PROJECT 1.1 – USA Crime Analysis

Introduction:

This dataset contains attributes related to crimes taking place in various areas like type of crime, FBI code related to that criminal case, arrest frequency, location of crime etc.

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

Sample Dataset:

10230953, HY418703, 09/10/2015 11:56:00 PM, 048XX W NORTH AVE, 0498, BATTERY, AGGRAVATED DOMESTIC BATTERY: HANDS/FIST/FEET SERIOUS INJURY, APARTMENT, true, true, 2533, 025, 37, 25, 04B, 1143637, 1910194, 2015, 09/17/2015 11:37:18 AM, 41.909605035, -87.747777145, "(41.909605035, -87.747777145)"

10230979, HY418750, 09/10/2015 11:55:00 PM, 120XX S PARNELL AVE, 0486, BATTERY, DOMESTIC BATTERY SIMPLE, ALLEY, true, true, 0523, 005, 34, 53, 08B, 1174806, 1825089, 2015, 09/17/2015 11:37:18 AM, 41.675427135, -87.63581257, "(41.675427135, -87.63581257)"

Load input dataset to HDFS:

Here is the command I have used to import this CSV file into HDFS:

$ hadoop fs –put /home/acadgild/USA_Crime_Analysis/Crimes_-_2001_to_present.csv /user/acadgild/

Verification of put command result:

```
[acadgild@localhost ~]$ hadoop fs -ls /user/acadgild/
18/01/26 06:40:52 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 9 items
-rw-r--r--   1 acadgild supergroup   69234930 2018-01-26 00:21 /user/acadgild/Crimes_-_2001_to_present.csv
drwxr-xr-x   - acadgild supergroup          0 2015-11-20 11:46 /user/acadgild/Pictures
drwxr-xr-x   - acadgild supergroup          0 2018-01-23 12:32 /user/acadgild/_sqoop
drwxr-xr-x   - acadgild supergroup          0 2018-01-23 10:13 /user/acadgild/emp_info
drwxr-xr-x   - acadgild supergroup          0 2018-01-25 22:08 /user/acadgild/employee_data
drwxr-xr-x   - acadgild supergroup          0 2015-11-17 02:03 /user/acadgild/oozie-acad
drwxr-xr-x   - acadgild supergroup          0 2015-11-17 02:00 /user/acadgild/share
drwxr-xr-x   - acadgild supergroup          0 2018-01-26 06:21 /user/acadgild/solution3
drwxr-xr-x   - acadgild supergroup          0 2018-01-25 18:19 /user/acadgild/twitter_data
[acadgild@localhost ~]$
```

Problem Statement:

1. Write a MapReduce/Pig program to calculate the number of cases investigated under each FBI code
2. Write a MapReduce/Pig program to calculate the number of cases investigated under FBI code 32.
3. Write a MapReduce/Pig program to calculate the number of arrests in theft district wise.
4. Write a MapReduce/Pig program to calculate the number of arrests done between October 2014 and October 2015.

Solution:

1. Here is the **pig script** to calculate the number of cases investigated under each FBI code:

pig_script_for_problem1.pig

```
crimes_data = LOAD '/user/acadgild/Crimes_-_2001_to_present.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX') AS (id:int, case_number:chararray, date:chararray, block:chararray, IUCR:int, primary_type:chararray, description:chararray, location_description:chararray, arrest:chararray, domestic:chararray, beat:int, district:int, ward:int, community_area:int, FBICode:chararray, x_coordinate:int, y_coordinate:int, year:int, updated_on:chararray, latitude:double, longitude:double, location:chararray);                                                    // line 1

crimes_data_filtered = FILTER crimes_data BY FBICode IS NOT NULL;          // line 2

crimes_data_group_by_FBI_code = GROUP crimes_data_filtered BY FBICode;  // line 3

FBI_code_with_cases = FOREACH crimes_data_group_by_FBI_code GENERATE group, COUNT (crimes_data_filtered.case_number);                                    // line 4

STORE FBI_code_with_cases INTO '/user/acadgild/solution1' USING PigStorage(',');     // line 5
```

Comments:

Line 1: loads the dataset 'Crimes_-_2001_to_present.csv' from HDFS path to the Pig environment. The column separator character here is comma (,). We have to provide the schema (column name and its datatype) for the dataset in this statement. The resulting dataset with the schema will be stored in a relation, in this case 'crimes_data'.

