

HADOOP PROJECT FOR CRIME DATA ANALYSIS

Name: Supriya Kulkarni

Net ID: sgk140430

Hadoop Program Details:

The crime data has fields like Crime ID, northing, easting, area code, reported by and Crime Type. The map reduce program here is run based on the coordinate system (northing and easting) fields and the crime Type. Three Hadoop programs are written based on the same concept but some variations.

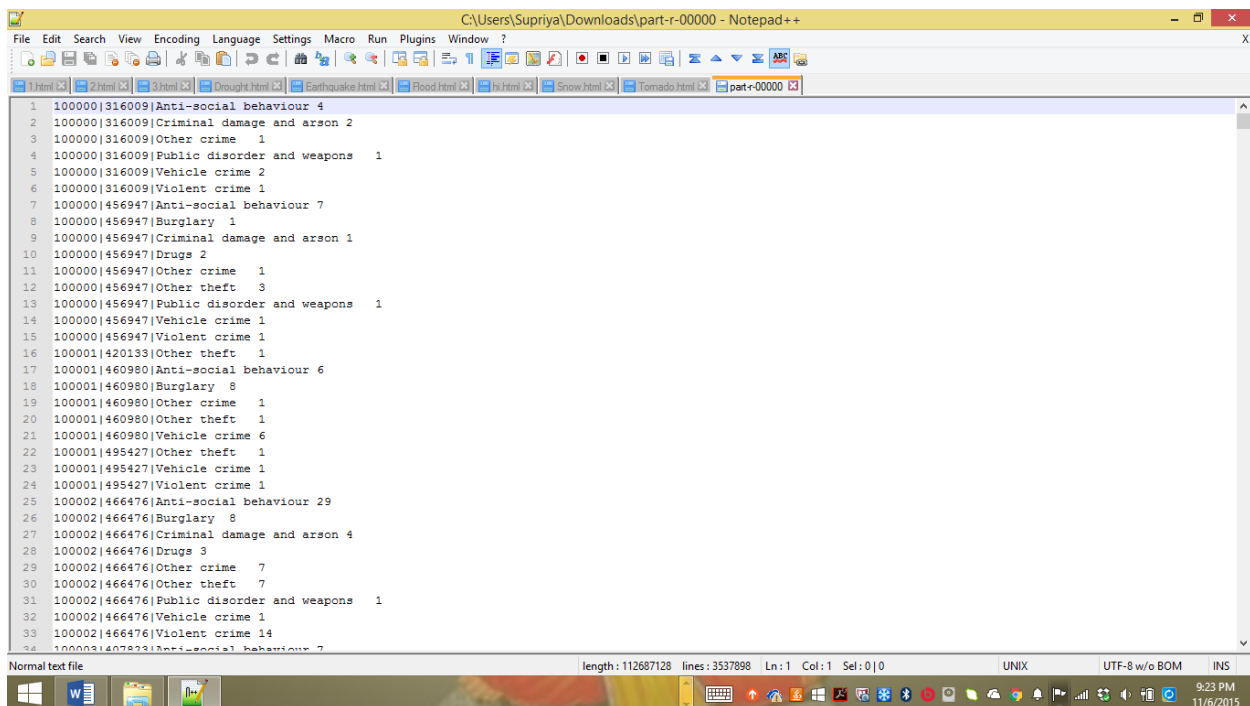
1) *Region 1*: Using the first character of north and eastern region code along with crime type as the key.

2) *Region 2*: Using the first three characters of north and eastern region code along with the crime type as the key.

3) *Region 3*: Using the all the characters from the north and eastern region code along with the crime type as the key.

In the above Hadoop programs in the mapper emits the key < (region code and crime type) ,one> and the reducer then groups the records with the similar keys and outputs the results.

