Big Data Analytics Assignment 2

Spark Streaming Algorithms

1. Kafka word Count with Spark Streaming (SCALA)

a. Go to http://aws.amazon.com/ and create an account.

b. Create an EC2 instance with Ubuntu machine (Take moderate machine requirements like 8 GB RAM, and makes the all connection open).

c. Download Putty on your local machine.

d. Do SSH to EC2 instance with our private key (It will open the EC2 Instance.)

e. Install JDK 1.7 , JRE and Kafka 2.11 version

- $ jps
- $ sudo apt-get install openjdk-7-jdk
- $ sudo apt-get update
- $ sudo apt-get install jre
- For Kakfa , run the following (http://www.bogotobogo.com/Hadoop/BigData_hadoop_Zookeeper_Kafka.php)
- $ wget http://www.webhostingjams.com/mirror/apache/kafka/0.8.2.1/kafka_2.11-0.8.2.1.tgz
- $ tar xvzf kafka_2.11-0.8.2.1.tgz
- Run the Kafka and Zookeeper server :
- $ cd kafka_2.11-0.8.2.1
- $ bin/zookeeper-server-start.sh config/zookeeper.properties (It will start the zoopkeeper on Zookeeper starts at localhost:2181)
- Open the duplicate instance of EC2.
- $ bin/kakfa-server-start.sh config/server.properties(It will start kafka on 9092 port.)
- $ bin/kafka-topics.sh --create --zookeeper ec2-54-208-147-4.compute-1.amazonaws.com:2181 --replication-factor 1 --partitions 1 --topic spark-topic

- $ bin/kafka-console-producer.sh --broker-list ec2-54-208-147-4.compute-1.amazonaws.com:9092 --topic spark-topic
- $ bin/kafka-console-consumer.sh --zookeeper ec2-54-208-147-4.compute-1.amazonaws.com:2181 --topic spark-topic --from-beginning

f. Go to Amazon services and create an EMR instance with SPARK 1.4.1 installed.

g. Connect the EMR through Putty by secure SSH connection

h. Go to spark directory under the /usr/lib/spark

i. Run the following commands :

- $ bin/run-example org.apache.spark.examples.streaming.KafkaWordCountProducer ec2-54-208-147-4.compute-1.amazonaws.com:9092 spark-topic 10 5
- Open the another instance of EMR by SSH through putty.
- $ bin/run-example org.apache.spark.examples.streaming.KafkaWordCount ec2-54-208-147-4.compute-1.amazonaws.com:2181 myconsumergroup spark-topic 1.

Note: ec2-54-208-147-4.compute-1.amazonaws.com is an EC2 instances, ie public IP address of EC2.

2. *Kafka word Count with Spark Streaming (Python)*

a. Follow the same step above, which was done for Scala code in the EC2 instance.

b. For EMR instance :

- $ go to spark directory : cd /usr/lib/spark
- Get the write permission on the directory:
- $ sudo chmod -R 777 $SPARK_HOME ($SPARK_HOME = *)
- $ wget http://central.maven.org/maven2/org/apache/spark/spark-streaming-kafka-assembly_2.10/1.4.1/spark-streaming-kafka-assembly_2.10-1.4.1.jar
- Run the kafka producer on the EC2 instance :
- $ bin/run-example org.apache.spark.examples.streaming.KafkaWordCountProducer <EC2 instance where Kafka is installed>:9092 spark-topic 10 5
- Run the below command on EMR again:
- $ bin/spark-submit --jars spark-streaming-kafka-assembly_2.10-1.4.1.jar examples/src/main/python/streaming/kafka_wordcount.py <EC2 instance where Kafka is installed>:2181 spark-topic

3. Amazon Kinesis word count with spark streaming :

a. Download the spark 1.4.1 source code from the spark website.
b. Go to Spark home.(cd c:/spark1.4.1/)
c. Build the code using Maven, with below command:
d. $ mvn -Pkinesis-asl -DskipTests clean package (It will take around an hour to build the binaries)
e. Follow the link to set up the Amazon Credentials to get your AWS_ACCESS_KEY_ID and AWS_SECRET_KEY. http://docs.aws.amazon.com/AWSSimpleQueueService/latest/SQSGettingStartedGuide/AWSCredentials.html
f. Set the Environment variable with AWS_ACCESS_KEY_ID and AWS_SECRET_KEY example :
g. Access key ID example: AKIAIOSFODNN7EXAMPLE
h. Secret access key example: wJalrXUtnFEMI/K7MDENG/bPxRfiCYEXAMPLEKEY
i. Follow the below link to create the Kinesis stream https://console.aws.amazon.com/kinesis/home?region=us-east-1
j. You can also follow the below link to find the correct parameter while creating the stream: http://docs.aws.amazon.com/kinesis/latest/dev/amazon-kinesis-streams.html
k. Once the stream is created go to your local machine where the spark 1.4.1 was build.
l. Run the following commands :
   - $ bin/run-example streaming.KinesisWordProducerASL mySparkStream https://kinesis.us-east-1.amazonaws.com 1000 10
   - $ bin/run-example streaming.KinesisWordCountASL myAppName mySparkStream https://kinesis.us-east-1.amazonaws.com

Note : mySparkStream is the name of the stream which you have created on the AWS.