Line 2: filters out NULL values for FBI Code. The result will be stored in an intermediate relation 'crimes_data_filtered'.

Line 3: groups data in the relation 'crimes_data_filtered' by field 'FBICode'. The result will be stored in another intermediate relation 'crimes_data_group_by_FBI_code'.

Line 4: generates a pair of values (FBI Code, <number of cases for this FBI code>).

Line 5: Writes result to specified path in HDFS.

```
[acadgild@localhost ~]$ cat pig_script_for_problem1.pig
crimes_data = LOAD '/user/acadgild/Crimes_-_2001_to_present.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','
NO_MULTILINE','UNIX') AS (id:int, case_number:chararray, date:chararray, block:chararray, IUCR:int, primary_type:chararray, d
escription:chararray, location_description:chararray, arrest:chararray, domestic:chararray, beat:int, district:int, ward:int,
community_area:int, FBICode:chararray, x_coordinate:int, y_coordinate:int, year:int, updated_on:chararray, latitude:double,
longitude:double, location:chararray);
crimes_data_filtered = FILTER crimes_data BY FBICode IS NOT NULL;
crimes_data_group_by_FBI_code = GROUP crimes_data_filtered BY FBICode;
FBI_code_with_cases = FOREACH crimes_data_group_by_FBI_code GENERATE group, COUNT(crimes_data_filtered.case_number);
STORE FBI_code_with_cases INTO '/user/acadgild/solution1' USING PigStorage(',');
[acadgild@localhost ~]$
```

Script Execution:

Here is the command to execute this Pig script in MapReduce mode:

