

# Task 1 – Spark Hive Integration

Start MySQL and Hive metastore

```
acadgild@localhost:~  
File Edit View Search Terminal Help  
[acadgild@localhost ~]$ sudo service mysqld start  
[sudo] password for acadgild:  
Starting mysqld: [ OK ]  
[acadgild@localhost ~]$ cd install/  
.....  
*****/  
2018-08-19T11:06:40,928 INFO [main] org.apache.hadoop.hive.metastore.HiveMetaStore - Starting hive metastore on port 9083  
2018-08-19T11:06:42,353 INFO [main] org.apache.hadoop.hive.metastore.HiveMetaStore - 0: Opening raw store with implementation class:org.apache.hadoop.hive.metastore.ObjectStore  
2018-08-19T11:07:12,852 INFO [main] org.apache.hadoop.hive.metastore.HiveMetaStore - Added admin role in metastore  
2018-08-19T11:07:12,868 INFO [main] org.apache.hadoop.hive.metastore.HiveMetaStore - Added public role in metastore  
2018-08-19T11:07:13,032 INFO [main] org.apache.hadoop.hive.metastore.HiveMetaStore - No user is added in admin role, since config is empty  
2018-08-19T11:07:14,429 INFO [main] org.apache.hadoop.hive.metastore.HiveMetaStore - Starting DB backed MetaStore Server with SetUGI enabled  
2018-08-19T11:07:14,507 INFO [main] org.apache.hadoop.hive.metastore.HiveMetaStore - Started the new metaserver on port [9083]...  
2018-08-19T11:07:14,507 INFO [main] org.apache.hadoop.hive.metastore.HiveMetaStore - Options.minWorkerThreads = 200  
2018-08-19T11:07:14,507 INFO [main] org.apache.hadoop.hive.metastore.HiveMetaStore - Options.maxWorkerThreads = 1000  
2018-08-19T11:07:14,507 INFO [main] org.apache.hadoop.hive.metastore.HiveMetaStore - TCP keepalive = true  
2018-08-19T13:47:08,451 INFO [pool-7-thread-1] org.apache.hadoop.hive.metastore.HiveMetaStore - 1: source:127.0.0.1 get_all_databases
```

Example code

```
build.sbt x SparkHiveTest.scala x SparkHbaseTest.scala x  
import org.apache.spark.sql.SparkSession  
  
object SparkHiveTest {  
  def main (args: Array[String]) : Unit = {  
    val sparkSession = SparkSession.builder.master( master = "local")  
      .appName( name = "spark session example").config("spark.sql.warehouse.dir", "/")  
    val listofDB = sparkSession.sql( sqlText = "show databases")  
    listofDB.show( numRows = 8, truncate = false)  
    println("test");  
  }  
}
```

## Output

```
18/08/19 13:48:28 INFO TaskSchedulerImpl: Removed taskSet 0.0, whose tasks have all completed, from pool
18/08/19 13:48:28 INFO DAGScheduler: ResultStage 0 (show at SparkHiveTest.scala:11) finished in 3.914 s
18/08/19 13:48:28 INFO DAGScheduler: Job 0 finished: show at SparkHiveTest.scala:11, took 8.171599 s
18/08/19 13:48:29 INFO CodeGenerator: Code generated in 109.806302 ms
+-----+
|databaseName|
+-----+
|default      |
+-----+

test
18/08/19 13:48:29 INFO SparkContext: Invoking stop() from shutdown hook
18/08/19 13:48:29 INFO SparkUI: Stopped Spark web UI at http://10.0.2.15:4040
18/08/19 13:48:29 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/08/19 13:48:29 INFO MemoryStore: MemoryStore cleared
18/08/19 13:48:29 INFO BlockManager: BlockManager stopped
18/08/19 13:48:29 INFO BlockManagerMaster: BlockManagerMaster stopped
18/08/19 13:48:29 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
Compilation completed successfully in 1 m 46 s 579 ms (12 minutes ago) 83:78 LF: UTF-8
```

## Task 2 – Spark Hbase Integration

Start all services including Hbase services

```
[acadgild@localhost ~]$ jps
6337 HMaster
6423 HRegionServer
6632 Jps
5416 SecondaryNameNode
5624 ResourceManager
5160 NameNode
5256 DataNode
5724 NodeManager
2956 Main
3085 RemoteMavenServer
[acadgild@localhost ~]$
```

