[2018-05-27 07:54:20,385] INFO Server environment:java.version=1.8.0_151 (org.apache.zookeeper.server.ZooKeeperServe
[2018-05-27 07:54:20,385] INFO Server environment:java.vendor=Oracle Corporation (org.apache.zookeeper.server.ZooKe
[2018-05-27 07:54:20,385] INFO Server environment:java.home=/usr/java/jdk1.8.0_151/jre (org.apache.zookeeper.server.
[2018-05-27 07:54:20,385] INFO Server environment:java.class.path=:/home/acadgild/install/kafka/kafka_2.12-0.10.1.1/
```

```
r:/home/acadgild/install/kafka/kafka_2.12-0.10.1.1/bin/../libs/reflections-0.9.10.jar:/home/acadgild/install/kafka/kafka_
/home/acadgild/install/kafka/kafka_2.12-0.10.1.1/bin/../libs/scala-library-2.12.1.jar:/home/acadgild/install/kafka/kafka_
ors_2.12-1.0.4.jar:/home/acadgild/install/kafka/kafka_2.12-0.10.1.1/bin/../libs/slf4j-api-1.7.21.jar:/home/acadgild/insta
log4j12-1.7.21.jar:/home/acadgild/install/kafka/kafka_2.12-0.10.1.1/bin/../libs/snappy-java-1.1.2.6.jar:/home/acadgild/in
idation-api-1.1.0.Final.jar:/home/acadgild/install/kafka/kafka_2.12-0.10.1.1/bin/../libs/zkclient-0.9.jar:/home/acadgild/
ookeeper-3.4.8.jar (org.apache.zookeeper.server.ZooKeeperServer)
[2018-05-27 07:54:20,386] INFO Server environment:java.library.path=/usr/java/packages/lib/amd64:/usr/lib64:/lib64:/lib:/
erver)
[2018-05-27 07:54:20,386] INFO Server environment:java.io.tmpdir=/tmp (org.apache.zookeeper.server.ZooKeeperServer)
[2018-05-27 07:54:20,386] INFO Server environment:java.compiler=<NA> (org.apache.zookeeper.server.ZooKeeperServer)
[2018-05-27 07:54:20,386] INFO Server environment:os.name=Linux (org.apache.zookeeper.server.ZooKeeperServer)
[2018-05-27 07:54:20,386] INFO Server environment:os.arch=amd64 (org.apache.zookeeper.server.ZooKeeperServer)
[2018-05-27 07:54:20,386] INFO Server environment:os.version=2.6.32-696.18.7.el6.x86_64 (org.apache.zookeeper.server.ZooK
[2018-05-27 07:54:20,386] INFO Server environment:user.name=acadgild (org.apache.zookeeper.server.ZooKeeperServer)
[2018-05-27 07:54:20,386] INFO Server environment:user.home=/home/acadgild (org.apache.zookeeper.server.ZooKeeperServer)
[2018-05-27 07:54:20,386] INFO Server environment:user.dir=/home/acadgild/install/kafka/kafka_2.12-0.10.1.1 (org.apache.z
[2018-05-27 07:54:20,400] INFO tickTime set to 3000 (org.apache.zookeeper.server.ZooKeeperServer)
[2018-05-27 07:54:20,400] INFO minSessionTimeout set to -1 (org.apache.zookeeper.server.ZooKeeperServer)
[2018-05-27 07:54:20,400] INFO maxSessionTimeout set to -1 (org.apache.zookeeper.server.ZooKeeperServer)
[2018-05-27 07:54:20,415] INFO binding to port 0.0.0.0/0.0.0.0:2181 (org.apache.zookeeper.server.NIOServerCnxnFactory)
```

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

```
acadgild@localhost:~/install/kafka/kafka_2.12-0.10.1.1
[acadgild@localhost ~]$
[acadgild@localhost ~]$ cd $KAFKA_HOME
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost kafka_2.12-0.10.1.1]$ ./bin/kafka-server-start.sh ./config/server.properties
[2018-05-27 08:03:30,705] INFO KafkaConfig values:
        advertised.host.name = null
        advertised.listeners = null
        advertised.port = null
        authorizer.class.name =
        auto.create.topics.enable = true
        auto.leader.rebalance.enable = true
        background.threads = 10
        broker.id = 0
        broker.id.generation.enable = true
        broker.rack = null
        compression.type = producer
        connections.max.idle.ms = 600000
        controlled.shutdown.enable = true
        controlled.shutdown.max.retries = 3
```