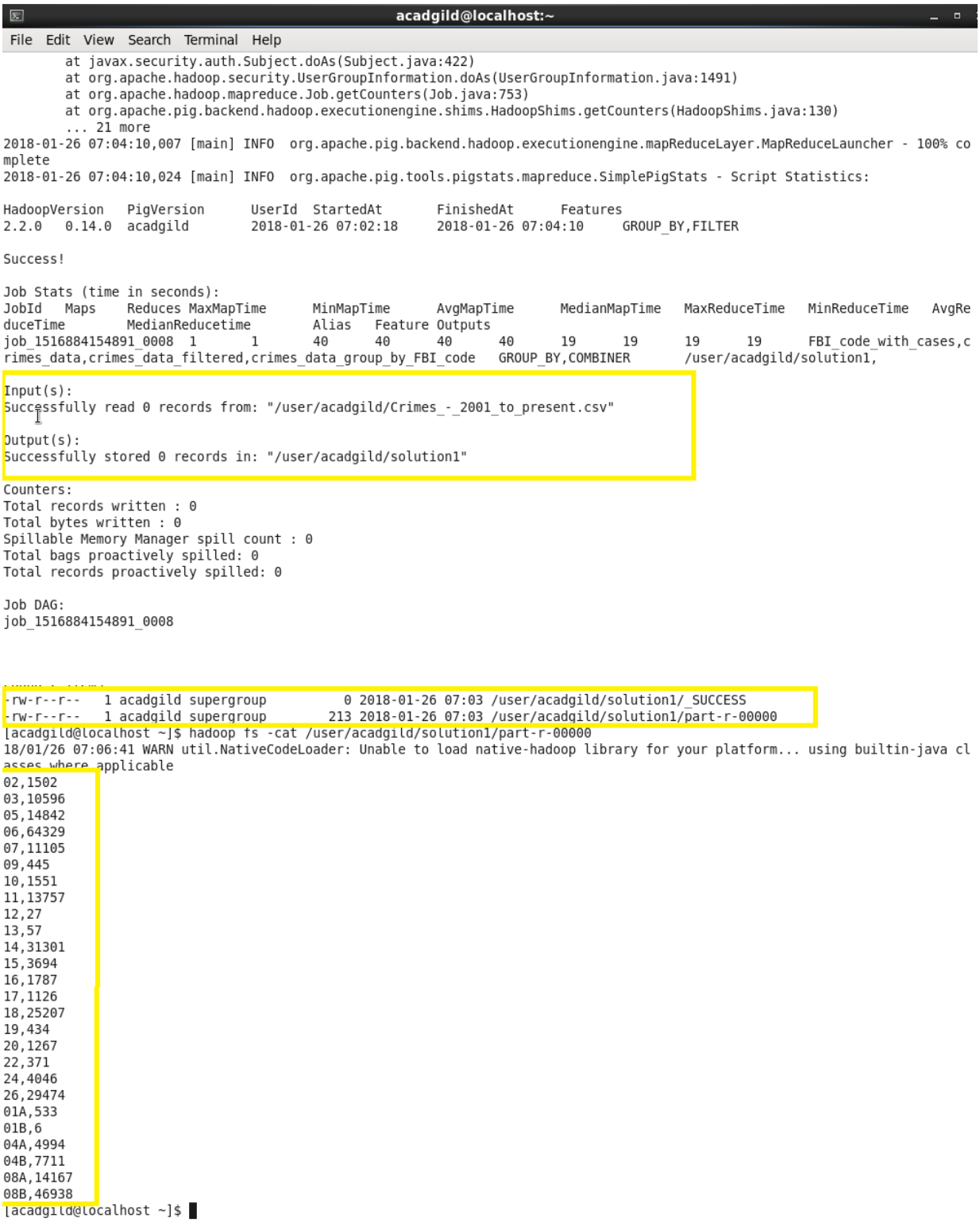
pig –f pig_script_for_problem1.pig

Note: The input dataset should be present in HDFS path before executing the script

We can also execute the same script in local mode using the command below:

pig –x local pig_script_for_problem1.pig

Output:


acadgild@localhost:~
File Edit View Search Terminal Help
at javax.security.auth.Subject.doAs(Subject.java:422)
at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1491)
at org.apache.hadoop.mapreduce.Job.getCounters(Job.java:753)
at org.apache.pig.backend.hadoop.executionengine.shims.HadoopShims.getCounters(HadoopShims.java:130)
... 21 more
2018-01-26 07:04:10,007 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2018-01-26 07:04:10,024 [main] INFO org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:
HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.2.0 0.14.0 acadgild 2018-01-26 07:02:18 2018-01-26 07:04:10 GROUP_BY,FILTER
Success!
Job Stats (time in seconds):
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime MedianReducetime Alias Feature Outputs
job_1516884154891_0008 1 1 40 40 40 40 19 19 19 19 FBI_code_with_cases,crimes_data,crimes_data_filtered,crimes_data_group_by_FBI_code GROUP_BY,COMBINER /user/acadgild/solution1,
Input(s):
Successfully read 0 records from: "/user/acadgild/Crimes_-_2001_to_present.csv"
Output(s):
Successfully stored 0 records in: "/user/acadgild/solution1"
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_1516884154891_0008
-rw-r--r-- 1 acadgild supergroup 0 2018-01-26 07:03 /user/acadgild/solution1/_SUCCESS
-rw-r--r-- 1 acadgild supergroup 213 2018-01-26 07:03 /user/acadgild/solution1/part-r-00000
[acadgild@localhost ~]$ hadoop fs -cat /user/acadgild/solution1/part-r-00000
18/01/26 07:06:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
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
[acadgild@localhost ~]$

2. Here is the **pig script** to calculate the number of cases investigated under FBI code 32:

pig_script_for_problem2.pig

```
crimes_data = LOAD '/user/acadgild/Crimes_-_2001_to_present.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX') AS (id:int, case_number:chararray, date:chararray, block:chararray, IUCR:int, primary_type:chararray, description:chararray, location_description:chararray, arrest:chararray, domestic:chararray, beat:int, district:int, ward:int, community_area:int, FBICode:chararray, x_coordinate:int, y_coordinate:int, year:int, updated_on:chararray, latitude:double, longitude:double, location:chararray);                                                    // line 1

crimes_data_FBI_code_32 = FILTER crimes_data BY FBICode == '32';    // line 2

crimes_data_group_by_FBI_code = GROUP crimes_data_FBI_code_32 BY FBICode;    // line 3

FBI_code_with_cases = FOREACH crimes_data_group_by_FBI_code GENERATE group, COUNT (crimes_data_FBI_code_32.case_number);                    // line 4

DUMP FBI_code_with_cases;                                                    // line 5
```

Note: I am not storing result into HDFS here since it returns a single key, value pair. I am printing it to the console.

