Project 1.1

ACADGILD

Big Data and Hadoop Development

Project 1.1 - USA Crime Analysis

Downloaded data from given link

1. Associated Data Files

https://drive.google.com/file/d/0B1QaXx7tpw3SaUJHOHBZclBXWG8/view?usp=sharing

Dataset Description:

ID,Case Number,Date,Block,IUCR,Primary Type,Description,Location Description,Arrest,Domestic,Beat,District,Ward,Community Area,FBICode,X Coordinate,Y Coordinate,Year,Updated On,Latitude,Longitude,Location

Problem Statement

1. ***Write a MapReduce/Pig program to calculate the number of cases investigated under each FBI code***

Code:-

```
grunt> A = load '/user/cloudera/crime.csv' using PigStorage(',') as(id:long,case_no:chararray,block:chararray,iucr:int,primary_type:chararray,desc:chararray,l_d
```

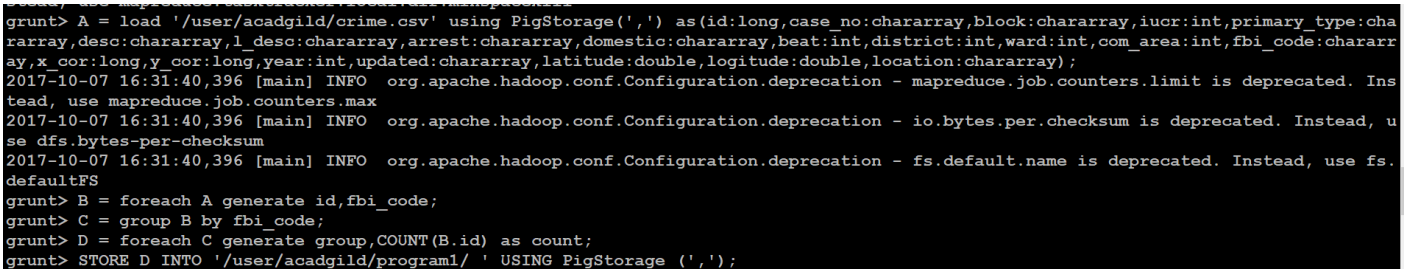
esc:chararray,arrest:chararray,domestic:chararray,beat:int,district:int,ward:int
,com_area:int,fbi_code:chararray,x_cor:long,y_cor:long,year:int,updated:chararra
y,latitude:double,logitude:double,location:chararray);

grunt> B = foreach A generate id,fbi_code;

grunt> C = group B by fbi_code;

grunt> D = foreach C generate group,COUNT(B.id) as count;

grunt> STORE D INTO '/user/acadgild/program1/ ' USING PigStorage (',');



```
Input(s):
Successfully read 0 records from: "/user/acadgild/crime.csv"

Output(s):
Successfully stored 0 records in: "/user/acadgild/program1"

Counters:
Total records written : 0
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1507368771118_0002
```

```
[acadgild@localhost ~]$ hadoop fs -cat /user/acadgild/program1/part-r-00000
17/10/07 16:35:50 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes wher
e applicable
0,2
1,3856
2,3395
3,3989
4,1936
5,1576
6,5963
7,3949
```

Output:-

1,3856

2,3395

3,3989

4,1936

5,1576

6,5963

7,3949

8,9592

9,279

10,1289

11,1272

12,495

13,840

14,2622

15,3637

16,3064

17,1759

18,620

19,5234

20,1827

21,2520

22,5219

23,9116

24,7289

25,19491

26,6336

27,5713

28,8620

29,8941

30,4697

31,2663

32,7851

33,1956

34,1197

35,2486

36,689

37,979

38,3347

39,1473

40,2862

41,1612

42,4262

43,10070

44,6652

45,1482

46,5602

47,406

48,1644

49,7349

50,1138

51,2243

52,1421

53,4419

54,1326

55,580

56,1943

57,956

58,2998

59,1167

60,1780

61,5356

62,1070

63,2568

64,1031

65,2236

66,6808

67,8077

68,7660

69,7134

70,2576

71,8273

72,1098

73,3355

74,619

75,2240

76,1810

77,2381

001,269

002,434

003,309

004,501

005,401

006,331

007,351

008,643

009,273

010,390

011,419

012,444

014,164

015,313

016,306

017,301

018,169

019,243

020,134

022,396

024,134

025,358

false,1

2. ***Write a MapReduce/Pig program to calculate the number of cases investigated under FBI code 32.***

