Session 25 – Big Data System Integration

Assignment 1

Task 1: Integrate Spark with Hive

On following the steps specified in the document:

Step 1:

```
[acadgild@localhost ~]$ cp /home/acadgild/install/hive/apache-hive-2.3.2-bin/conf/hive-site.xml /home/acadgild/install/spark/spark-2.2.1-bin-hadoop2.7/conf/
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ ls /home/acadgild/install/spark/spark-2.2.1-bin-hadoop2.7/conf/
docker.properties.template | hive-site.xml 2 | metrics.properties.template | spark-defaults.conf.template | fairscheduler.xml.template | log4].properties.template | slaves.template | spark-env.sh.template | [acadgild@localhost ~]$ gedit /home/acadgild/install/spark/spark-2.2.1-bin-hadoop2.7/conf/hive-site.xml 3 | [acadgild@localhost ~]$ | Spark HIVE INTEGRATION tel Compatibility Model. |
```

In reference with the above screenshot,

- 1: Copying hive-site.xml from hive home to spark home
- 2: hive-site.xml is in the spark home folder.
- 3: editing the hive-site xml to add the property related to hive metastore.

Step 2:

```
hive-site.xml 💥
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
property>
  <name>javax.jdo.option.ConnectionURL</name>
  <value>jdbc:mysql://localhost/metastore?createDatabaseIfNotExist=true</value>
<value>jdbc:mysql://localhost/metastore?useSSL=false</value>
property>
  <name>javax.jdo.option.ConnectionDriverName</name>
  <value>com.mysql.jdbc.Driver</value>
</property>
cproperty>
  <name>javax.jdo.option.ConnectionUserName</name>
  <value>root</value>
</property>
property>
  <name>javax.jdo.option.ConnectionPassword
  <value>Root@123</value>
</property>
  <name>datanucleus.autoCreateSchema</name>
  <value>true</value>
 /propertv>
```

Editing the hive-stie.xml and adding the below property.

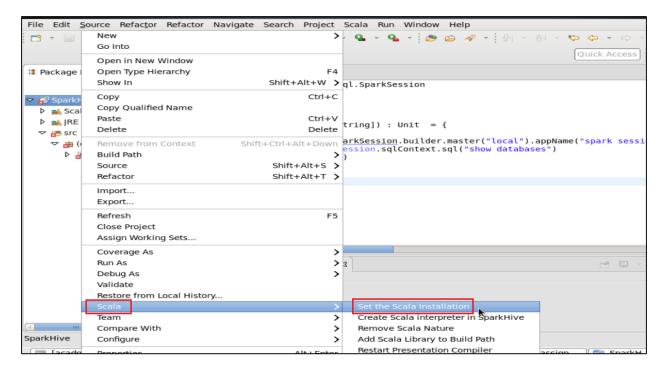
```
property>
```

```
<name>hive.metastore.uris</name>
  <value>thrift://localhost:9083</value>
  <description>password for connecting to mysql server</description>
</property>
```

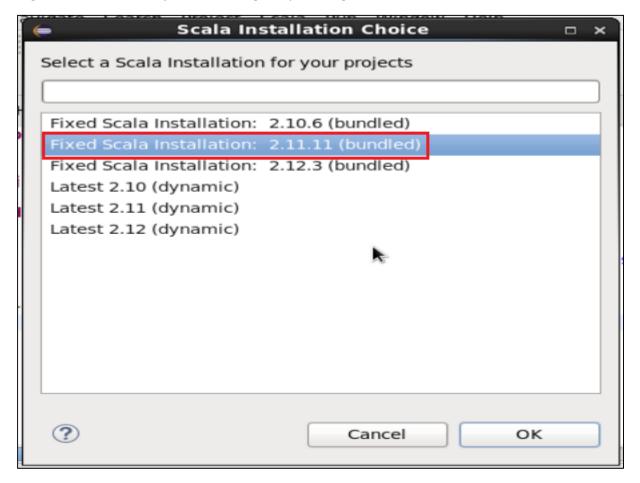
Step 3:

Step 4:

Adding the given code to Scala IDE. Since none of the required libraries are added, the project is with errors in the IDE.