Comments:

Line 1: loads the dataset ‘Crimes_-_2001_to_present.csv’ from specified file system path to the Pig environment. The column separator character here is comma (,). We have to provide the schema (column name and its datatype) for the dataset in this statement. The resulting dataset with the schema will be stored in a relation, in this case ‘crimes_data’.

Line 2: applies filter operation to get only those records with FBI Code being 32. The result will be stored in an intermediate relation ‘crimes_data_FBI_code_32’.

Line 3: groups data in the relation ‘crimes_data_FBI_code_32’ by field ‘FBICode’. The result will be stored in another intermediate relation ‘crimes_data_group_by_FBI_code’.

Line 4: generates a pair of values (FBI Code, <number of cases for this FBI code>).

Line 5: Prints the result on console.

```
acadgild@localhost:~
File  Edit  View  Search  Terminal  Help
[acadgild@localhost ~]$ cat pig_script_for_problem2.pig
crimes_data = LOAD '/user/acadgild/Crimes_-_2001_to_present.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','
NO_MULTILINE','UNIX') AS (id:int, case_number:chararray, date:chararray, block:chararray, IUCR:int, primary_type:chararray, d
escription:chararray, location_description:chararray, arrest:chararray, domestic:chararray, beat:int, district:int, ward:int,
 community_area:int, FBICode:chararray, x_coordinate:int, y_coordinate:int, year:int, updated_on:chararray, latitude:double,
longitude:double, location:chararray);
crimes_data_FBI_code_32 = FILTER crimes_data BY FBICode == '32';
crimes_data_group_by_FBI_code = GROUP crimes_data_FBI_code_32 BY FBICode;
FBI_code_with_cases = FOREACH crimes_data_group_by_FBI_code GENERATE group, COUNT(crimes_data_FBI_code_32.case_number);
DUMP FBI_code_with_cases;
[acadgild@localhost ~]$
```

Output:

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

Job DAG:
job_local1029573782_0003


2018-01-26 02:41:45,002 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-01-26 02:41:45,006 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-01-26 02:41:45,008 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNam
e=JobTracker, sessionId= - already initialized
2018-01-26 02:41:45,017 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encount
ered Warning FIELD_DISCARDED_TYPE_CONVERSION_FAILED 38460 time(s).
2018-01-26 02:41:45,017 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success
!
2018-01-26 02:41:45,068 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
Instead, use dfs.bytes-per-checksum
2018-01-26 02:41:45,069 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2018-01-26 02:41:45,069 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is depre
cated. Instead, use mapreduce.job.counters.max
2018-01-26 02:41:45,069 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2018-01-26 02:41:45,238 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-01-26 02:41:45,238 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(32,76)
```

We can conclude that **there are 76 cases** under FBI code 32.

3. Here is the **pig script** to calculate the number of arrests in theft district wise:

pig_script_for_problem3.pig

crimes_data = LOAD '/user/acadgild/Crimes_-_2001_to_present.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX') AS (id:int, case_number:chararray, date:chararray, block:chararray, IUCR:int, primary_type:chararray, description:chararray, location_description:chararray, arrest:chararray, domestic:chararray, beat:int, district:int, ward:int, community_area:int, FBICode:chararray, x_coordinate:int, y_coordinate:int, year:int, updated_on:chararray, latitude:double, longitude:double, location:chararray); // line 1

crimes_data_for_theft_cases = FILTER crimes_data_filtered BY primary_type == 'THEFT' AND arrest == 'true'; // line 2

crimes_data_grouped_by_district = GROUP crimes_data_for_theft_cases BY district; // line 3

district_with_number_of_arrests = FOREACH crimes_data_grouped_by_district GENERATE group, COUNT (crimes_data_for_theft_cases.arrest); // line 4

STORE district_with_number_of_arrests INTO '/user/acadgild/solution3' USING PigStorage(','); // line 5

Comments:

Line 1: loads the dataset ‘Crimes_-_2001_to_present.csv’ from specified file system path to the Pig environment. The column separator character here is comma (,). We have to provide the schema (column name and its datatype) for the dataset in this statement. The resulting dataset with the schema will be stored in a relation, in this case ‘crimes_data’.

Line 2: applies filter operation to get only those records with primary case type being theft and the arrest field set to true. The result will be stored in an intermediate relation ‘crimes_data_for_theft_cases’.

Line 3: groups data in the relation ‘crimes_data_for_theft_cases’ by district. The result will be stored in another intermediate relation ‘district_with_number_of_arrests’.

Line 4: generates a pair of values (district, <number of arrests for this district>).

Line 5: Writes result to specified path in HDFS.