Code:-

```
grunt> A = load '/user/acadgild/crime.csv' using PigStorage(',')
as(id:long,case_no:chararray,block:chararray,iucr:int,primary_type:chararray,desc
:chararray,l_desc:chararray,arrest:chararray,domestic:chararray,beat:int,district:i
nt,ward:int,com_area:int,fbi_code:chararray,x_cor:long,y_cor:long,year:int,updat
ed:chararray,latitude:double,logitude:double,location:chararray);
```

grunt> B = foreach A generate id,fbi_code;

grunt> C = filter B by fbi_code=='32';

grunt> D = group C by fbi_code;

grunt> E = foreach D generate group,COUNT(C.id) as count;

grunt> STORE E INTO '/user/acadgild/program2/ ' USING PigStorage (',');

```
grunt> A = load '/user/acadgild/crime.csv' using PigStorage(',') as(id:long,case_no:chararray,block:chararray,iucr:int,primary_type:cha
rarray,desc:chararray,l_desc:chararray,arrest:chararray,domestic:chararray,beat:int,district:int,ward:int,com_area:int,fbi_code:chararr
ay,x_cor:long,y_cor:long,year:int,updated:chararray,latitude:double,logitude:double,location:chararray);
2017-10-07 16:39:58,640 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Ins
tead, use mapreduce.job.counters.max
2017-10-07 16:39:58,640 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, u
se dfs.bytes-per-checksum
2017-10-07 16:39:58,640 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.
defaultFS
grunt> B = foreach A generate id,fbi_code;
grunt> C = filter B by fbi_code=='32';
grunt> D = group C by fbi_code;
grunt> E = foreach D generate group,COUNT(C.id) as count;
grunt> STORE E INTO '/user/acadgild/program2/ ' USING PigStorage (',');
```

```
Input(s):
Successfully read 0 records from: "/user/acadgild/crime.csv"

Output(s):
Successfully stored 0 records in: "/user/acadgild/program2"

Counters:
Total records written : 0
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1507368771118_0003
```

```
[acadgild@localhost ~]$ hadoop fs -ls /user/acadgild/program2
17/10/07 16:42:06 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform..
e applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2017-10-07 16:41 /user/acadgild/program2/_SUCCESS
-rw-r--r--   1 acadgild supergroup          8 2017-10-07 16:41 /user/acadgild/program2/part-r-00000
[acadgild@localhost ~]$ hadoop fs -cat /user/acadgild/program2/part-r-00000
17/10/07 16:42:23 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform..
e applicable
32,7851
```

Output:-32,7851

3. ***Write a MapReduce/Pig program to calculate the number of arrests in theft district wise.***

Code: -

```
grunt> A = load '/user/acadgild/crime.csv' using PigStorage(',')
as(id:long,case_no:chararray,block:chararray,iucr:int,primary_type:chararray,desc
:chararray,l_desc:chararray,arrest:chararray,domestic:chararray,beat:int,district:i
nt,ward:int,com_area:int,fbi_code:chararray,x_cor:long,y_cor:long,year:int,updat
ed:chararray,latitude:double,logitude:double,location:chararray);

grunt> B = foreach A generate primary_type,arrest,district;

grunt> C = filter B by primary_type matches '.*(THEFT).*' AND arrest=='true';

grunt> D = group C by district;

grunt> E = foreach D generate group,COUNT(C.primary_type) as count;

grunt> STORE E INTO '/user/acadgild/program3/ ' USING PigStorage (',');
```

```
grunt> A = load '/user/acadgild/crime.csv' using PigStorage(',') as(id:long,case_no:chararray,block:chararray,iucr:int,primary_type:cha
rarray,desc:chararray,l_desc:chararray,arrest:chararray,domestic:chararray,beat:int,district:int,ward:int,com_area:int,fbi_code:chararr
ay,x_cor:long,y_cor:long,year:int,updated:chararray,latitude:double,logitude:double,location:chararray);
2017-10-07 16:58:14,397 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Ins
tead, use mapreduce.job.counters.max
2017-10-07 16:58:14,397 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, u
se dfs.bytes-per-checksum
2017-10-07 16:58:14,398 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.
defaultFS
grunt> B = foreach A generate primary_type,arrest,district;
grunt> C = filter B by primary_type matches '.*(THEFT).*' AND arrest=='true';
grunt> D = group C by district;
grunt> E = foreach D generate group,COUNT(C.primary_type) as count;
grunt> STORE E INTO '/user/acadgild/program3/ ' USING PigStorage (',');
```

