Team 8: Amjad Alqahtani, Caleb Alva, Nicholas Winkelmann

Cloud Computing- 5573

Dr. Glenn Brown

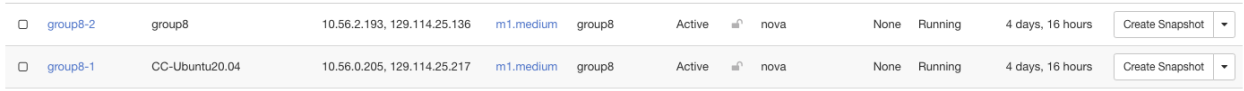
29 September 2024

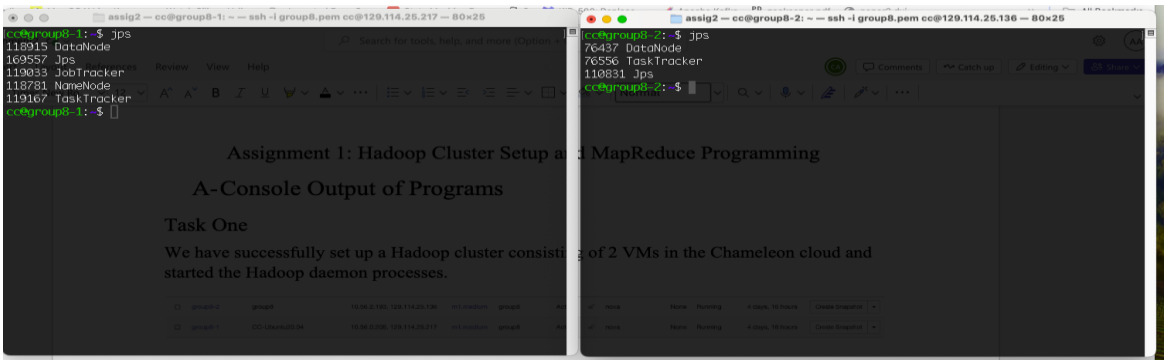
Assignment 1: Hadoop Cluster Setup and MapReduce Programming

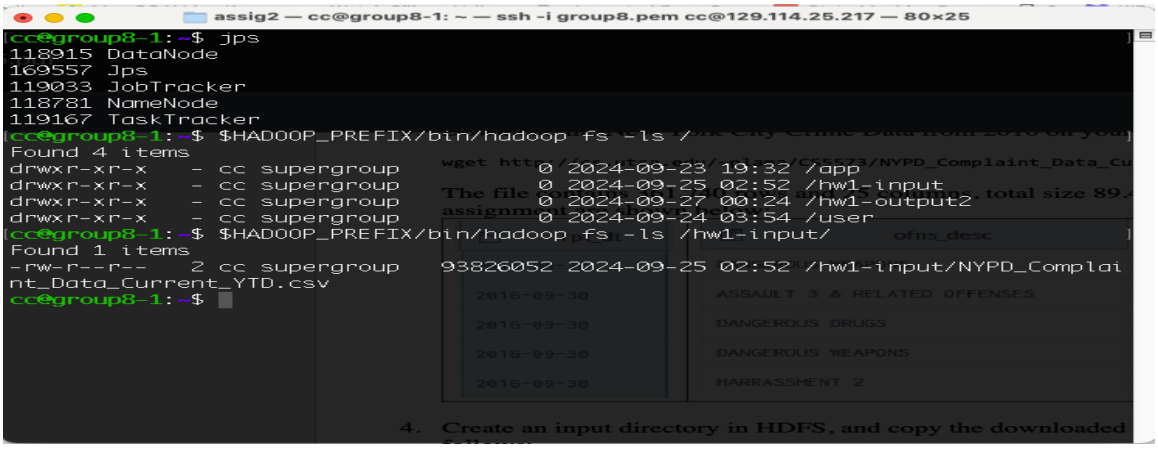
A-Console Output of Programs

Task One

We have successfully set up a Hadoop cluster with 2 VMs in the Chameleon cloud, started the Hadoop daemon processes, created the necessary directories in HDFS, and added the dataset.