Example code:

```

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

  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( original = "row" + x).getBytes());
    p.add("cf".getBytes(), "column1".getBytes(), new String( original = "value" + x).getBytes());
    table.put(p);
    print("Data Entered In Table")
  }
  val hBaseRDD = sc.newAPIHadoopRDD(conf, classOf[TableInputFormat], classOf[ImmutableBytesWritable], classOf[TableOutputFormat])
  print("RecordCount->"+hBaseRDD.count())
  sc.stop()
}
}

```

Output:

```

18/08/19 20:36:19 INFO ClientCnxn: Opening socket connection to server localhost/127.0.0.1:2181. Will not attempt to authenticate using
18/08/19 20:36:19 INFO ClientCnxn: Socket connection established to localhost/127.0.0.1:2181, initiating session
18/08/19 20:36:19 INFO ClientCnxn: Session establishment complete on server localhost/127.0.0.1:2181, sessionId = 0x16552b76fce0006, neg
creating table:SparkHBasesTable 18/08/19 20:36:41 INFO HBaseAdmin: Created SparkHBasesTable
Data Entered In TableData Entered In TableData Entered In TableData Entered In TableData Entered In TableData Entered In TableData Enter
18/08/19 20:36:46 INFO MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 19.3 KB, free 111.0 MB)
18/08/19 20:36:46 INFO BlockManagerInfo: Added broadcast_0_piece0 in memory on 10.0.2.15:44056 (size: 19.3 KB, free: 111.2 MB)
18/08/19 20:36:46 INFO SparkContext: Created broadcast 0 from newAPIHadoopRDD at SparkHbaseTest.scala:42
18/08/19 20:36:46 INFO RecoverableZooKeeper: Process identifier=hconnection-0x1d6bc91c connecting to ZooKeeper ensemble=localhost:2181
18/08/19 20:36:46 INFO ZooKeeper: Initiating client connection, connectString=localhost:2181 sessionTimeout=90000 watcher=hconnection-0x
18/08/19 20:36:46 INFO ClientCnxn: Opening socket connection to server localhost/0:0:0:0:0:0:1:2181. Will not attempt to authenticate
18/08/19 20:36:46 INFO ClientCnxn: Socket connection established to localhost/0:0:0:0:0:0:1:2181, initiating session
18/08/19 20:36:46 INFO ClientCnxn: Session establishment complete on server localhost/0:0:0:0:0:0:1:2181, sessionId = 0x16552b76fce0006
18/08/19 20:36:52 INFO Executor: Finished task 0.0 in stage 0.0 (TID 0). 1035 bytes result sent to driver
18/08/19 20:36:52 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 2575 ms on localhost (1/1)
18/08/19 20:36:52 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
18/08/19 20:36:52 INFO DAGScheduler: ResultStage 0 (count at SparkHbaseTest.scala:43) finished in 3.000 s
18/08/19 20:36:52 INFO DAGScheduler: Job 0 finished: count at SparkHbaseTest.scala:43, took 4.717643 s
RecordCount->1018/08/19 20:36:53 INFO SparkUI: Stopped Spark web UI at http://10.0.2.15:4040
18/08/19 20:36:53 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/08/19 20:36:53 INFO MemoryStore: MemoryStore cleared
18/08/19 20:36:53 INFO BlockManager: BlockManager stopped
18/08/19 20:36:53 INFO BlockManagerMaster: BlockManagerMaster stopped
18/08/19 20:36:53 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
18/08/19 20:36:53 INFO SparkContext: Successfully stopped SparkContext
18/08/19 20:36:53 INFO ShutdownHookManager: Shutdown hook called
18/08/19 20:36:53 INFO ShutdownHookManager: Deleting directory /tmp/spark-0430a337-ecdd-4bd3-8ee3-0d064bc02b00

