Input File : crime.csv

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.

==========

HDFS Input Commands to put input files/pig scripts/jar files on HDFS from local file system:

```
hadoop fs -mkdir '/user/cloudera/chhaya/pig_first_project/'

hadoop fs -copyFromLocal  '/home/cloudera/chhaya/pig_first_project/piggybank-0.17.0.jar' '/user/cloudera/chhaya/pig_first_project/'

hadoop fs -copyFromLocal  '/home/cloudera/chhaya/pig_first_project/crime.csv' '/user/cloudera/chhaya/pig_first_project/'

hadoop fs -copyFromLocal  '/home/cloudera/chhaya/pig_first_project/project11.pig' '/user/cloudera/chhaya/pig_first_project/'

hadoop fs -copyFromLocal  '/home/cloudera/chhaya/pig_first_project/project12.pig' '/user/cloudera/chhaya/pig_first_project/'

hadoop fs -copyFromLocal  '/home/cloudera/chhaya/pig_first_project/project13.pig' '/user/cloudera/chhaya/pig_first_project/'

hadoop fs -copyFromLocal  '/home/cloudera/chhaya/pig_first_project/project14.pig' '/user/cloudera/chhaya/pig_first_project/'

hadoop fs -ls  '/user/cloudera/chhaya/pig_first_project/'


//pig commands to execute the pig scripts at HDFS

pig -x mapreduce project11.pig
```

//project11.pig description start//

```
//registering the piggybank jar for apache pig operations

REGISTER 'piggybank-0.17.0.jar';

//defining the class for data storage in CSV EXCEL files.

DEFINE CSVExcelStorage org.apache.pig.piggybank.storage.CSVExcelStorage();

//Load statement declaration for the file in apache pig which is present at hdfs

A = LOAD '/user/cloudera/chhaya/pig_first_project/crime.csv' USING CSVExcelStorage(',') AS (

id:int,

case_number:chararray,

dated:chararray,

block:chararray,

iucr:int,
```

```
primary_type:chararray,
description:chararray,
location_description:chararray,
arrest:boolean,
domestic:boolean,
beat:int,
district:int,
ward:int,
community_area:int,
fbicode:chararray,
x_oordinate:int,
y_coordinate:int,
year:int,
updated_on:chararray,
latitude:float,
longitude:float,
location:chararray
);
// generate statement to store selected columns in an alias
B = FOREACH A GENERATE id as id,case_number AS case_number, fbicode as fbicode;
// filtering out null values from an alias
C = FILTER B BY fbicode IS NOT NULL AND id IS NOT NULL AND case_number IS NOT NULL ;
//grouping the values by a column in an alias
D = GROUP C BY fbicode;
// counting the records for each group in bag
E = FOREACH D GENERATE group,COUNT(C.fbicode);
// storing the final output in a file at HDFS
STORE E INTO '/user/cloudera/chhaya/pig_first_project/pigqueryoutput1.txt';

// dump statement to display the final output
DUMP E;
```

//project11.pig description end//

Query 1 Input Commands' Screenshots:

```
[cloudera@quickstart pig_project_1]$ hadoop fs -copyFromLocal  '/home/cloudera/chhaya/pig_first_project/crime.csv' '/user/cloudera/chhaya/pig_first_project/'
[cloudera@quickstart pig_project_1]$ hadoop fs -ls  '/user/cloudera/chhaya/pig_first_project/'
Found 7 items
-rw-r--r--   1 cloudera cloudera   69234933 2017-11-05 14:20 /user/cloudera/chhaya/pig_first_project/crime.csv
-rw-r--r--   1 cloudera cloudera   69234933 2017-11-05 14:18 /user/cloudera/chhaya/pig_first_project/crimes.csv
-rw-r--r--   1 cloudera cloudera     396335 2017-11-05 14:17 /user/cloudera/chhaya/pig_first_project/piggybank-0.17.0.jar
-rw-r--r--   1 cloudera cloudera        913 2017-11-05 14:17 /user/cloudera/chhaya/pig_first_project/project11.pig
-rw-r--r--   1 cloudera cloudera       1076 2017-11-05 14:17 /user/cloudera/chhaya/pig_first_project/project12.pig
-rw-r--r--   1 cloudera cloudera       1013 2017-11-05 14:17 /user/cloudera/chhaya/pig_first_project/project13.pig
-rw-r--r--   1 cloudera cloudera       1351 2017-11-05 14:17 /user/cloudera/chhaya/pig_first_project/project14.pig
[cloudera@quickstart pig_project_1]$
```

```
[cloudera@quickstart pig_project_1]$ pig -x mapreduce project11.pig
```

