Project 1.1

|  |  |
| --- | --- |
| **Prepared For:** | AcadGild |
|  |  |
| **Document Approval:** | **AcadGild** |
|  |  |
|  |  |
|  |  |
|  |  |
| **Project Title:** | Project 1.1 |
|  |  |
| **Prepared By:** | Duncan Burgess |
|  |  |
|  | dburgess@duncb.com |
|  |  |
| **Primary Engineer:** | Duncan Burgess |
|  |  |
| **Document Reference:** | **Project 1.1** |
|  |  |
| **Start Date:** | 06/10/2017 |
|  |  |
|  |  |



# 

# Contents

[Contents 2](#_Toc495056294)

[Change History 3](#_Toc495056295)

[1. Problem Statement 4](#_Toc495056296)

[1.1. Dataset 5](#_Toc495056297)

[2. Setting environment 6](#_Toc495056298)

[3. Solutions 7](#_Toc495056299)

[3.1. Write a MapReduce/Pig program to calculate the number of cases investigated under each FBI code 7](#_Toc495056300)

[3.2. Write a MapReduce/Pig program to calculate the number of cases investigated under FBI code 32. 8](#_Toc495056301)

[3.3. Write a MapReduce/Pig program to calculate the number of arrests in theft district wise. 9](#_Toc495056302)

[3.4. Write a MapReduce/Pig program to calculate the number of arrests done between October 2014 and October 2015 10](#_Toc495056303)

# Change History

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Document Revision** | **Date** | **Authored By** | **Authorised By** | **Sections Affected** | **Reason for Change** |
| Rev 01 | 06/10/2017 | Duncan Burgess |  | All | Initial release. |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

# 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.

## Dataset

The dataset Crimes\_-\_2001\_to\_present.csv renamed to Crimes.csv

**Dataset Description:**

|  |  |
| --- | --- |
| Column Number | Column Description |
| 0 | ID |
| 1 | Case\_Number |
| 2 | Date |
| 3 | Block |
| 4 | IUCR |
| 5 | Primary\_Type |
| 6 | Description |
| 7 | Location\_Description |
| 8 | Arrest |
| 9 | Domestic |
| 10 | Beat |
| 11 | District |
| 12 | Ward |
| 13 | Community\_Area |
| 14 | FBICode |
| 15 | X\_Coordinate |
| 16 | Y\_Coordinate |
| 17 | Year |
| 18 | Updated\_On |
| 19 | Latitude |
| 20 | Longitude |
| 21 | Location |

**Sample of Data**

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)"

10231208,HY418843,09/10/2015 11:50:00 PM,021XX W BERWYN AVE,0820,THEFT,$500 AND UNDER,STREET,false,false,2012,020,40,4,06,1161036,1935171,2015,09/17/2015 11:37:18 AM,41.97779966,-87.683164484,"(41.97779966, -87.683164484)"

10230943,HY418702,09/10/2015 11:45:00 PM,009XX N DRAKE AVE,0486,BATTERY,DOMESTIC BATTERY SIMPLE,APARTMENT,true,true,1121,011,27,23,08B,1152539,1906092,2015,09/17/2015 11:37:18 AM,41.898177341,-87.71518334,"(41.898177341, -87.71518334)"

10230974,HY418690,09/10/2015 11:35:00 PM,038XX W HARRISON ST,0337,ROBBERY,ATTEMPT: ARMED-OTHER DANG WEAP,STREET,false,false,1133,011,24,26,03,1151141,1897093,2015,09/17/2015 11:37:18 AM,41.873510659,-87.720554136,"(41.873510659, -87.720554136)"

10231069,HY418804,09/10/2015 11:30:00 PM,016XX W 79TH ST,0890,THEFT,FROM BUILDING,RESIDENCE,false,false,0611,006,21,71,06,1166608,1852321,2015,09/17/2015 11:37:18 AM,41.75033427,-87.665045312,"(41.75033427, -87.665045312)"

10230986,HY418698,09/10/2015 11:30:00 PM,016XX E 67TH ST,0560,ASSAULT,SIMPLE,SIDEWALK,false,false,0331,003,5,42,08A,1188443,1860888,2015,09/17/2015 11:37:18 AM,41.773349622,-87.584759546,"(41.773349622, -87.584759546)"

10233462,HY421628,09/10/2015 11:30:00 PM,049XX W GLADYS AVE,1154,DECEPTIVE PRACTICE,FINANCIAL IDENTITY THEFT $300 AND UNDER,RESIDENCE,false,false,1533,015,24,25,11,1143652,1897891,2015,09/17/2015 11:37:18 AM,41.875843892,-87.748030348,"(41.875843892, -87.748030348)"

10231724,HY419244,09/10/2015 11:30:00 PM,002XX S LOTUS AVE,1320,CRIMINAL DAMAGE,TO VEHICLE,STREET,false,false,1522,015,29,25,14,1139915,1898582,2015,09/17/2015 11:37:18 AM,41.877809197,-87.761734553,"(41.877809197, -87.761734553)"

# Setting environment

Dataset stored in '/home/cloudera/Datasets/crime.csv'

Jar file stored in '/home/cloudera/lib/piggybank.jar'

To begin the entire required problem the first step is to register the piggybank.jar

**REGISTER '/home/cloudera/lib/piggybank.jar';**

Then define the storage method CSVExcelStorage to use class from Jar

**DEFINE CSVExcelStorage org.apache.pig.piggybank.storage.CSVExcelStorage();** to allow us to import CSV formatted file.

**For all the problems we upload the dataset Crimes.csv**

CrimeData = LOAD '/home/cloudera/Datasets/Crimes.csv' USING CSVExcelStorage(',','NO\_MULTILINE','UNIX');

I have not defined the column types as this will be performed during loading the required data, therefore CrimeData has no schema.