```



```
[acadgild@localhost ~]$ hbase shell
2018-08-19 21:44:23,827 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using b
uilt-in java classes where applicable
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.jar!/org/slf4j/impl/Static
LoggerBinder.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.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
SparkHBasesTable
1 row(s) in 2.1340 seconds

=> ["SparkHBasesTable"]
hbase(main):002:0> scan 'SparkHBasesTable'
ROW COLUMN+CELL
row1 column=cf:column1, timestamp=1534691202638, value=value1
row10 column=cf:column1, timestamp=1534691202798, value=value10
row2 column=cf:column1, timestamp=1534691202713, value=value2
row3 column=cf:column1, timestamp=1534691202724, value=value3
row4 column=cf:column1, timestamp=1534691202737, value=value4
row5 column=cf:column1, timestamp=1534691202743, value=value5
row6 column=cf:column1, timestamp=1534691202751, value=value6
row7 column=cf:column1, timestamp=1534691202762, value=value7
row8 column=cf:column1, timestamp=1534691202773, value=value8
row9 column=cf:column1, timestamp=1534691202782, value=value9
10 row(s) in 1.7790 seconds


hbase(main):003:0> █
```

## Task 3 – Spark Kafka Integration

Start zookeeper

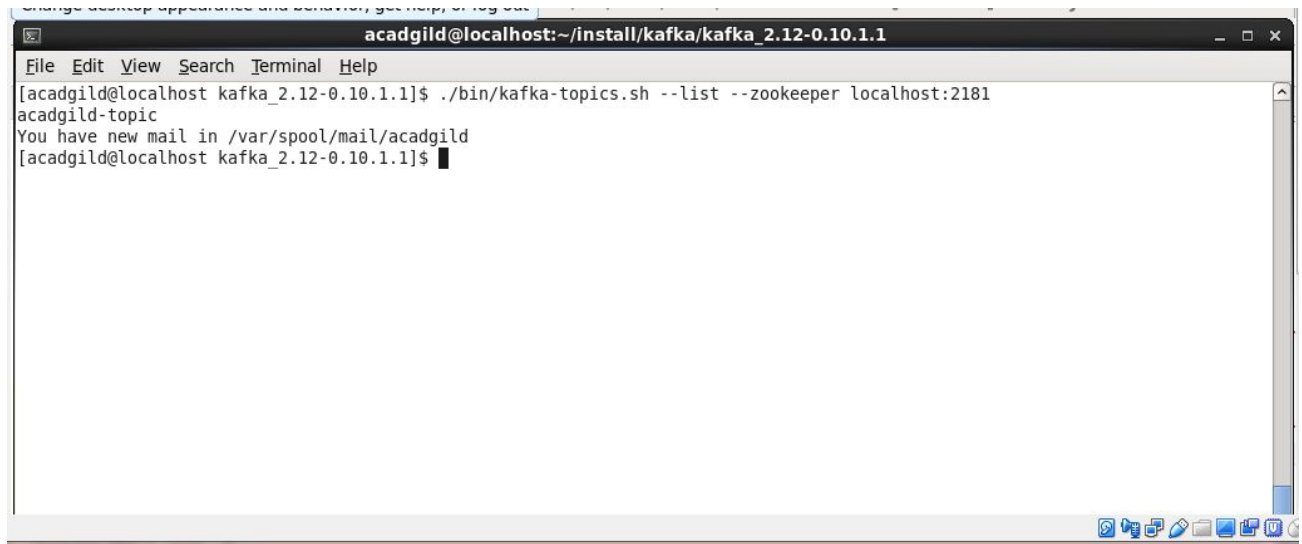
```
acadgild@localhost: ~/install/kafka/kafka_2.12-0.10.1.1
File Edit View Search Terminal Help
[acadgild@localhost ~]$ cd $kafka_home
[acadgild@localhost ~]$ cd $KAFKA_HOME
[acadgild@localhost kafka_2.12-0.10.1.1]$ ./bin/zookeeper-se
zookeeper-security-migration.sh zookeeper-server-stop.sh
zookeeper-server-start.sh
[acadgild@localhost kafka_2.12-0.10.1.1]$ ./bin/zookeeper-server-start.sh config/zookeeper.properties
[2018-08-21 22:09:48,382] INFO Reading configuration from: config/zookeeper.properties (org.apache.zookeeper.server.quor
um.QuorumPeerConfig)
[2018-08-21 22:09:48,473] INFO autopurge.snapRetainCount set to 3 (org.apache.zookeeper.server.DataDirCleanupManager)
[2018-08-21 22:09:48,473] INFO autopurge.purgeInterval set to 0 (org.apache.zookeeper.server.DataDirCleanupManager)
[2018-08-21 22:09:48,473] INFO Purge task is not scheduled. (org.apache.zookeeper.server.DataDirCleanupManager)
[2018-08-21 22:09:48,474] WARN Either no config or no quorum defined in config, running in standalone mode (org.apache.
zookeeper.server.quorum.QuorumPeerMain)
[2018-08-21 22:09:49,071] INFO Reading configuration from: config/zookeeper.properties (org.apache.zookeeper.server.quor
um.QuorumPeerConfig)
[2018-08-21 22:09:49,079] INFO Starting server (org.apache.zookeeper.server.ZooKeeperServerMain)
[2018-08-21 22:09:49,353] INFO Server environment:zookeeper.version=3.4.8--1, built on 02/06/2016 03:18 GMT (org.apache.
zookeeper.server.ZooKeeperServer)
[2018-08-21 22:09:49,353] INFO Server environment:host.name=localhost (org.apache.zookeeper.server.ZooKeeperServer)
[2018-08-21 22:09:49,354] INFO Server environment:java.version=1.8.0_151 (org.apache.zookeeper.server.ZooKeeperServer)
[2018-08-21 22:09:49,354] INFO Server environment:java.vendor=Oracle Corporation (org.apache.zookeeper.server.ZooKeeperS
erver)
[2018-08-21 22:09:49,354] INFO Server environment:java.home=/usr/java/jdk1.8.0_151/jre (org.apache.zookeeper.server.ZooK
eeperServer)
```

