CS 4593/5463 Cloud and Big Data

Assignment 3: Spark Setup, and Programming
Due Midnight Monday, Mar 27, 2017

1. Continue with your hadoop cluster setup from Assignment 2. Make sure that the Namenode, and Datanodes are running. Make sure that the Austin police reports (2008-2011) are already uploaded to HDFS. Stop the JobTracker and Tasktrackers.

   ```
   $HADOOP_PREFIX/bin/stop-mapred.sh
   ```

   (Note: You have to be logged in as hduser)

2. Install Spark **on each** of your VMs (logging in as ubuntu)

   ```
   cd /usr/local
   sudo wget http://d3kbcqa49mib13.cloudfront.net/spark-1.6.1-bin-hadoop1.tgz
   sudo tar -xzf spark-1.6.1-bin-hadoop1.tgz
   sudo chown -R hduser:hadoop spark-1.6.1-bin-hadoop1
   ```

3. Configure Spark as follows:
   a) Login to the master VM

   ```
   su - hduser
   cd /usr/local/spark-1.6.1-bin-hadoop1
   cp ./conf/spark-env.sh.template ./conf/spark-env.sh
   cp ./conf/spark-defaults.conf.template ./conf/spark-defaults.conf
   cp ./conf/slaves.template ./conf/slaves
   ```

   b) Edit spark-env.sh file as follows:

   ```
   HADOOP_CONF_DIR=/usr/local/hadoop-1.2.1/conf
   SPARK_MASTER_IP=<ipaddress of your master VM>
   SPARK_MASTER_PORT=7077
   ```

   c) Edit spark-defaults.conf as follows:

   ```
   spark.eventLog.enabled=true
   spark.eventLog.dir=hdfs://vm:54310/user/spark/applicationHistory
   spark.history.fs.logDirectory=hdfs://vm:54310/user/spark/applicationHistory
   ```

   d) Edit slaves file inside the conf folder to include the hostnames or ip addresses of your slave VMs

e) Copy spark-env.sh and spark-defaults.sh files to all other VMs (using scp command).

f) Using HDFS command, create a directory as follows:

```
$HADOOP_PREFIX/bin/hadoop fs -mkdir /user/spark/applicationHistory
```

g) Edit the /etc/hosts file in the master VM to make sure that it includes the following line (in addition to other details).

```
127.0.0.1 localhost
```

4. Start Spark daemon processes while logged in to the master VM as hduser.

```
cd /usr/local/spark-1.6.1-bin-hadoop1
./sbin/start-master.sh
./sbin/start-slaves.sh
```

5. Write a Spark program (in Python) that will apply the k-means clustering algorithm to group together the data items of Austin Police Reports (2008-2011) according to the number of occurence of two crime types, i.e THEFT, and BURGLARY on each address. Assume that any crime type that has the word "BURGLARY" is considered to belong to the same category. The value of k can be chosen to be 5.

Important Note:

Job history can also be obtained from the Spark web interface:

```
http://129.115.xx.xx:8080

(Note: Edit security group in OpenStack cloud to add a new rule that allows
incoming traffic on port 8080)
```

Submission Policy and Deliverables

Only one submission per group is required. Submission should include the following.

1. Spark program (python files)
2. A PDF report that includes:

    a. One representative Screenshot of the console output when you execute the program.

b. One representative Screenshot of the Spark’s web interface which shows the status of running/completed jobs.
c. Output of your Spark program.
d. Describe how the work was divided among your group members.