Output:-

1 1132

2 282

3 208

4 260

5 305

6 696

7 227

8 545

9 375

10 223

11 246

12 393

14 254

15 162

16 197

17 246

18 749

19 521

20 252

22 238

24 238

25 660

4. ***Write a MapReduce/Pig program to calculate the number of arrests done between October 2014 and October 2015.***

Code: -

```
grunt> A = load '/user/acadgild/crime.csv' using PigStorage(',')
as(id:long,case_no:chararray,date:chararray,block:chararray,iucr:int,parrest:chara
rray,domestic:chararray,beat:int,district:int,ward:int,com_area:int,fbi_code:chara
rray,x_cor:long,y_cor:long,year:int,updatchararray);

grunt> B = foreach A generate arrest, ToDate(date, 'MM/dd/yyyy HH:mm:ss aa')
as (date_mod:datetime);

grunt>  C = filter B by date_mod >= ToDate('10/01/2014 12:00:00
AM','MM/dd/yyyy hh:mm:ss aa') AND date_mod <= ToDate('11/01/2015 12:00:

grunt> D = group C all;

grunt> E = foreach D generate COUNT(C.arrest) as count;

grunt> STORE E INTO '/user/acadgild/program4/ ' USING PigStorage (',');
```

```
grunt> A = load '/user/acadgild/crime.csv' using PigStorage(',') as(id:long,case_no:chararray,date:chararray,block:chararray,iucr:int,p
arrest:chararray,domestic:chararray,beat:int,district:int,ward:int,com_area:int,fbi_code:chararray,x_cor:long,y_cor:long,year:int,updat
chararray);
2017-10-07 15:11:50,517 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Ins
2017-10-07 15:11:50,522 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, u
2017-10-07 15:11:50,522 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.
grunt> B = foreach A generate arrest, ToDate(date, 'MM/dd/yyyy HH:mm:ss aa') as (date_mod:datetime);
grunt>  C = filter B by date_mod >= ToDate('10/01/2014 12:00:00 AM','MM/dd/yyyy hh:mm:ss aa') AND date_mod <= ToDate('11/01/2015 12:00:
grunt> D = group C all;
grunt> E = foreach D generate COUNT(C.arrest) as count;
grunt> STORE E INTO '/user/acadgild/program4/ ' USING PigStorage (',');
2017-10-07 15:13:18,725 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Ins
```

```
2017-10-07 15:13:53,328 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.2.0   0.14.0  acadgild        2017-10-07 15:13:19     2017-10-07 15:13:53     GROUP_BY,FILTER

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime
job_1507368771118_0001  1       1       6       6       6       6       4       4       4       4       A,B,C,D,E       GROUP_BY,COMBIN

Input(s):
Successfully read 0 records from: "/user/acadgild/crime.csv"

Output(s):
Successfully stored 0 records in: "/user/acadgild/program4"

Counters:
Total records written : 0
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1507368771118_0001
```

Output:-63173

```
[acadgild@localhost ~]$ hadoop fs -cat /user/acadgild/program4/part-r-00000
17/10/07 15:14:50 WARN util.NativeCodeLoader: Unable to load native-hadoop library
63173
```