Project 1 – USA Crime Data Analysis

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

FBI code

```
REGISTER '/home/maria_dev/acadgild/assignments/projects/project1/piggybank-0.17.0.jar' ;
-- Load Crime data from csv

crime_data = LOAD '/home/maria_dev/acadgild/assignments/projects/project1/Crimes2001_to_present.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULT
ILINE','UNIX','SKIP_INPUT_HEADER');

-- select the data columns on which we have to work
qualifying_data = foreach crime_data generate (chararray)$0 as caseid, (chararray)$14 as fbicode;

-- group the data by fbicode
group_qualifying_data = GROUP qualifying_data by fbicode ;
--dump group_qualifying_data;

-- get count
count_qualifying_data = foreach group_qualifying_data GENERATE group,COUNT(qualifying_data) ;

-- display output
dump count_qualifying_data ;
```

pig -x local project1_1.pig

```
time(s).
2017-09-19 06:57:58,617 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-09-19 06:57:58,622 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-09-19 06:57:58,669 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-09-19 06:57:58,670 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(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,7710)
(08A,14167)
(08B,46938)
(,1)
2017-09-19 06:57:58,751 [main] INFO  org.apache.pig.Main - Pig script completed in 16 seconds and 223 milliseconds (16223 ms)
[maria_dev@sandbox project1]$
```

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

code 32.

```
--##################################################
-- GET THE COUNT OF ALL THE CASES SOLVED BY FBI CODE 32
--##################################################

REGISTER '/home/maria_dev/acadgild/assignments/projects/project1/piggybank-0.17.0.jar' ;

-- Load Crime data from csv

crime_data = LOAD '/home/maria_dev/acadgild/assignments/projects/project1/Crimes2001_to_present.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULT
ILINE','UNIX','SKIP_INPUT_HEADER');

-- select the data columns on which we have to work
dataset = foreach crime_data generate (chararray)$14 as fbicode;

-- filter dataset
elg_dataset = filter dataset  by fbicode == '32' ;

elg_dataset_groupall = group elg_dataset ALL ;

-- get count
count_elg_dataset = foreach elg_dataset_groupall GENERATE COUNT(elg_dataset.fbicode) ;

-- display output
dump count_elg_dataset ;
```

pig -x local project1_2.pig

```
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_local1665927608_0001


2017-09-19 06:59:38,458 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initiali
zed
2017-09-19 06:59:38,460 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initiali
zed
2017-09-19 06:59:38,463 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initiali
zed
2017-09-19 06:59:38,482 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning ACCESSING_NON_EXISTENT_FIELD 1
 time(s).
2017-09-19 06:59:38,482 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-09-19 06:59:38,487 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-09-19 06:59:38,522 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-09-19 06:59:38,523 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2017-09-19 06:59:38,577 [main] INFO  org.apache.pig.Main - Pig script completed in 8 seconds and 224 milliseconds (8224 ms)
[maria_dev@sandbox project1]$
```

There is no data satisfying the criteria.

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

```
--##################################################
-- GET DISTRICT WISE COUNT OF ARRESTS DUE TO THEFT
--##################################################

REGISTER '/home/maria_dev/acadgild/assignments/projects/project1/piggybank-0.17.0.jar' ;
-- Load Crime data from csv

crime_data = LOAD '/home/maria_dev/acadgild/assignments/projects/project1/Crimes2001_to_present.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULT
ILINE','UNIX','SKIP_INPUT_HEADER');

-- select the data columns on which we have to work

qualifying_data = foreach crime_data generate (chararray)$5 as crimetype, (chararray)$8 as arrests, (chararray)$11 as district ;

--filter data
filter_data = filter qualifying_data by crimetype == 'THEFT' AND arrests == 'true' ;
--dump filter_data ;

-- group the data by fbicode
group_qualifying_data = GROUP filter_data by district ;

-- get count
count_qualifying_data = foreach group_qualifying_data GENERATE group,COUNT(filter_data) ;

-- display output
dump count_qualifying_data ;
```

pig -x local project1_3.pig

```
 time(s).
2017-09-19 07:01:12,425 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-09-19 07:01:12,430 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-09-19 07:01:12,464 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-09-19 07:01:12,464 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process
(001,1124)
(002,227)
(003,162)
(004,230)
(005,286)
(006,652)
(007,176)
(008,471)
(009,320)
(010,170)
(011,178)
(012,360)
(014,228)
(015,115)
(016,177)
(017,237)
(018,734)
(019,501)
(020,244)
(022,220)
(024,226)
(025,596)
2017-09-19 07:01:12,530 [main] INFO  org.apache.pig.Main - Pig script completed in 8 seconds and 457 milliseconds (8457 ms)
[maria_dev@sandbox project1]$
```

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

maria_dev@sandbox:~/acadgild/assignments/projects/project1

```
--################################################
-- GET count of arrests between Oct 2014 and Oct 2015
--################################################

REGISTER '/home/maria_dev/acadgild/assignments/projects/project1/piggybank-0.17.0.jar' ;
-- Load Crime data from csv

crime_data = LOAD '/home/maria_dev/acadgild/assignments/projects/project1/Crimes2001_to_present.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULT
ILINE','UNIX','SKIP_INPUT_HEADER') ;

-- select the data columns on which we have to work

qualifying_data = foreach crime_data generate ToDate($2,'M/d/yyyy HH:mm:ss a') as (casedate:DateTime), (chararray)$8 as arrests ;

--filter data



filter_data = filter qualifying_data by arrests == 'true' AND casedate >= ToDate('10/01/2014','M/d/yyyy') AND casedate <= ToDate('10/31/2015','M/d/yyyy') ;
--dump filter_data ;

-- group the data by fbicode
group_qualifying_data = GROUP filter_data ALL ;

-- get count
count_qualifying_data = foreach group_qualifying_data GENERATE COUNT(filter_data) ;

-- display output
dump count_qualifying_data ;
```

```
2017-09-19 07:05:33,713 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initiali
zed
2017-09-19 07:05:33,715 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initiali
zed
2017-09-19 07:05:33,717 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initiali
zed
2017-09-19 07:05:33,726 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning ACCESSING_NON_EXISTENT_FIELD 2
 time(s).
2017-09-19 07:05:33,726 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-09-19 07:05:33,735 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-09-19 07:05:33,848 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-09-19 07:05:33,848 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(65027)
2017-09-19 07:05:33,928 [main] INFO  org.apache.pig.Main - Pig script completed in 12 seconds and 186 milliseconds (12186 ms)
[maria_dev@sandbox project1]$
```