STORE COMMAND FINAL Output 1 screenshot:

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.6.0-cdh5.12.0 0.12.0-cdh5.12.0        cloudera        2017-11-05 14:23:37     2017-11-05 14:25:38     GROUP_BY,FILTER

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Featu
re      Outputs
job_1509735440995_0007  1       1       32      32      32      32      14      14      14      14      A,B,C,D,E       GROUP_BY,COMBINER       /user/cloudera/chhaya/pig_fir
st_project/pigqueryoutput1.txt,

Input(s):
Successfully read 291268 records (69235332 bytes) from: "/user/cloudera/chhaya/pig_first_project/crime.csv"

Output(s):
Successfully stored 26 records (213 bytes) in: "/user/cloudera/chhaya/pig_first_project/pigqueryoutput1.txt"

Counters:
Total records written : 26
Total bytes written : 213
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
```

```
[cloudera@quickstart pig_project_1]$ hadoop fs -cat /user/cloudera/chhaya/pig_first_project/pigqueryoutput1.txt/part-r-00000
02      1502
03      10596
05      14842
06      64329
07      11105
09      445
10      1551
11      13757
12      27
13      57
14      31301
15      3694
16      1787
17      1126
18      25207
19      434
20      1267
22      371
24      4046
26      29474
01A     533
01B     6
04A     4994
04B     7711
08A     14167
08B     46938
```

==========

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

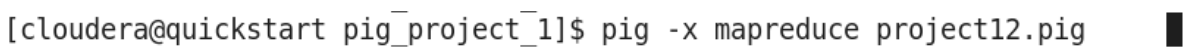
```
//pig commands to execute the pig scripts at HDFS
pig -x mapreduce project12.pig
//project12.pig description start//
//registering the piggybank jar for apache pig operations
REGISTER 'piggybank-0.17.0.jar';
//defining the class for data storage in CSV EXCEL files.
DEFINE CSVExcelStorage org.apache.pig.piggybank.storage.CSVExcelStorage();
//LOAD statement to load file present at HDFS in an alias
A = LOAD '/user/cloudera/chhaya/pig_first_project/crime.csv' USING CSVExcelStorage(',') AS (
id:int,
case_number:chararray,
dated:chararray,
block:chararray,
iucr:int,
primary_type:chararray,
description:chararray,
location_description:chararray,
arrest:boolean,
domestic:boolean,
beat:int,
district:int,
ward:int,
community_area:int,
fbicode:chararray,
x_oordinate:int,
y_coordinate:int,
year:int,
updated_on:chararray,
latitude:float,
longitude:float,
location:chararray
);
// foreach statement to store selected columns in an alias
B = FOREACH A GENERATE id as id,case_number AS cr, fbicode as fc;
```

```
// filtering out the null values and store in an alias
C = FILTER B BY id IS NOT NULL AND cr IS NOT NULL AND fc IS NOT NULL;
// group by data by fbi code and store it an an alias
D = GROUP C BY fc;
// counting the total number of cases registered against each fbi code
E = FOREACH D GENERATE group as fcode,COUNT(C.fc) as totalcount ;
// filtering the records where fbi code is 32 and store it an alias.
F = FILTER E BY fcode  == '32';
// storing the final output in an hdfs file
STORE F INTO '/user/cloudera/chhaya/pig_first_project/pigqueryoutput2.txt';
// dump statement to display the final output
DUMP F;
```

//project12.pig description end //

==========

Query 2 Input Commands' Screenshot:



STORE COMMAND Output 2 Screenshot:

```
Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Featu
re      Outputs
job_1509735440995_0012  1       1       13      13      13      13      9       9       9       9       A,B,C,D,E       GROUP_BY,COMBINER       hdfs://quickstart.cloudera:80
20/tmp/temp-1427728381/tmp-1811013530,

Input(s):
Successfully read 291268 records (69235332 bytes) from: "/user/cloudera/chhaya/pig_first_project/crime.csv"

Output(s):
Successfully stored 0 records in: "hdfs://quickstart.cloudera:8020/tmp/temp-1427728381/tmp-1811013530"

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_1509735440995_0012


2017-11-05 14:39:09,855 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning ACCESSING_NON_EXISTENT_FIELD 2 time
(s).
2017-11-05 14:39:09,855 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
```

==========

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

```
//pig commands to execute the pig scripts
pig -x mapreduce project13.pig
//project13.pig description start //
//registering the piggybank jar for apache pig operations
REGISTER 'piggybank-0.17.0.jar';
//defining the class for data storage in CSV EXCEL files.
DEFINE CSVExcelStorage org.apache.pig.piggybank.storage.CSVExcelStorage();
// loading the hdfs file in an alias
A = LOAD '/user/cloudera/chhaya/pig_first_project/crime.csv' USING CSVExcelStorage(',') AS (
id:int,
case_number:chararray,
dated:chararray,
block:chararray,
iucr:int,
primary_type:chararray,
description:chararray,
location_description:chararray,
arrest:boolean,
domestic:boolean,
beat:int,
district:int,
ward:int,
community_area:int,
fbicode:chararray,
x_oordinate:int,
y_coordinate:int,
year:int,
updated_on:chararray,
latitude:float,
longitude:float,
location:chararray
);
//storing selected columns in an alias
B = FOREACH A GENERATE district as district,primary_type as primary_type,arrest as arrest;
```

```
// filtering out the null values in a column

C = FILTER B BY district IS NOT NULL;

//  filtering only the values for only theft cases as per the usecase

D = FILTER C BY primary_type == 'THEFT';

// filtering only those values where arrest has been happened in theft cases

E = FILTER D BY arrest ;

// storing only district and theft cases only and storing it an alias

F = FOREACH E GENERATE district as district,primary_type as primary_type ;

// grouping the data districtwise

G = GROUP F BY district;

// generating the bag of grouped data districtwise with total count of occurence

H = FOREACH G GENERATE group,COUNT(F);

// storing the final output in an hdfs file

STORE H INTO '/user/cloudera/chhaya/pig_first_project/pigqueryoutput3.txt';

//dump statement to display final output

DUMP H;
```

//project13.pig description end//

==========

Query 3 Input Commands' Screenshots:

```
[cloudera@quickstart pig_project_1]$ pig -x mapreduce project13.pig
```

STORE COMMAND Output 3 Screenshot:

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.6.0-cdh5.12.0 0.12.0-cdh5.12.0        cloudera        2017-11-05 14:41:56     2017-11-05 14:43:07     GROUP_BY,FILTER

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Featu
re      Outputs
job_1509735440995_0013  1       1       24      24      24      24      12      12      12      12      A,B,C,F,G,H     GROUP_BY,COMBINER       /user/cloudera/chhaya/pig_fir
st_project/pigqueryoutput3.txt,

Input(s):
Successfully read 291268 records (69235332 bytes) from: "/user/cloudera/chhaya/pig_first_project/crime.csv"

Output(s):
Successfully stored 22 records (146 bytes) in: "/user/cloudera/chhaya/pig_first_project/pigqueryoutput3.txt"

Counters:
Total records written : 22
Total bytes written : 146
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
```

```
(1,1124)
(2,227)
(3,162)
(4,230)
(5,286)
(6,652)
(7,176)
(8,471)
(9,320)
(10,170)
(11,178)
(12,360)
(14,228)
(15,115)
(16,177)
(17,237)
(18,734)
(19,501)
(20,244)
(22,220)
(24,226)
(25,596)
2017-11-05 14:44:06,094 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-11-05 14:44:06,094 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
[cloudera@quickstart pig_project_1]$
```

==========

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

//pig commands to execute the pig scripts

pig -x mapreduce project14.pig

//project14.pig description start//

//registering the piggybank jar for apache pig operations

REGISTER 'piggybank-0.17.0.jar';

//defining the class for data storage in CSV EXCEL files.

