SPARK STREAMING: REAL TIME STOCK MARKET DATA PROCESSING (HANDS-ON LAB)

By www.HadoopExam.com

Note: These instructions should be used with the HadoopExam Apache Spark: Professional Trainings. Where it is executed and you can do hands on with trainer.

Problem statement: In stock market when trading is going on, and we wanted to find all the trade which are in doubt, for this we need to investigate all the trades in real-time.

In this example, we will write a Scala streaming application, which will find all the trades. Which has a trade volume of more than 500 and store in HDFS file system.

Input data format

```
Symbol,DATE,TIME,PRV_CLOSE,BID_PRICE,ASK_PRICE,SELL_PRICE,VOLUME
```

Sample data

```
IBM,03/10/14,9:01,500,501,504,499,1500
GOOGLE,03/10/14,9:01,600,602,609,601,2000
APPL,03/10/14,9:01,448,447,449,440,1600
HP,03/10/14,9:01,200,202,209,201,700
EMC,03/10/14,9:01,250,248,255,248,400
```

Solution:

1. **Import statements**

```
import org.apache.spark.SparkConf
import org.apache.spark.streaming.{Seconds, StreamingContext}
import StreamingContext._
```

2. **Create Case classes to represent our data**

```
case class Trade(symbol: String, date: String, time: String, prvclose: Double, bidprice: Double,
askprice: Double, sellprice: Double, volume: Double)extends Serializable
```

3. **Create StreamingContext (Entry point for the Streaming application) using SparkContext instance sc. We will be having batch interval as 10 seconds.**

```
val ssc = new StreamingContext(sc,Seconds(10))
```

4. **Create input stream, which will read data from hdfs directory "/user/cloudera/stream". As soon as new file arrived in this directory, it will start creating DStream (series of rdds).**

```
val tradeDStream = ssc.textFileStream("/user/cloudera/stream")
```

5. **Print DStream content**

```
tradeDStream.print()
```

6. **As we know DStream represent collections of RDDs. Hence, process each rdd as below.**

```
tradeDStream.foreachRDD(rdd=>{
val tradeRDD =rdd.map(_.split(",")).map(t => Trade(t(0), t(1), t(2), t(3).toDouble, t(4).toDouble,
t(5).toDouble, t(6).toDouble, t(7).toDouble))

//print first 10 records from each RDD
tradeRDD.take(10).foreach(println)

//filter all the trades , which has volume is more than 500 stocks
 val bigTrade = tradeRDD.filter(trade => trade.volume > 500)

//print filtered RDD
  bigTrade.take(2).foreach(println)

//Save filtered RDD in hdfs file system for further inspection
  bigTrade.saveAsTextFile("/user/cloudera/stream_out")

} )
```

7. Start streaming application to receive data, and await for computation to finish.

```
ssc.start()
ssc.awaitTermination()
```

8. Create file using following data in hdfs directory **"/user/cloudera/stream". Follow instructor, how he has created new files using Hue.**

```
IBM,03/10/14,9:01,500,501,504,499,1500
GOOGLE,03/10/14,9:01,600,602,609,601,2000
APPL,03/10/14,9:01,448,447,449,440,1600
HP,03/10/14,9:01,200,202,209,201,700
EMC,03/10/14,9:01,250,248,255,248,400
```

```
IBM,03/10/14,9:02,500,501,504,499,1000
GOOGLE,03/10/14,9:02,600,602,609,601,2000
APPL,03/10/14,9:02,448,447,449,440,780
HP,03/10/14,9:02,200,202,209,201,2000
EMC,03/10/14,9:02,250,248,255,248,1200
```

```
IBM,03/10/14,9:03,500,501,504,499,1200
GOOGLE,03/10/14,9:03,600,602,609,601,700
APPL,03/10/14,9:03,448,447,449,440,800
HP,03/10/14,9:03,200,202,209,201,1500
EMC,03/10/14,9:03,250,248,255,248,1600
```

```
IBM,03/10/14,9:04,500,501,504,499,502
GOOGLE,03/10/14,9:04,600,602,609,601,607
```

```
APPL,03/10/14,9:04,448,447,449,440,448
HP,03/10/14,9:04,200,202,209,201,206
EMC,03/10/14,9:04,250,248,255,248,253
```

```
IBM,03/10/14,9:05,500,501,504,499,502
GOOGLE,03/10/14,9:05,600,602,609,601,607
APPL,03/10/14,9:05,448,447,449,440,448
HP,03/10/14,9:05,200,202,209,201,206
EMC,03/10/14,9:05,250,248,255,248,253
```

9. Check the content in output directory on hdfs. We must be able to see all the trades, which has volume more than 500.

```
"/user/cloudera/stream_out"
```

HadoopExam Learning Resource provides the following material for the Advanced Technologies.
Please visit www.HadoopExam.com for more detail this is just a few products from portfolio.

Price start for training with Just $79/3500INR

Apache Spark Professional Training with HandsOn Session + Certification Material

Hadoop Professional Training with HandsOn Sessioon + Certification Material

HBase Professional Training with HandsOn Session + Certification Material

Certification Material

Certification Material

Certification Material

Certification Material

Certification Material

Certification Material

Certification Material