CS 4843 Cloud Computing

Assignment 3: Spark Setup, and Programming

Due Midnight Tuesday, April 16, 2020

1. Continue with your hadoop cluster setup from Assignment 2. Make sure that the Namenode, and Datanodes are running. Make sure that the NYPD crime report is uploaded to HDFS. The crime report is available at
http://cs.utsa.edu/~plama/CS4843/NYPD_Complaint_Data_Current_YTD.csv

2. Follow the video lecture and related PowerPoint slides available on blackboard to setup the Spark cluster.

3. Write a Spark program (one program only) to answer the following:

   What are the top 3 crime types (use OFNS_DESC) that were reported in the month of July (use RPT_DT)? Crime types should be ranked based on the number of crimes reported in the month of July.

   How many crimes of type DANGEROUS WEAPONS were reported in the month of July ?

4. [Extra Credits] Write a Spark program that will apply the k-means clustering algorithm to group together various locations (PREM_TYPE_DESC) from NYPD crime report based on their similarity in the number of occurence of two crime types (OFNS_DESC), i.e DANGEROUS DRUGS, and DANGEROUS WEAPONS. The output of the program should include the following:

   a. Display 4 cluster centers. A cluster center is denoted by (x,y) where x is the number of DANGEROUS DRUGS on that location, and y is the number of DANGEROUS WEAPONS on that location.
   b. Find out which cluster does a location belong to if there were 10 DANGEROUS DRUGS and 50 DANGEROUS WEAPONS reported in that location.

   Hints:
   - Use Spark MLlib library
   - Your Spark program should perform the following steps:
     c. Prepare the data in the following key-value pair format

     (PREM_TYPE_DESC, (number of DANGEROUS DRUGS, number of DANGEROUS weapons))

     where PREM_TYPE_DESC is the key, and the tuple (number of DANGEROUS DRUGS, number of DANGEROUS weapons) is the value.

   The following transformations will be useful to prepare the data.
   filter, groupByKey, mapValues etc.

d. Extract the Values from the above key-value pairs, convert them into numpy array and call the KMeans.train( ) function (from the MLLib library) by passing the numpy array as a parameter as follows:

```
model=KMeans.train(data_array,4,maxIterations=20,
initializationMode="random")
```

Reading CSV file:

The NYPD police report is a CSV file. Please note that some of the comma separated values in this file have commas embedded inside double quotes. Therefore, a simple split(“,”) function will incorrectly split those special values. In order to avoid this issue, you need to import and use Python’s CSV module as follows:

```
from csv import reader
from pyspark.mllib.clustering import KMeans
from pyspark import SparkContext
import numpy as np

sc = SparkContext(appName="MySparkProg")
sc.setLogLevel("ERROR")

data = sc.textFile("hdfs://ipaddr:54310/hw2-input/")

# use csv reader to split each line of file into a list of elements.
# this will automatically split the csv data correctly.

splitdata = data.mapPartitions(lambda x: reader(x))

# If you need to extract PREM_TYPE_DES (location) and OFNS_DESC (crime
# type), it can be done as follows:

splitdata.map(lambda x: (x[16], x[7]))
```

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. Representative Screenshots of the console output when you execute the program.
   b. Output of your Spark program.
   c. Describe the contribution of each group members.