Task Two

Write a MapReduce program in Python 3 (hw1-mapper1.py and hw1-reducer1.py) that will answer the following based on New York City Crime Data 2016. Run the program with only one reduce task.

a) Where is most of the crime happening in New York? (e.g BRONX, QUEENS, BROOKLYN, etc.)

b) What is the total number of crimes reported in that location ?

c) What types of crime are happening in that location (show unique crime types) ?

Below is the terminal screen of us running our hw1-mapper1.py and hw1-reducer1.py on HADOOP with:

$HADOOP_PREFIX/bin/hadoop jar $HADOOP_PREFIX/contrib/streaming/hadoop-streaming-.jar -input /hw1-input -output /hw1-output -file /home/cc/hw1-mapper1.py -mapper /home/cc/hw1-mapper1.py -file /home/cc/hw1-reducer1.py -reducer /home/cc/hw1-reducer1.py*

You might have to *zoom in*.

```
cc@group8-1:~$ $HADOOP_PREFIX/bin/hadoop jar $HADOOP_PREFIX/contrib/streaming/hadoop-streaming-*.jar -input /hw1-input -output /hw1-output1 -file /home/cc/hw1-mapper1.py -mapper /home/cc/hw1-mapper1
.py -file /home/cc/hw1-reducer1.py -reducer /home/cc/hw1-reducer1.py
packageJobJar: [/home/cc/hw1-mapper1.py, /home/cc/hw1-reducer1.py, /app/hadoop/tmp/hadoop-unjar9219138039190260841/] [] /tmp/streamjob7357702889915292651.jar tmpDir=null
24/09/26 01:29:06 INFO util.NativeCodeLoader: Loaded the native-hadoop library
24/09/26 01:29:06 WARN snappy.LoadSnappy: Snappy native library not loaded
24/09/26 01:29:06 INFO mapred.FileInputFormat: Total input paths to process : 1
24/09/26 01:29:07 INFO streaming.StreamJob: getLocalDirs(): [/app/hadoop/tmp/mapred/local]
24/09/26 01:29:07 INFO streaming.StreamJob: Running job: job_202409251822_0054
24/09/26 01:29:07 INFO streaming.StreamJob: To kill this job, run:
24/09/26 01:29:07 INFO streaming.StreamJob: /usr/local/hadoop-1.2.1/libexec/../bin/hadoop job  -Dmapred.job.tracker=10.56.0.205:54311 -kill job_202409251822_0054
24/09/26 01:29:07 INFO streaming.StreamJob: Tracking URL: http://group8-1:50030/jobdetails.jsp?jobid=job_202409251822_0054
24/09/26 01:29:08 INFO streaming.StreamJob:  map 0%  reduce 0%
24/09/26 01:29:12 INFO streaming.StreamJob:  map 50%  reduce 0%
24/09/26 01:29:13 INFO streaming.StreamJob:  map 100%  reduce 0%
24/09/26 01:29:19 INFO streaming.StreamJob:  map 100%  reduce 33%
24/09/26 01:29:21 INFO streaming.StreamJob:  map 100%  reduce 100%
24/09/26 01:29:22 INFO streaming.StreamJob: Job complete: job_202409251822_0054
24/09/26 01:29:22 INFO streaming.StreamJob: Output: /hw1-output1
```

