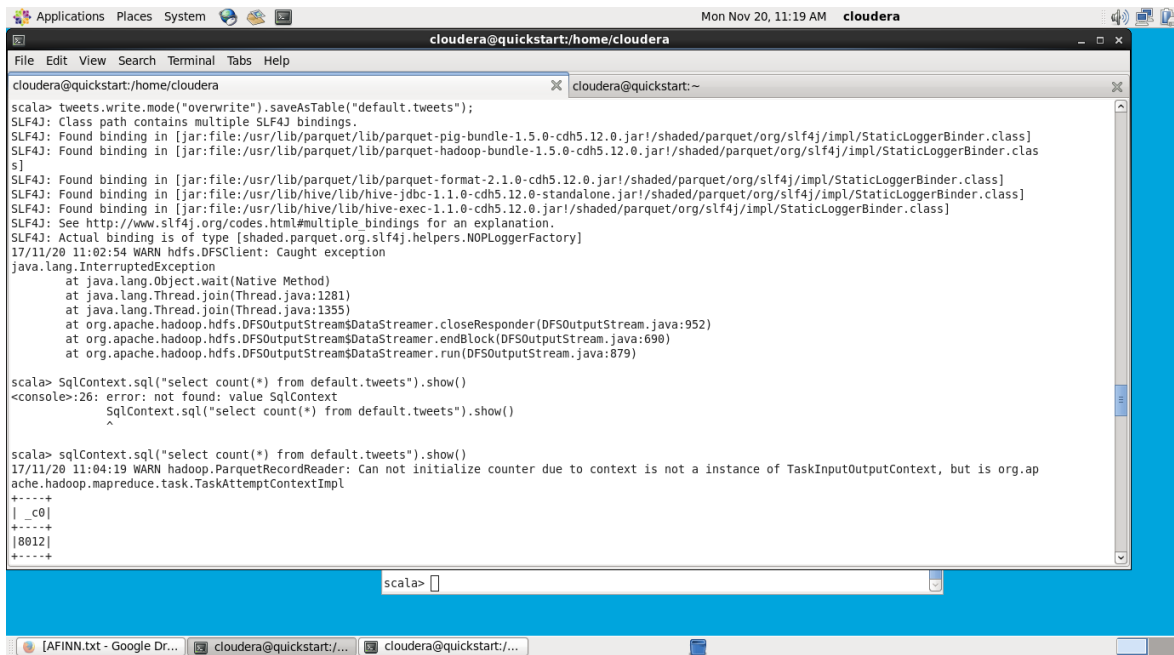


3. Create table from data frame created from step 2



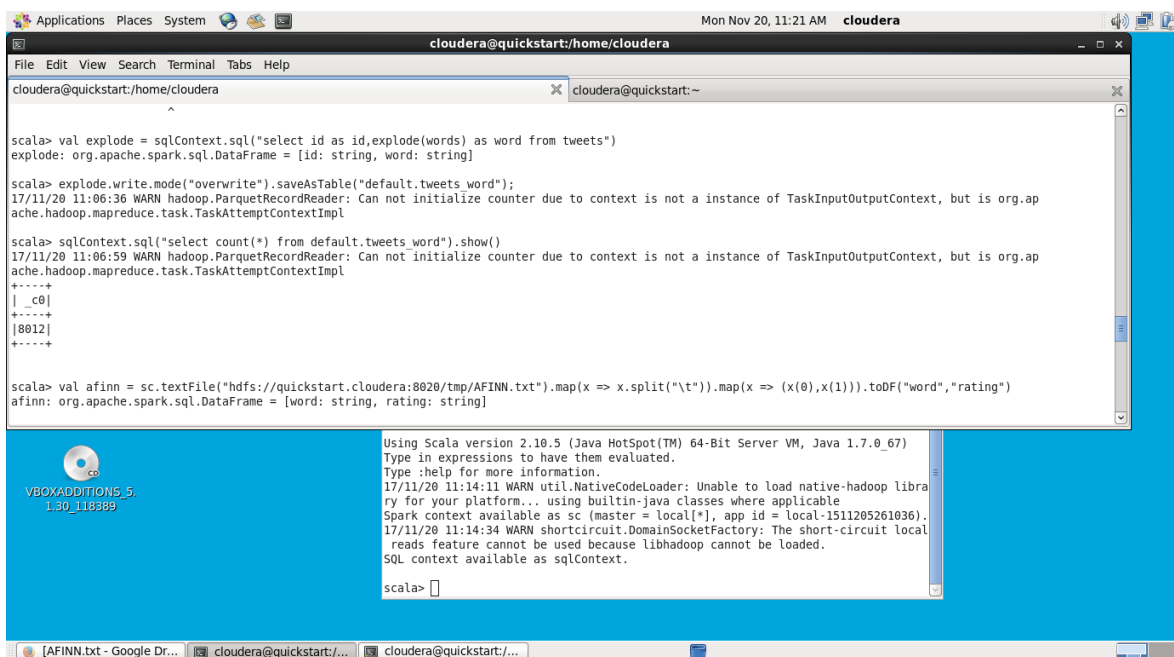
```
cloudera@quickstart:/home/cloudera
scala> tweets.write.mode("overwrite").saveAsTable("default.tweets");
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/parquet/lib/parquet-pig-bundle-1.5.0-cdh5.12.0.jar!/shaded/parquet/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/parquet/lib/parquet-hadoop-bundle-1.5.0-cdh5.12.0.jar!/shaded/parquet/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/parquet/lib/parquet-format-2.1.0-cdh5.12.0.jar!/shaded/parquet/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/hive-jdbc-1.1.0-cdh5.12.0-standalone.jar!/shaded/parquet/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/hive-exec-1.1.0-cdh5.12.0.jar!/shaded/parquet/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [shaded.parquet.org.slf4j.helpers.NOPLoggerFactory]
17/11/20 11:02:54 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:952)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:690)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:879)

scala> sqlContext.sql("select count(*) from default.tweets").show()
<console>:26: error: not found: value SqlContext
    sqlContext.sql("select count(*) from default.tweets").show()
    ^

scala> sqlContext.sql("select count(*) from default.tweets").show()
17/11/20 11:04:19 WARN hadoop.ParquetRecordReader: Can not initialize counter due to context is not a instance of TaskInputOutputContext, but is org.ap
ache.hadoop.mapreduce.task.TaskAttemptContextImpl
+----+
| _c0 |
+----+
| 8012 |
+----+

scala>
```