Below is a screenshot of the output part file.



```
1 100000|316009|Anti-social behaviour 4
2 100000|316009|Criminal damage and arson 2
3 100000|316009|Other crime 1
4 100000|316009|Public disorder and weapons 1
5 100000|316009|Vehicle crime 2
6 100000|316009|Violent crime 1
7 100000|456947|Anti-social behaviour 7
8 100000|456947|Burglary 1
9 100000|456947|Criminal damage and arson 1
10 100000|456947|Drugs 2
11 100000|456947|Other crime 1
12 100000|456947|Other theft 3
13 100000|456947|Public disorder and weapons 1
14 100000|456947|Vehicle crime 1
15 100000|456947|Violent crime 1
16 100001|420133|Other theft 1
17 100001|460980|Anti-social behaviour 6
18 100001|460980|Burglary 8
19 100001|460980|Other crime 1
20 100001|460980|Other theft 1
21 100001|460980|Vehicle crime 6
22 100001|495427|Other theft 1
23 100001|495427|Vehicle crime 1
24 100001|495427|Violent crime 1
25 100002|466476|Anti-social behaviour 29
26 100002|466476|Burglary 8
27 100002|466476|Criminal damage and arson 4
28 100002|466476|Drugs 3
29 100002|466476|Other crime 7
30 100002|466476|Other theft 7
31 100002|466476|Public disorder and weapons 1
32 100002|466476|Vehicle crime 1
33 100002|466476|Violent crime 14
```

The following are the observations made on a single node cluster with 1 reducer on CPU time, mapper execution times, reducer execution times on all the three programs mentioned above.

Observations:

1)

Program Type	MapExecution time(in milliseconds)	Reducer Execution time(in milliseconds)
Region 1	3978102	858309
Region2	3936484	872655
Region3	4073797	892069

The below is the screenshot of the output of the program:

```
Launched reduce tasks=1
Data-local map tasks=1341
Total time spent by all maps in occupied slots (ms)=4073797
Total time spent by all reduces in occupied slots (ms)=892069
Total time spent by all map tasks (ms)=4073797
Total time spent by all reduce tasks (ms)=892069
Total vcore-seconds taken by all map tasks=4073797
Total vcore-seconds taken by all reduce tasks=892069
Total megabyte-seconds taken by all map tasks=4171568128
Total megabyte-seconds taken by all reduce tasks=913478656
Map-Reduce Framework
  Map input records=15669890
  Map output records=15669890
  Map output bytes=516961488
  Map output materialized bytes=362720072
  Input split bytes=171543
  Combine input records=15669890
  Combine output records=10436131
  Reduce input groups=3537897
  Reduce shuffle bytes=362720072
  Reduce input records=10436131
  Reduce output records=3537897
  Spilled Records=20872262
  Shuffled Maps =1341
  Failed Shuffles=0
  Merged Map outputs=1341
  GC time elapsed (ms)=34927
  CPU time spent (ms)=1823490
  Physical memory (bytes) snapshot=355882766336
  Virtual memory (bytes) snapshot=1387707883520
  Total committed heap usage (bytes)=269903462400
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=2183910056
File Output Format Counters
  Bytes Written=105611334
```

It can be observed that the map execution time of all the programs remain almost the same as we are using the same input for all the programs but the sorting and shuffling phase takes longer for region 3 hence the reducer execution time will be longer for region 3 when compared to the other two (region 3 > region 2 > region 1) Likewise the reducer execution time of region 2 is greater than 1. The number of keys are highest in the region 3 program because all the characters of the codes are used and then region 2 has the next highest keys as we are using 3 characters of the north and eastern codes and then the least keys are present in region 1.

2)

Now, using the single node environment the number of reducers were increased in the Region and the execution times of the them are as below:

Program Type	No.of Reducers	MapExecution time(in milliseconds)	Reducer Execution time(in milliseconds)
Region 3	1	4073797	892069
Region 3	10	4105679	5657123
Region3	25	4204579	8178123

As the number of the reducers increases the reducer execution time increases. Because as the reducers increase the sorting and shuffling phase will be longer because it has to deal with many reducers.

3)

Multiple node cluster was setup with one name node and two datanodes and the program Region 3 was run with different number of reducers and the following observed:

Program Type	No.of Reducers	MapExecution time(in milliseconds)	Reducer Execution time(in milliseconds)
Region 3	1	4905363	827431
Region 3	10	5305679	3976189
Region3	25	5904579	7778276

It was observed that the with multi node cluster the reducer times was decreased when compared to the single node cluster with the same program and same number of reducers.