After our output is created, we see that it gives us the desired statistics:

```
cc@group8-1:~$ $HADOOP_PREFIX/bin/hadoop fs -get /hw1-output1/part-00000 /home/cc
cc@group8-1:~$ cat part-00000
Most of the crimes were reported in BROOKLYN.
Total number of crimes reported in BROOKLYN is 106202.
Crime types reported in BROOKLYN are VEHICLE AND TRAFFIC LAWS, FRAUDULENT ACCOSTING, AGRICULTURE & MRKTS LAW-UNCLASSIFIED, CHILD ABANDONMENT/NON SUPPORT, OFFENSES RELATED TO CHILDREN, ADMINISTRATIVE
 CODE, ROBBERY, JOSTLING, OFFENSES INVOLVING FRAUD, INTOXICATED & IMPAIRED DRIVING, OFFENSES AGAINST THE PERSON, PETIT LARCENY OF MOTOR VEHICLE, OTHER OFFENSES RELATED TO THEF, FORGERY, OTHER STATE
LAWS (NON PENAL LAW), CRIMINAL MISCHIEF & RELATED OF, HOMICIDE-NEGLIGENT, RAPE, ESCAPE 3, GRAND LARCENY OF MOTOR VEHICLE, OTHER STATE LAWS (NON PENAL LA, THEFT OF SERVICES, OFF. AGNST PUB ORD SENSBL
TY &, NEW YORK CITY HEALTH CODE, PETIT LARCENY, ASSAULT 3 & RELATED OFFENSES, BURGLARY, GAMBLING, OFFENSES AGAINST PUBLIC ADMINI, ARSON, CRIMINAL TRESPASS, FRAUDS, DANGEROUS DRUGS, UNLAWFUL POSS. WE
AP. ON SCHOOL, HARRASSMENT 2, LOITERING/GAMBLING (CARDS, MISCELLANEOUS PENAL LAW, OFFENSES AGAINST PUBLIC SAFETY, NYS LAWS-UNCLASSIFIED VIOLATION, POSSESSION OF STOLEN PROPERTY, PROSTITUTION & RELAT
ED OFFENSES, FELONY ASSAULT, BURGLAR'S TOOLS, DISORDERLY CONDUCT, DANGEROUS WEAPONS, KIDNAPPING, MURDER & NON-NEGL. MANSLAUGHTER, SEX CRIMES, THEFT-FRAUD, NYS LAWS-UNCLASSIFIED FELONY, UNAUTHORIZED
USE OF A VEHICLE, ALCOHOLIC BEVERAGE CONTROL LAW, HOMICIDE-NEGLIGENT-VEHICLE, DISRUPTION OF A RELIGIOUS SERV, GRAND LARCENY, KIDNAPPING & RELATED OFFENSES, ENDAN WELFARE INCOMP, OTHER STATE LAWS.
```

Task Three

Write a MapReduce program in Python 3 (hw1-mapper2.py and hw1 reducer2.py) that will answer the following based on New York City Crime Data 2016. Run the program with two reduce tasks.

How many crimes of type “DANGEROUS WEAPONS” were reported on each month of the year 2016?

Below is the terminal screen of us running our hw1-mapper2.py and hw1-reducer2.py:

$HADOOP_PREFIX/bin/hadoop jar $HADOOP_PREFIX/contrib/streaming/hadoop-streaming-.jar -D mapred.reduce.tasks=2 -input /hw1-input -output /hw1-output2 -file /home/cc/hw1-mapper2.py -mapper /home/cc/hw1-mapper2.py -file /home/cc/hw1-reducer2.py -reducer /home/cc/hw1-reducer2.py*

```
cc@group8-1:~$ $HADOOP_PREFIX/bin/hadoop jar $HADOOP_PREFIX/contrib/streaming/hadoop-streaming-*.jar -D mapred.reduce.tasks=2 -input /hw1-input -output /hw1-output2 -file /home/cc/hw1-mapper2.py -mapper /home/cc/hw1-mapper2.py -file /home/cc/hw1-reducer2.py -reducer /home/cc/hw1-reducer2.py
packageJobJar: [/home/cc/hw1-mapper2.py, /home/cc/hw1-reducer2.py, /app/hadoop/tmp/hadoop-unjar3136327769938985O2/] [] /tmp/streamjob8317796231859681603.jar tmpDir=null
24/09/27 00:24:15 INFO util.NativeCodeLoader: Loaded the native-hadoop library
24/09/27 00:24:15 WARN snappy.LoadSnappy: Snappy native library not loaded
24/09/27 00:24:15 INFO mapred.FileInputFormat: Total input paths to process : 1
24/09/27 00:24:15 INFO streaming.StreamJob: getLocalDirs(): [/app/hadoop/tmp/mapred/local]
24/09/27 00:24:15 INFO streaming.StreamJob: Running job: job_202409251822_0061
24/09/27 00:24:15 INFO streaming.StreamJob: To kill this job, run:
24/09/27 00:24:15 INFO streaming.StreamJob: /usr/local/hadoop-1.2.1/libexec/../bin/hadoop job  -Dmapred.job.tracker=10.56.0.205:54311 -kill job_202409251822_0061
24/09/27 00:24:15 INFO streaming.StreamJob: Tracking URL: http://group8-1:50030/jobdetails.jsp?jobid=job_202409251822_0061
24/09/27 00:24:16 INFO streaming.StreamJob:  map 0%  reduce 0%
24/09/27 00:24:21 INFO streaming.StreamJob:  map 100%  reduce 0%
24/09/27 00:24:27 INFO streaming.StreamJob:  map 100%  reduce 17%
24/09/27 00:24:28 INFO streaming.StreamJob:  map 100%  reduce 33%
24/09/27 00:24:29 INFO streaming.StreamJob:  map 100%  reduce 67%
24/09/27 00:24:30 INFO streaming.StreamJob:  map 100%  reduce 100%
24/09/27 00:24:32 INFO streaming.StreamJob: Job complete: job_202409251822_0061
24/09/27 00:24:32 INFO streaming.StreamJob: Output: /hw1-output2
```

Below is the output of our program which gives us the desired statistics, split into two outputs based on the two reduce jobs we specified.

```
cc@group8-1:~$ $HADOOP_PREFIX/bin/hadoop fs -ls /hw1-output2
Found 4 items
-rw-r--r--   2 cc supergroup          0 2024-09-27 00:24 /hw1-output2/_SUCCESS
drwxr-xr-x   - cc supergroup          0 2024-09-27 00:24 /hw1-output2/_logs
-rw-r--r--   2 cc supergroup        156 2024-09-27 00:24 /hw1-output2/part-00000
-rw-r--r--   2 cc supergroup        156 2024-09-27 00:24 /hw1-output2/part-00001
cc@group8-1:~$ $HADOOP_PREFIX/bin/hadoop fs -cat /hw1-output2/part-00000
DANGEROUS WEAPONS reported per month:
January: 379
February: 446
March: 493
April: 565
May: 477
June: 489
July: 332
August: 453
September: 376

cc@group8-1:~$ $HADOOP_PREFIX/bin/hadoop fs -cat /hw1-output2/part-00001
DANGEROUS WEAPONS reported per month:
January: 488
February: 423
March: 518
April: 479
May: 482
June: 365
July: 369
August: 442
September: 482
```

B- Group Tasks

Amjad Alqahtani – wkh221

1. I contributed to setting up Hadoop environment and executing data processing tasks, ensuring the system is up and running
2. I contributed to debugging issues that arose during testing
3. I documented our processes, including setup instructions
4. I actively participated in group meetings and sharing insights
5. I contributed to the final report

Nicholas Winkelmann

1- Created group discord server for stream-lined communication and file sharing
2- Watched the lecture video and researched information based on HADOOP inquires
3- Assisted in group brainstorming
4- Wrote the code for hw1-mapper1.py and hw1-reducer1.py
5- Help write and organize final documentation.

Caleb Alva - uxy606

1- Helped debug the Hadoop environment during initial setup
2- Debugged issues during code deployments
3- Fixed formatting on hw1-reducer1.py to match output described in HW
4- Wrote the code for hw1-mapper2.py and hw1-reducer2.py
5- Provided screenshots of commands/ outputs for the report.