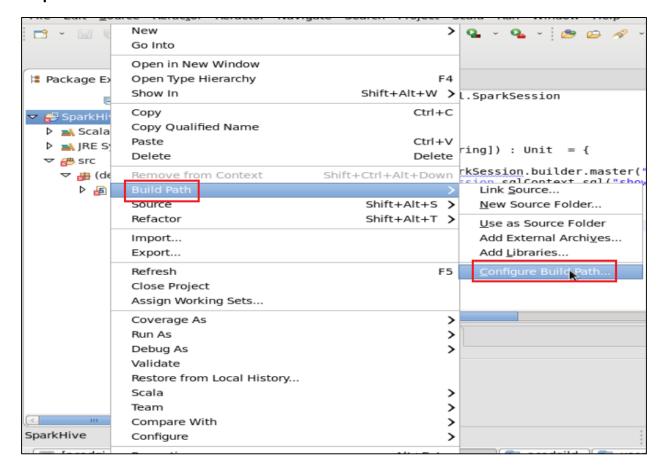


Right click on the Project in Package Explorer, go to Scala -> Set the Scala Intallation.



Select the specified Scala version (2.11) and click on OK.

Step 5:



To add required jars, right click on the Project in Package Explorer, got to Build Path -> Configure Build Path...

Navigate to the jars location of Spark(\$SPARK_HOME/jars) and Hive(\$HIVE_HOME/lib) and select all the jars into the build path. Click on Apply and Close. The Project is built and all the errors are removed from the code.

Step 6:

```
[acadgild@localhost ~]$ hive --service metastore
2018-07-18 14:01:35: Starting Hive Metastore Server
/home/acadgild/install/hive/apache-hive-2.3.2-bin/bin/ext/metastore.sh: line 29: export: `-Dproc_metastore -Dlog4j.configur
ationFile=hive-log4j2.properties '-Djava.util.logging.config.file=/home/acadgild/install/hive/apache-hive-2.3.2-bin/conf/parq
uet-logging.properties ': not a valid identifier
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]
2018-07-18T14:01:39,220 INFO [main] org.apache.hadoop.hive.conf.HiveConf - Found configuration file file:/home/acadgild/install/hive/apache-hive-2.3.2-bin/conf/hive-site.xml
2018-07-18T14:01:42,797 INFO [main] org.apache.hadoop.hive.metastore.HiveMetaStore - STARTUP_MSG:
STARTUP_MSG: dasspath = /home/acadgild/install/hive/apache-hive-2.3.2-bin/conf:/home/acadgild/install/hive/apache-hive-2.3
STARTUP_MSG: classpath = /home/acadgild/install/hive/apache-hive-2.3.2-bin/lib/accumulo-fate-1.6.0.jar:/home/acadgild/install/hive/apache-hive-2.3.2-bin/lib/accumulo-fate-1.6.0.jar:/home/acadgild/install/hive/apache-hive-2.3.2-bin/lib/accumulo-fate-1.6.0.jar:/home/acadgild/install/hive/apache-hive-2.3.2-bin/lib/accumulo-fate-1.6.0.jar:/home/acadgild/install/hive/apache-hive-2.3.2-bin/lib/accumulo-fate-1.6.0.jar:/home/acadgild/install/hive/apache-hive-2.3.2-bin/lib/accumulo-fate-1.6.0.jar:/home/acadgild/install/hive/apache-hive-2.3.2-bin/lib/accumulo-fate-1.6.0.jar:/home/acadgild/install/hive/apache-hive-2.3.2-bin/lib/accumulo-fate-1.6.0.jar:/home/acadgild/install/hive/apache-hive-2.3.2-bin/lib/accumulo-f
```

```
File Edit View Search Terminal Help

mer/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/lib/commons-io-2.4.jar:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/lib/leveldbjni-all-1.8.jar:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/lib/levey-core-1.9.jar:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/lib/levey-guice-1.9.jar:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/lib/levey-guice-1.9.jar:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/lib/levey-guice-1.9.jar:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/lib/jackson-mapper-ast-1.9.jar:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/lib/jackson-mapper-ast-1.9.jar:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/lib/guice-servlet-3.6.jar:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/lib/guice-servlet-3.6.2.final-jar:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/lib/guice-servlet-3.6.2.final-jar:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/lib/guice-sure-lib/guice-sure-lib/guice-3.6.2.final-jar:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/lib/guice-3.6.2.final-jar:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/lib/guice-3.6.2.final-jar:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/lib/guice-3.6.2.final-jar:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/mapreduce/lib/guice-3.6.5/share/hadoop/mapreduce/lib/guice-3.6.5/share/hadoop/mapreduce/lib/guice-3.6.5/share/hadoop/mapreduce/lib/guice-3.6.5/share/hadoop/mapreduce/hadoop-3.6.5/share/hadoop/mapreduce/hadoop-3.6.5/share/hadoop/mapreduce/hadoop-3.6.5/share/hadoop/mapreduce/hadoop-3.6.5/share/hadoop/mapreduce/hadoop-3.6.5/share/hadoop/mapreduce/hadoop-3.6.5/share/hadoop/mapreduce/hadoop-3.6.5/share/hadoop/mapreduce/hadoop-3.6.5/share/hadoop/mapreduce/hadoop-3.6.5/share/hadoop/mapreduce/hadoop-3.6.5/share/hadoop/mapreduce/hado
```

Hive Metastore is started using the following command.

hive --service metastore

As seen in the above screen shot, Hive metastore is successfully started.