## Start kafka broker



```
acadgild@localhost:~/install/kafka/kafka_2.12-0.10.1.1
File Edit View Search Terminal Help
[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
[2018-08-21 22:12:46,934] 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
    controlled.shutdown.retry.backoff.ms = 5000
    controller.socket.timeout.ms = 30000
    default.replication.factor = 1
    delete.topic.enable = false
    fetch.purgatory.purge.interval.requests = 1000
```

## Create a topic with name acadgild-topic



```
acadgild@localhost:~/install/kafka/kafka_2.12-0.10.1.1
File Edit View Search Terminal Help
[acadgild@localhost kafka_2.12-0.10.1.1]$ ./bin/kafka-topics.sh --list --zookeeper localhost:2181
acadgild-topic
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost kafka_2.12-0.10.1.1]$
```

Code :

```

4 object KafkaAndSparkStreaming {
5
6   def main(args: Array[String]) : Unit {
7     Logger.getRootLogger.setLevel(Level.WARN)
8     val topics = Set("acadgild-topic")
9     val kafkaParams = Map[String, String](elems = "metadata.broker.list" -> "localhost:9092")
10
11     val sparkConf = new SparkConf().setAppName("KafkaAndSparkStreaming").setMaster("local[*]")
12     val ssc = new StreamingContext(sparkConf, Seconds(30))
13     ssc.checkpoint(directory = "checkpoint")
14     val rawStream = KafkaUtils.createDirectStream[String, String, StringDecoder, StringDecoder](ssc, kafkaParams)
15     rawStream.print()
16     val words = rawStream.map(_._2).flatMap(x => x.split(regex = " "))
17     val wc = words.map(x => (x,1)).reduceByKey(_+_ )
18     //wc.print()
19
20     rawStream.foreachRDD(rdd =>
21       | val sqlContext = SQLContext.getOrCreate(rdd.sparkContext)
22       import sqlContext.implicits._
23       val wordsDF = rdd.toDF(colNames = "dummy","eachword")
24       wordsDF.registerTempTable(tableName = "words")
25
26       val wcDF = sqlContext.sql(sqlText = "select eachword, count(*) as total from words group by eachword")
27       wcDF.show()
28     )
29     ssc.start()
30     ssc.awaitTermination()
31   }

```

## Input

```

acadgild@localhost:~/install/kafka/kafka_2.12-0.10.1.1
File Edit View Search Terminal Help
[acadgild@localhost kafka_2.12-0.10.1.1]$ ./bin/kafka-topics.sh --list --zookeeper localhost:2181
acadgild-topic
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost kafka_2.12-0.10.1.1]$ ./bin/kafka-console-producer.sh --broker-list localhost:9092 --topic acadgild-topic
kafka
testing kafka spark integration
spark
^CYou have new mail in /var/spool/mail/acadgild
[acadgild@localhost kafka_2.12-0.10.1.1]$ ./bin/kafka-console-producer.sh --broker-list localhost:9092 --topic acadgild-topic
testing kafka spark integration
kafka
spark
kafka
spark
spark
kafka
testing
testing kafka spark integration
kafka
spark