```
        controller.socket.timeout.ms = 30000
        default.replication.factor = 1
        delete.topic.enable = false
        fetch.purgatory.purge.interval.requests = 1000
        group.max.session.timeout.ms = 300000
        group.min.session.timeout.ms = 6000
        host.name =
        inter.broker.protocol.version = 0.10.1-IV2
        leader.imbalance.check.interval.seconds = 300
        leader.imbalance.per.broker.percentage = 10
        listeners = null
        log.cleaner.backoff.ms = 15000
        log.cleaner.dedupe.buffer.size = 134217728
        log.cleaner.delete.retention.ms = 86400000
        log.cleaner.enable = true
        log.cleaner.io.buffer.load.factor = 0.9
        log.cleaner.io.buffer.size = 524288
        log.cleaner.io.max.bytes.per.second = 1.7976931348623157E308
        log.cleaner.min.cleanable.ratio = 0.5
        log.cleaner.min.compaction.lag.ms = 0
        log.cleaner.threads = 1
```

```
[2018-05-27 08:03:31,768] INFO [ExpirationReaper-0], Starting  (kafka.server.DelayedOperationPurgatory$ExpiredOperationReaper)
[2018-05-27 08:03:31,771] INFO [ExpirationReaper-0], Starting  (kafka.server.DelayedOperationPurgatory$ExpiredOperationReaper)
[2018-05-27 08:03:31,860] INFO Creating /controller (is it secure? false) (kafka.utils.ZKCheckedEphemeral)
[2018-05-27 08:03:31,877] INFO Result of znode creation is: OK (kafka.utils.ZKCheckedEphemeral)
[2018-05-27 08:03:31,878] INFO 0 successfully elected as leader (kafka.server.ZookeeperLeaderElector)
[2018-05-27 08:03:32,176] INFO New leader is 0 (kafka.server.ZookeeperLeaderElector$LeaderChangeListener)
[2018-05-27 08:03:32,183] INFO [ExpirationReaper-0], Starting  (kafka.server.DelayedOperationPurgatory$ExpiredOperationReaper)
[2018-05-27 08:03:32,192] INFO [ExpirationReaper-0], Starting  (kafka.server.DelayedOperationPurgatory$ExpiredOperationReaper)
[2018-05-27 08:03:32,201] INFO [ExpirationReaper-0], Starting  (kafka.server.DelayedOperationPurgatory$ExpiredOperationReaper)
[2018-05-27 08:03:32,221] INFO [GroupCoordinator 0]: Starting up. (kafka.coordinator.GroupCoordinator)
[2018-05-27 08:03:32,223] INFO [GroupCoordinator 0]: Startup complete. (kafka.coordinator.GroupCoordinator)
[2018-05-27 08:03:32,243] INFO [Group Metadata Manager on Broker 0]: Removed 0 expired offsets in 11 milliseconds. (kafka.coordinat
[2018-05-27 08:03:32,269] INFO Will not load MX4J, mx4j-tools.jar is not in the classpath (kafka.utils.Mx4jLoader$)
[2018-05-27 08:03:32,373] INFO Creating /brokers/ids/0 (is it secure? false) (kafka.utils.ZKCheckedEphemeral)
[2018-05-27 08:03:32,396] INFO Result of znode creation is: OK (kafka.utils.ZKCheckedEphemeral)
[2018-05-27 08:03:32,400] INFO Registered broker 0 at path /brokers/ids/0 with addresses: PLAINTEXT -> EndPoint(localhost,9092,PLAI
[2018-05-27 08:03:32,441] INFO Kafka version : 0.10.1.1 (org.apache.kafka.common.utils.AppInfoParser)
[2018-05-27 08:03:32,441] INFO Kafka commitId : f10ef2720b03b247 (org.apache.kafka.common.utils.AppInfoParser)
[2018-05-27 08:03:32,442] INFO [Kafka Server 0], started (kafka.server.KafkaServer)
[2018-05-27 08:03:32,931] INFO [ReplicaFetcherManager on broker 0] Removed fetcher for partitions KeyedTopic-0,TestTopic-0,TestTopi
ger)
[2018-05-27 08:03:33,003] INFO [ReplicaFetcherManager on broker 0] Removed fetcher for partitions KeyedTopic-0,TestTopic-0,TestTopi
ger)
```

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 --topic sample_topic --zookeeper localhost:2181 --partitions 1 --replication-factor 1

```
acadgild@localhost:~/install/kafka/kafka_2.12-0.10.1.1
[acadgild@localhost ~]$ cd $KAFKA_HOME
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost kafka_2.12-0.10.1.1]$ ./bin/kafka-topics.sh --create --topic sample_topic --zookeeper localhost:2181 --partitions 1 --replication-factor 1
WARNING: Due to limitations in metric names, topics with a period ('.') or underscore ('_') could collide. To avoid issues it is best to use either, but not bo
Created topic "sample_topic".
[acadgild@localhost kafka_2.12-0.10.1.1]$
```

After creating topic we will get a message **Created Topic "sample_topic"**

You can also check the topic list using the following command:

./bin/kafka-topics.sh --list --zookeeper localhost:2181

