CS 4843 Cloud Computing

Assignment 3: Spark Setup, and Programming

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 ?

Hints

- The following Spark transformations and actions will be useful.
  Transformations : filter, map, reduceByKey, sortByKey etc.
  Actions: take, count etc.

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))
```

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.