```

## Output

```

18/08/21 23:46:26 INFO JobScheduler: Finished job streaming job 1534875360000 ms.0 from job set of time 1534875360000 ms
18/08/21 23:46:26 INFO JobScheduler: Starting job streaming job 1534875360000 ms.1 from job set of time 1534875360000 ms
-----
Time: 1534875360000 ms
-----
(null,spark)
(null,kafka)
(null,testing)

18/08/21 23:46:26 INFO ContextCleaner: Cleaned accumulator 14
18/08/21 23:46:26 INFO BlockManagerInfo: Removed broadcast_3_piece0 on localhost:45352 in memory (size: 1265.0 B, free: 139.0 MB)
18/08/21 23:46:28 INFO SparkContext: Starting job: show at KafkaAndSparkStreaming.scala:47

```

```

18/08/21 23:46:55 INFO JobScheduler: Finished job streaming job 1534875360000 ms.1 from job set of time 1534875360000 ms
18/08/21 23:46:55 INFO JobScheduler: Total delay: 55.187 s for time 1534875360000 ms (execution: 29.091 s)
+-----+-----+
|eachword|total|
+-----+-----+
|   kafka|    1|
| testing|    1|
|   spark|    1|
+-----+-----+

18/08/21 23:46:55 INFO KafkaRDD: Removing RDD 0 from persistence list
18/08/21 23:46:55 INFO JobScheduler: Starting job streaming job 1534875390000 ms.0 from job set of time 1534875390000 ms
18/08/21 23:46:55 INFO BlockManager: Removing RDD 0

```

```

18/08/21 23:46:55 INFO DAGScheduler: Job 5 finished: print at KafkaAndSparkStreaming.scala:32, took 0.061488 s
18/08/21 23:46:55 INFO JobScheduler: Finished job streaming job 1534875390000 ms.0 from job set of time 1534875390000 ms
18/08/21 23:46:55 INFO JobScheduler: Starting job streaming job 1534875390000 ms.1 from job set of time 1534875390000 ms
-----
Time: 1534875390000 ms
-----
(null,testing kafka spark integration)

18/08/21 23:46:55 INFO CheckpointWriter: Checkpoint for time 1534875360000 ms saved to file 'file:/home/acadgild/IdeaProjects/untitled1,
18/08/21 23:46:55 INFO DStreamGraph: Clearing checkpoint data for time 1534875360000 ms
18/08/21 23:46:55 INFO DStreamGraph: Cleared checkpoint data for time 1534875360000 ms

```

```

18/08/21 23:47:00 INFO KafkaRDD: Computing topic acadgild-topic, partition 0 offsets 48 -> 50
18/08/21 23:47:00 INFO VerifiableProperties: Verifying properties
18/08/21 23:47:00 INFO VerifiableProperties: Property group.id is overridden to
18/08/21 23:47:00 INFO VerifiableProperties: Property zookeeper.connect is overridden to
+-----+-----+
|           eachword|total|
+-----+-----+
|testing kafka spa...|    1|
+-----+-----+

-----
Time: 1534875420000 ms
-----
(null,kafka)
(null,spark)

18/08/21 23:47:01 INFO BlockManagerInfo: Removed broadcast_6_piece0 on localhost:45352 in memory (size: 6.1 KB, free: 139.0 MB)

```

```

18/08/21 23:47:05 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/08/21 23:47:05 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/08/21 23:47:05 INFO Executor: Finished task 176.0 in stage 18.0 (TID 784). 1652 bytes result sent to driver
+-----+-----+
|eachword|total|
+-----+-----+
|   kafka|    1|
|   spark|    1|
+-----+-----+

18/08/21 23:47:05 INFO TaskSetManager: Starting task 177.0 in stage 18.0 (TID 785, localhost, partition 178,NODE_LOCAL, 1918 bytes)
18/08/21 23:47:05 INFO TaskSetManager: Finished task 176.0 in stage 18.0 (TID 784) in 19 ms on localhost (177/199)
18/08/21 23:47:05 INFO Executor: Running task 177.0 in stage 18.0 (TID 785)
18/08/21 23:47:05 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks

```