4. explode words from the step 3 and create another table



```
cloudera@quickstart:/home/cloudera
scala> val explode = sqlContext.sql("select id as id,explode(words) as word from tweets")
explode: org.apache.spark.sql.DataFrame = [id: string, word: string]

scala> explode.write.mode("overwrite").saveAsTable("default.tweets_word");
17/11/20 11:06:36 WARN hadoop.ParquetRecordReader: Can not initialize counter due to context is not a instance of TaskInputOutputContext, but is org.ap
ache.hadoop.mapreduce.task.TaskAttemptContextImpl

scala> sqlContext.sql("select count(*) from default.tweets_word").show()
17/11/20 11:06:59 WARN hadoop.ParquetRecordReader: Can not initialize counter due to context is not a instance of TaskInputOutputContext, but is org.ap
ache.hadoop.mapreduce.task.TaskAttemptContextImpl
+----+
| _c0 |
+----+
| 8012 |
+----+

scala> val afinn = sc.textFile("hdfs://quickstart.cloudera:8020/tmp/AFINN.txt").map(x => x.split("\t")).map(x => (x(0),x(1))).toDF("word","rating")
afinn: org.apache.spark.sql.DataFrame = [word: string, rating: string]

Using Scala version 2.10.5 (Java HotSpot(TM) 64-Bit Server VM, Java 1.7.0_67)
Type in expressions to have them evaluated.
Type :help for more information.
17/11/20 11:14:11 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
Spark context available as sc (master = local[*], app id = local-1511205261036).
17/11/20 11:14:34 WARN shortcircuit.DomainSocketFactory: The short-circuit local
reads feature cannot be used because libhadoop cannot be loaded.
SQL context available as sqlContext.

scala>
```

5. Read another text file in to Dataframe and create a table from the data frame

```
cloudera@quickstart:/home/cloudera
scala> val affinn = sc.textFile("hdfs://quickstart.cloudera:8020/tmp/AFINN.txt").map(x => x.split("\t")).map(x => (x(0),x(1))).toDF("word","rating")
affinn: org.apache.spark.sql.DataFrame = [word: string, rating: string]

scala> affinn.write.mode("overwrite").saveAsTable("default.affinn");
<console>:26: error: not found: value affinn
    affinn.write.mode("overwrite").saveAsTable("default.affinn");
    ^

scala> affinn.write.mode("overwrite").saveAsTable("default.affinn");
17/11/20 11:09:02 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:952)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:690)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:879)

scala> val join = spark.sql("select t.id,AVG(a.rating) as rating from default.tweet_word t join default.affinn a on t.word=a.word group by t.id order by
Using Scala version 2.10.5 (Java HotSpot(TM) 64-Bit Server VM, Java 1.7.0_67)
Type in expressions to have them evaluated.
Type :help for more information.
17/11/20 11:14:11 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Spark context available as sc (master = local[*], app id = local-1511205261036).
17/11/20 11:14:34 WARN shortcircuit.DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be loaded.
SQL context available as sqlContext.

scala>
```

6. Select average rating by going the tweet words and affinn words

```
cloudera@quickstart:/home/cloudera
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:606)
at org.apache.spark.deploy.SparkSubmit$.org$apache$spark$deploy$SparkSubmit$$runMain(SparkSubmit.scala:730)
at org.apache.spark.deploy.SparkSubmit$.doRunMain$1(SparkSubmit.scala:181)
at org.apache.spark.deploy.SparkSubmit$.submit(SparkSubmit.scala:206)
at org.apache.spark.deploy.SparkSubmit$.main(SparkSubmit.scala:121)
at org.apache.spark.deploy.SparkSubmit.main(SparkSubmit.scala)

scala> val join = sqlContext.sql("select t.id,AVG(a.rating) as rating from default.tweets_word t join default.affinn a on t.word=a.word group by t.id or
der by rating desc").show
17/11/20 11:10:36 WARN parquet.CorruptStatistics: Ignoring statistics because created by is null or empty! See PARQUET-251 and PARQUET-297
+-----+
| id|rating|
+-----+
|7357| 2.0|
|5332| 1.0|
+-----+

join: Unit = ()

[root@quickstart cloudera]#
Using Scala version 2.10.5 (Java HotSpot(TM) 64-Bit Server VM, Java 1.7.0_67)
Type in expressions to have them evaluated.
Type :help for more information.
17/11/20 11:14:11 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Spark context available as sc (master = local[*], app id = local-1511205261036).
17/11/20 11:14:34 WARN shortcircuit.DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be loaded.
SQL context available as sqlContext.

scala>
```