4. MQTT word count with spark streaming :

a. GO to your Amazon EC2 instance and Run the following commands:

- $ sudo apt-get install mosquito (It will install the mqtt message broker server on your machine)
- Go to $ Spark_Home and run the following commands
- $ bin/run-example org.apache.spark.examples.streaming.MQTTPublisher tcp://localhost:1883 foo
- $ bin/run-example org.apache.spark.examples.streaming.MQTTWordCount tcp://localhost:1883 foo

5. ZeroMQ word count with spark streaming :

a. Go to Amazon EC2 instance and run the following steps:

- Download the ZeroMQ server (2.1.10) using the below link.
- $ wget http://download.zeromq.org/zeromq-2.1.10.tar.gz
- $ tar zxf zeromq-2.1.10.tar.gz
- $ cd zeromq-2.1.10
- Install the following using the commands :
- $ apt-get install make
- $ apt-get install libtool ,
- $ apt-get install pkg-config,
- $ apt-get install build-essential ,
- $ apt-get install autoconf,
- $ apt-get install automake,
- $ apt-get install uuid-dev,

- To install ZeroMQ system-wide Run
- $ sudo make install
- Run $ sudo ldconfig
- Now go to Spark home and run the following on two different instances of EC2:
- $ bin/run-example org.apache.spark.examples.streaming.SimpleZeroMQPublisher tcp://127.0.1.1:1234 foo.bar
- $ bin/run-example org.apache.spark.examples.streaming.ZeroMQWordCount tcp://127.0.1.1:1234 foo

6. HDFS word count with spark streaming (Scala):

a. Create the AWS EMR instance
b. Do secure SSH with Putty using the using machine hostname.
c. Run the following code:
    - $ hadoop fs -mkdir /input
    - Go to Spark home (/usr/lib/spark)
    - $ bin/run-example org.apache.spark.examples.streaming.HdfsWordCount hdfs:///input
    - Put the files in the HDFS input folder using duplicate instance of EMR using the below command.
    - $ hadoop fs -put /usr/lib/spark/data/mllib/*.txt hdfs:///input
    - $ hadoop fs -put /usr/lib/spark/examples/src/main/resources/*.txt hdfs:///input

7. HDFS word count with spark streaming (Python):

a. Create the AWS EMR instance

b. Do secure SSH with Putty using the using machine hostname.

c. Run the following code:

   - $ hadoop fs -mkdir /input
   - Go to Spark home (/usr/lib/spark)
   - $ bin/spark-submit examples/src/main/python/streaming/hdfs_wordcount.py hdfs:///input
   - Put the files in the HDFS input folder using duplicate instance of EMR using the below command.
   - $ hadoop fs -put /usr/lib/spark/data/mllib/*.txt hdfs:///input
   - $ hadoop fs -put /usr/lib/spark/examples/src/main/resources/*.txt hdfs:///input

8. Twitter Popular Tags with spark streaming

a. Create a twitter account.

b. Create the Twitter application to generate keys.

c. Open the application and get the consumer key, consumer secret, access token, access token secret

d. Run the following code:

   - Go to spark home(/usr/lib/spark)
   - $ bin/run-example org.apache.spark.examples.streaming.TwitterPopularTags < consumer key > < consumer secret > < access token > < access token secret >

For Example:

< consumer key > = 6RXPWt4yCQBbStPcibCbuDGKx

< consumer secret > = TI2uTXuhI3qkQmAOzLyMOJJ1RClsQ33O6FucvJAi1h46YaZa6C

< access token > = 220439848-NRnkROffJRONrWDB4GN9W0aCbELtBrpyMX6463jH

< access token secret > = r6VRyb6nZBRvR12rpGdBUKG7gmjcoorCMp2Jqa7HiUdkU

9. Network Word Count using Spark Streaming

a. Download Spark 1.4.1 with Hadoop binaries.
b. Open the terminal and create the data server $ nc -lk {port number}
c. Open another Terminal and Run the following code.
    - Go to spark home(/usr/lib/spark)
    - $ ./bin/run-example streaming.NetworkWordCount localhost portnumber
    - Now enter words in the data server file and press enter.
    - Now navigate back to example terminal and see the word count of the words written in the data server.

10. *Flume Event Count using Spark Streaming*

a. GO to your Amazon EC2 instance and Run the following commands:

b. Download the Apache flume server using the command,

c. $ wget http://supergsego.com/apache/flume/1.6.0/apache-flume-1.6.0-bin.tar.gz

d. tar xvzf apache-flume-1.6.0-bin

e. Go to Flume/conf and create the flume.conf file.
   Add the following properties in flume.conf:
   Sources , Sink and Channel properties from
   https://github.com/abhinavg6/spark-flume-stream

f. Run the following command on three different instances of EC2:
   - $ bin/flume-ng agent --conf ./conf/ -f conf/flume.conf -n agent1
   - $ bin/run-example org.apache.spark.examples.streaming.FlumeEventCount localhost 43333
   - $ ./flume-ng avro-client -c . -H localhost -p 43333