CS 5573/EE 5523/IS 6973 Cloud Computing

Assignment 2: Spark Setup and Programming

1. Continue with your hadoop cluster setup from Assignment 1. Make sure that the Namenode, and Datanodes are running. Make sure that the NYPD crime report is already uploaded to HDFS. The crime report is available at

   http://cs.utsa.edu/~plama/CS5573/NYPD_Complaint_Data_Current_YTD.csv

   If the file is already present in Hadoop (/hw1-input), then you can use the same input directory for this assignment as well.

2. Run the following command to stop MapReduce processes (JobTracker and TaskTracker) to free up memory.

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

3. Follow the video lecture and related PowerPoint slides available on Canvas to setup the Spark cluster.

4. Install python2.7 on both VMs. This is needed since the Spark version used in this class is not fully compatible with python3.

   ```
   sudo apt install python2.7
   ```

5. The port number used by Spark process is blocked by Chameleon Cloud's strict firewall policy. You need to unblock them by running the following commands on both VMs.

   ```
   sudo firewall-cmd --zone=public --add-port=7077/tcp
   sudo firewall-cmd --zone=public --add-port=8080/tcp
   ```

6. On the master VM, edit the *.bashrc* file located in the home directory (/home/cc) by adding the following two lines.

   ```
   export PYSPARK_DRIVER_PYTHON=python2.7
   export PYSPARK_PYTHON=python2.7
   ```

   Save the file, and run the following command:

   ```
   source .bashrc
   ```

   This will force Spark to use python2.7 instead of python3. This is needed since the Spark version used in this class is not fully compatible with python3.

7. Write a Spark program to answer all of the following:

   a) *Where is most of the crime happening in New York? And what is the total number of crimes reported in that location ?*

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

   c) *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, sortBy etc.
  Actions: take, count etc.

  pyspark.RDD.sortBy — PySpark 3.1.2 documentation (apache.org)
  pyspark.RDD.reduceByKey — PySpark 3.1.2 documentation (apache.org)

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 import SparkContext

sc = SparkContext(appName="hw2")
sc.setLogLevel("ERROR")

# read input data from HDFS to create an RDD
data = sc.textFile("hdfs://groupX-1:54310/hw1-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))

# use filter to select only those rows in which crime type is not blank
splitdata = splitdata.filter(lambda x: x[7])

…rest of your code goes here…
```

(Note: replace groupX-1 with your group’s VM name.)

Program Execution:

(a) Test and debug your code using **pyspark** shell.

```
cd /usr/local/spark-1.6.1-bin-hadoop1
bin/pyspark --master spark://groupX-1:7077
```

(b) Executing your Spark program.

```
cd /usr/local/spark-1.6.1-bin-hadoop1
bin/spark-submit --master spark://groupX-1:7077 /home/cc/hw2.py
```

Submission Policy and Deliverables

1. Group Submission should include the following.
    a. Spark program (1 python file: hw2.py)
    b. A PDF report that includes:
        i. Representative Screenshots of the console output when you execute the program.
        ii. Describe the contribution of each group members.

2. Each student needs to submit a Self & Peer Evaluation form available on Canvas. The evaluation will be considered for assignment grading.