```
acadgild@localhost:~/install/kafka/kafka_2.12-0.10.1.1
[acadgild@localhost kafka_2.12-0.10.1.1]$ ./bin/kafka-topics.sh  --list  --zookeeper  localhost:2181
KeyedTopic
TestTopic
TestTopic1
sample_topic
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost kafka_2.12-0.10.1.1]$
```

Now in Spark, we will develop an application to consume the data that will do the word count for us. Our Spark application is as follows:

kafka_WordCount.scala | Spark_Kafka_Integration/pom.xml

```scala
package SparkIntegration

import org.apache.spark._
import org.apache.spark.streaming.StreamingContext
import org.apache.spark.streaming.Seconds
import org.apache.spark.streaming.kafka.KafkaUtils


object kafka_WordCount {

   def main( args:Array[String] ){

    val conf = new SparkConf().setMaster("local[*]").setAppName("KafkaReceiver")

    val ssc = new StreamingContext(conf, Seconds(10))

    val kafkaStream = KafkaUtils.createStream(ssc, "localhost:2181","spark-streaming-consumer-group", Map("sample_topic" -> 5))

    //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(_ + _)

    //prints the stream of data received
    kafkaStream.print()

    //prints the word count result of the stream
    wordCounts.print()

    ssc.start()
    ssc.awaitTermination()
  }

}
```

Dependencies for above file present in pom.xml file are as shown below

kafka_WordCount.scala | Spark_Kafka_Integration/pom.xml

```xml
<project xmlns="http://maven.apache.org/POM/4.0.0" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://maven.apache.org/POM/4.0.0 http://maven.apache.org/xsd/maven-4.0.0.xsd">
  <modelVersion>4.0.0</modelVersion>

  <groupId>Spark_Streaming</groupId>
  <artifactId>Spark_Kafka_Integration</artifactId>
  <version>0.0.1-SNAPSHOT</version>
  <packaging>jar</packaging>

  <name>Spark_Kafka_Integration</name>
  <url>http://maven.apache.org</url>

  <properties>
    <project.build.sourceEncoding>UTF-8</project.build.sourceEncoding>
  </properties>

  <dependencies>
  <dependency>
        <groupId>org.apache.spark</groupId>
        <artifactId>spark-streaming_2.11</artifactId>
        <version>1.6.3</version>
    </dependency>
    <!-- https://mvnrepository.com/artifact/org.apache.spark/spark-streaming-kafka_2.11 -->
<dependency>
    <groupId>org.apache.spark</groupId>
    <artifactId>spark-streaming-kafka_2.11</artifactId>
    <version>1.6.3</version>
</dependency>
    <dependency>
      <groupId>junit</groupId>
      <artifactId>junit</artifactId>
      <version>3.8.1</version>
      <scope>test</scope>
    </dependency>
```

Now for sending messages to this topic, you can use the console producer and send messages continuously. You can use the following commands to start the console producer.

./bin/kafka-console-producer.sh --broker-list localhost: 9092 -—topic sample_topic

```
>
>
>This is spark and kafka integration session
>kafka and spark is used for count the words
>This will show integration between spark and kafka
>
```

We are sending a message from the console producer and the Spark job will do the word count instantly and return the results as shown in the screenshot below:

```
18/06/19 17:59:40 INFO TaskSetManager: Finished task 0.0 in stage 10.0 (TID 1
18/06/19 17:59:40 INFO TaskSchedulerImpl: Removed TaskSet 10.0, whose tasks h
18/06/19 17:59:40 INFO DAGScheduler: ResultStage 10 (print at kafka_WordCount
-------------------------------------------
Time: 1529411380000 ms
-------------------------------------------
(integration,2)
(is,2)
(between,1)
(will,1)
(session,1)
(,3)
(This,2)
(kafka,3)
(spark,3)
(show,1)
...

18/06/19 17:59:40 INFO DAGScheduler: Job 6 finished: print at kafka_WordCount
18/06/19 17:59:40 INFO JobScheduler: Finished job streaming job 1529411380000
18/06/19 17:59:40 INFO JobScheduler: Total delay: 0.643 s for time 1529411380
18/06/19 17:59:40 INFO BlockRDD: Removing RDD 1 from persistence list
18/06/19 17:59:40 INFO KafkaInputDStream: Removing blocks of RDD BlockRDD[1]
```