Step 7:

On running the given script, we can see that the list of databases created in Hive are displayed.

Task 2: Integrate Spark with HBase

Following a similar procedure as with Hive, below steps are performed.

Step 1:

With respect to the screeshot above,

- 1: Ensuring that hadoop services are up.
- 2: Starting the hbase services.

Step 2:

The given code is created as a project in Scala IDE.

Step 3:

Scala installation is set 2.11

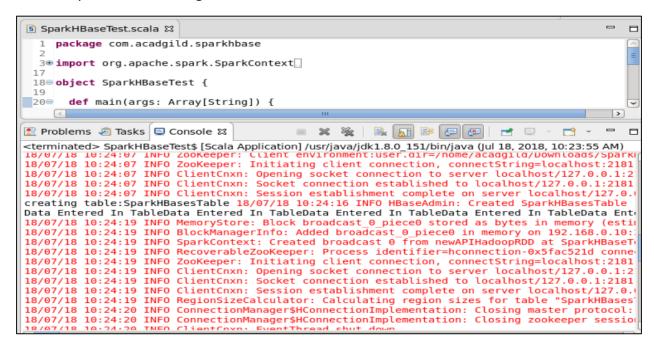
Step 4:

All the necessary jars related to Spark and Hbase are included in the Build Path.

```
package com.acadgild.sparkhbase
   3⊕ import org.apache.spark.SparkContext∏
  18@ object SparkHBaseTest {
         def main(args: Array[String]) {
  200
            // Create a SparkContext using every core of the local machine, named RatingsCounter val sc = new SparkContext("local[*]", "SparkHBaseTest")
            val conf = HBaseConfiguration.create()
             conf.set(TableInputFormat.INPUT TABLE,tablename)
  26
            cont.set(labteinputroimat.infor_index,
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()));
  29
30
  31
              admin.createTable(tableDescription);
} else {
                 print("table already exists")
  36
            val table = new HTable(conf, tablename);
for(x <- 1 to 10){
    var n = new Dut(new String("row" + x) getBytes());</pre>
 38
🖹 Problems 🥒 Tasks 📮 Console 🛭
                                                                                                                                                   <u></u>
No consoles to display at this time
```

Step 5:

The table 'SparkHBasesTable' is created in hbase and data is inserted into the table. The same is specified in the log as shown below



The Record Count of the table is also displayed.

```
18/07/18 10:24:22 INFO ConnectionManager$HConnectionImplementation: Closing zookeeper session 18/07/18 10:24:22 INFO ZooKeeper: Session: 0x164abb667d9000d closed 18/07/18 10:24:22 INFO ClientCnxn: EventThread shut down 18/07/18 10:24:22 INFO Executor: Finished task 0.0 in stage 0.0 (TID 0). 875 bytes result sen 18/07/18 10:24:22 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 1168 ms on 18/07/18 10:24:22 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 1168 ms on 18/07/18 10:24:22 INFO DAGSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completer 18/07/18 10:24:22 INFO DAGScheduler: ResultStage 0 (count at SparkHBaseTest.scala:45) finish 18/07/18 10:24:22 INFO DAGScheduler: Job 0 finished: count at SparkHBaseTest.scala:45, took 18/07/18 10:24:23 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopper 18/07/18 10:24:23 INFO MemoryStore: MemoryStore cleared 18/07/18 10:24:23 INFO BlockManager: BlockManager stopped 18/07/18 10:24:23 INFO BlockManagerMaster: BlockManagerMaster stopped 18/07/18 10:24:23 INFO BlockManagerMaster: BlockManagerMaster stopped 18/07/18 10:24:23 INFO SparkContext: Successfully stopped SparkContext 18/07/18 10:24:23 INFO SparkContext: Successfully stopped SparkContext 18/07/18 10:24:23 INFO ShutdownHookManager: Deleting directory /tmp/spark-e90e4ab7-90fb-42f3
```