```
[acadgild@localhost ~]$ cat pig_script_for_problem3.pig
crimes_data = LOAD '/user/acadgild/Crimes_-_2001_to_present.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','
NO_MULTILINE','UNIX') AS (id:int, case_number:chararray, date:chararray, block:chararray, IUCR:int, primary_type:chararray, d
escription:chararray, location_description:chararray, arrest:chararray, domestic:chararray, beat:int, district:int, ward:int,
 community_area:int, FBICode:chararray, x_coordinate:int, y_coordinate:int, year:int, updated_on:chararray, latitude:double,
longitude:double, location:chararray);
crimes_data_for_theft_cases = FILTER crimes_data BY primary_type == 'THEFT' AND arrest == 'true';
crimes_data_grouped_by_district = GROUP crimes_data_for_theft_cases BY district;
district_with_number_of_arrests = FOREACH crimes_data_grouped_by_district GENERATE group, COUNT(crimes_data_for_theft_cases.a
rrest);
STORE district_with_number_of_arrests INTO '/user/acadgild/solution3' USING PigStorage(',');
[acadgild@localhost ~]$
```

Output:

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.2.0   0.14.0  acadgild        2018-01-26 06:19:42     2018-01-26 06:21:50     GROUP_BY,FILTER

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgRe
duceTime        MedianReducetime        Alias   Feature Outputs
job_1516884154891_0007  1       1       24      24      24      24      18      18      18      18      crimes_data,crimes_da
ta_for_theft_cases,crimes_data_grouped_by_district,district_with_number_of_arrests      GROUP_BY,COMBINER       /user/acadgil
d/solution3,

Input(s):
Successfully read 0 records from: "/user/acadgild/Crimes_-_2001_to_present.csv"

Output(s):
Successfully stored 0 records in: "/user/acadgild/solution3"

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
```

```
[acadgild@localhost ~]$ hadoop fs -ls /user/acadgild/solution3
18/01/26 06:22:34 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2018-01-26 06:21 /user/acadgild/solution3/_SUCCESS
-rw-r--r--   1 acadgild supergroup        146 2018-01-26 06:21 /user/acadgild/solution3/part-r-00000
[acadgild@localhost ~]$ hadoop fs -cat /user/acadgild/solution3/part-r-00000
18/01/26 06:23:09 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
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
[acadgild@localhost ~]$
```

4. Here is the **pig script** to calculate the number of arrests done between October 2014 and October 2015.

pig_script_for_problem4.pig

```
crimes_data = LOAD '/user/acadgild/Crimes_-_2001_to_present.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX') AS (id:int, case_number:chararray, date:chararray, block:chararray, IUCR:int, primary_type:chararray, description:chararray, location_description:chararray, arrest:chararray, domestic:chararray, beat:int, district:int, ward:int, community_area:int, FBICode:chararray, x_coordinate:int, y_coordinate:int, year:int, updated_on:chararray, latitude:double, longitude:double, location:chararray);                                                    // line 1

crimes_data_filtered = FILTER crimes_data BY arrest == 'true' AND ((ToDate(updated_on,'MM/dd/yyyy hh:mm:ss aa') >= ToDate('10/01/2014','MM/dd/yyyy')) AND (ToDate(updated_on,'MM/dd/yyyy hh:mm:ss aa') <= ToDate('10/31/2015','MM/dd/yyyy')));
                                                                                    // line 2
crimes_data_grouped_by_all = GROUP crimes_data_filtered ALL;          // line 3

number_of_arrests = FOREACH crimes_data_grouped_by_all GENERATE COUNT (crimes_data_filtered.arrest);                                                    // line 4

STORE number_of_arrests into '/user/acadgild/solution4' USING PigStorage(',');  // line 5
```