DEFINE CSVExcelStorage org.apache.pig.piggybank.storage.CSVExcelStorage();

// loading the hdfs file an alias

A = LOAD '/user/cloudera/chhaya/pig_first_project/crime.csv' USING CSVExcelStorage(',') AS (

id:int,

case_number:chararray,

dated:chararray,

block:chararray,

iucr:int,

primary_type:chararray,

description:chararray,

location_description:chararray,

arrest:boolean,

domestic:boolean,

```
beat:int,
district:int,
ward:int,
community_area:int,
fbicode:chararray,
x_oordinate:int,
y_coordinate:int,
year:int,
updated_on:chararray,
latitude:float,
longitude:float,
location:chararray
);
// store selected columns in an alias
B = FOREACH A GENERATE dated as date ,primary_type as primary_type,arrest as arrest;
// filtering the THEFT cases
D = FILTER B BY primary_type == 'THEFT';
// filtering the THEFT cases where arrest has happened and storing the resultant in an alias.
E = FILTER D BY arrest ;
// FILTERING out the null values and storing the resultant in an alias
F = FILTER E BY date IS NOT NULL;
//  generating only date column and storing the resultant in an alias
G = FOREACH F GENERATE date;
// formatting the dates in same format(YYYYMMDD) as there are multiple date formats present in input hdfs file and storing them in an alias
H = FOREACH G GENERATE (
INDEXOF(date,'-',0)==2 ?
CONCAT(SUBSTRING(date,6,10),CONCAT(SUBSTRING(date,3,5),SUBSTRING(date,0,2))):
(INDEXOF(date,'/',0)==2 ?
CONCAT(SUBSTRING(date,6,10),CONCAT(SUBSTRING(date,0,2),SUBSTRING(date,3,5))):
SUBSTRING(date,0,10))
)
AS yyyymmdd;
// conversion of date into standard date with Builtin ToDate function
I = FOREACH H GENERATE ToDate(yyyymmdd,'YYYYMMDD') AS dt;
```

```
// filtering the values where cases where registered between Oct  2014 and Oct 2015
J = FILTER I BY dt > ToDate('2014-09-30') AND dt < ToDate('2015-11-01');
// grouping the data by date
K = GROUP J ALL;
// total occurrence of cases
L = FOREACH K GENERATE COUNT(J.dt);
// store statement to store the final output in an hdfs file
STORE L INTO '/user/cloudera/chhaya/pig_first_project/pigqueryoutput4.txt';
// Dump statement to display the final output
DUMP L;
```

//project14.pig description end//

Query 4 Input Commands' Screenshots:

```
[cloudera@quickstart pig_project_1]$ pig -x mapreduce project14.pig
```

STORE COMMAND Output 4 Screenshots:

```
Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   Me
re      Outputs
job_1509735440995_0016  1       1       15      15      15      15      12      12      12      12      A,B,D,H,I,J,K,L GROUP_BY,COMBINER
20/tmp/temp-474900874/tmp-1617713692,

Input(s):
Successfully read 291268 records (69235332 bytes) from: "/user/cloudera/chhaya/pig_first_project/crime.csv"

Output(s):
Successfully stored 1 records (7 bytes) in: "hdfs://quickstart.cloudera:8020/tmp/temp-474900874/tmp-1617713692"

Counters:
Total records written : 1
Total bytes written : 7
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
```

```
[cloudera@quickstart pig_project_1]$ hadoop fs -cat /user/cloudera/chhaya/pig_first_project/pigqueryoutput4.txt/part-r-00000
4563
[cloudera@quickstart pig_project_1]$
```

==========