Comments:

Line 1: loads the dataset ‘Crimes_-_2001_to_present.csv’ from specified file system path to the Pig environment. The column separator character here is comma (,). We have to provide the schema (column name and its datatype) for the dataset in this statement. The resulting dataset with the schema will be stored in a relation, in this case ‘crimes_data’.

Line 2: applies filter operation to get only those records with ‘arrest’ field set to true and the date falls between October 01, 1024 and October 31, 2015. The result will be stored in an intermediate relation ‘crimes_data_filtered’.

Line 3: groups data in the relation ‘crimes_data_filtered’ by all fields, which means each row is formed as a group here. The result will be stored in another intermediate relation ‘crimes_data_grouped_by_all’.

Line 4: generates number of arrests by counting all the rows from ‘crimes_data_grouped_by_all’.

Line 5: Writes result to specified path in HDFS.

```
[acadgild@localhost ~]$ cat pig_script_for_problem4.pig
crimes_data = LOAD '/user/acadgild/Crimes_-_2001_to_present.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','
NO_MULTILINE','UNIX') AS (id:int, case_number:chararray, date:chararray, block:chararray, IUCR:int, primary_type:chararray, d
escription:chararray, location_description:chararray, arrest:chararray, domestic:chararray, beat:int, district:int, ward:int,
 community_area:int, FBICode:chararray, x_coordinate:int, y_coordinate:int, year:int, updated_on:chararray, latitude:double,
longitude:double, location:chararray);

crimes_data_filtered = FILTER crimes_data BY arrest == 'true' AND ((ToDate(updated_on,'MM/dd/yyyy hh:mm:ss aa') >= ToDate('10
/01/2014','MM/dd/yyyy')) AND (ToDate(updated_on,'MM/dd/yyyy hh:mm:ss aa') <= ToDate('10/31/2015','MM/dd/yyyy')));

crimes_data_grouped_by_all = GROUP crimes_data_filtered ALL;

number_of_arrests = FOREACH crimes_data_grouped_by_all GENERATE COUNT(crimes_data_filtered.arrest);

STORE number_of_arrests into '/user/acadgild/solution4' USING PigStorage(',');
[acadgild@localhost ~]$
```

Output:

```
2018-01-26 08:56:11,293 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% co
mplete
2018-01-26 08:56:11,304 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.2.0   0.14.0  acadgild        2018-01-26 08:52:25     2018-01-26 08:56:11     GROUP_BY,FILTER

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgRe
duceTime        MedianReducetime        Alias   Feature Outputs
job_1516884154891_0011  1       1       33      33      33      33      14      14      14      14      crimes_data,crimes_da
ta_filtered,crimes_data_grouped_by_all,number_of_arrests        GROUP_BY,COMBINER       /user/acadgild/solution4,

Input(s):
Successfully read 0 records from: "/user/acadgild/Crimes_-_2001_to_present.csv"

Output(s):
Successfully stored 0 records in: "/user/acadgild/solution4"

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_1516884154891_0011


[acadgild@localhost ~]$ hadoop fs -ls /user/acadgild/solution4
18/01/26 08:57:12 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2018-01-26 08:55 /user/acadgild/solution4/_SUCCESS
-rw-r--r--   1 acadgild supergroup          6 2018-01-26 08:55 /user/acadgild/solution4/part-r-00000
[acadgild@localhost ~]$ hadoop fs -cat /user/acadgild/solution4/part-r-00000
18/01/26 08:57:26 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
68258
[acadgild@localhost ~]$
```

Conclusion: The number of arrests done between October 2014 and October 2015 are **68,258**.