*grunt> describe CrimeData;*

*Schema for CrimeData unknown.*

**Note** : For all the solutions the output can be stored if required in a file using “store” command.

# Solutions

**The following are the four required problems.**

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

**Code created**

*REGISTER '/home/cloudera/lib/piggybank.jar';*

*CrimeData = LOAD '/home/cloudera/Datasets/Crimes.csv' USING CSVExcelStorage(',','NO\_MULTILINE','UNIX');*

*A = FOREACH CrimeData GENERATE (chararray)$1 AS caseNum, (chararray)$4 AS iucr, (chararray)$14 AS fbiCode;*

*B = GROUP A BY fbiCode;*

*NumberByFBICode = FOREACH B GENERATE group, COUNT(B.caseNum);*

*DUMP NumberByFBICode;*

**Results**

*(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)*

*(,0)*

**Explanation**

1. Loaded data from crimes.csv into “CrimeData” relation. Here ‘,’ is the delimiter and schema is provided as mentioned in the question
2. Then, all records are grouped on basis of FBICode.
3. For each FBICode (group-key) calculating the values – number of cases investigated under that particular FBICode using COUNT function.
4. Outputting only “FBICode and number of cases investigated under that FBICode” as per Question using dump command.

Output displayed: <FBICode> <number\_of\_cases>

**Note**  There is one invalid (last) entry (only ID = 9 it has) in the dataset with no FBICode – the last ( ,1) corresponds to that which can be neglected .

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

**Code created**

*REGISTER '/home/cloudera/lib/piggybank.jar';*

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

*CrimeData = LOAD '/home/cloudera/Datasets/Crimes.csv' USING CSVExcelStorage(',','NO\_MULTILINE','UNIX');*

*A = FOREACH CrimeData GENERATE (chararray)$1 AS caseNum, (chararray)$4 AS iucr, (chararray)$12 AS Ward;*

*B = FILTER A BY Ward=='32';*

*C = GROUP B BY Ward;*

*CasesWard32 = FOREACH C GENERATE group, COUNT(B.caseNum);*

*dump CasesWard32;*

**Result**

*(32,4592)*

**Explanation**

1. Loaded data from crime.csv into “CrimeData” relation. Here ‘,’ is the delimiter and schema is provided as mentioned in the question
2. Then, all records are filtered on the basis of expression Ward ==’32’.
3. Now, for counting the number of cases investigated under Ward 32, we group all the above filtered records and then using COUNT function, we calculate the number of cases investigated under Ward 32.
4. Outputting only and the number of cases investigated under Ward 32” as per Question using dump command.

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

**Code created**

*REGISTER '/home/cloudera/lib/piggybank.jar';*

*DEFINE CSVExcelStorage org.apache.pig.piggybank.storage.CSVExcelStorage;*

*CrimeData = LOAD '/home/cloudera/Datasets/Crimes.csv' USING CSVExcelStorage(',','NO\_MULTILINE','UNIX');*

*A = FOREACH CrimeData GENERATE (chararray)$1 AS caseNum, (chararray)$5 AS iucr, (boolean)$8 AS arrest, (chararray)$11 AS dist;*

*B = FILTER A BY (arrest==true) AND (iucr=='THEFT' OR iucr=='MOTOR VEHICLE THEFT');*

*C = GROUP B BY dist;*

*TheftPerDistrict = FOREACH C GENERATE group, COUNT(B.caseNum);*

*dump TheftPerDistrict;*

**Results**

*(001,1137)*

*(002,289)*

*(003,217)*

*(004,269)*

*(005,318)*

*(006,699)*

*(007,232)*

*(008,561)*

*(009,377)*

*(010,229)*

*(011,253)*

*(012,405)*

*(014,255)*

*(015,168)*

*(016,203)*

*(017,256)*

*(018,753)*

*(019,524)*

*(020,256)*

*(022,252)*

*(024,240)*

*(025,667)*

**Explanation**

1. Loaded data from crime.csv into “CrimeData”
2. All the records are filtered based on the conditions – PrimaryType == **THEFT** or **MOTOR VEHICLE THEFT** as they are both theft’s and Arrest==’true’. This gives us all theft cases in which arrest have happened.
3. Now, grouping the theft\_arrest\_cases with respect to District.
4. Then, foreach key in the group, we calculated the number of values – theft\_arrest\_cases that happened corresponding to that District using COUNT function.
5. Outputting the “District and corresponding number of theft arrest cases” using dump command.

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

**Code created**

*REGISTER '/home/acadgild/project/piggybank.jar';*

*DEFINE CSVExcelStorage org.apache.pig.piggybank.storage.CSVExcelStorage;*

*CrimeData = LOAD '/user/acadgild/project/crimes.csv' USING CSVExcelStorage(',','NO\_MULTILINE','UNIX');*

*A = FOREACH CrimeData GENERATE (chararray)$2 AS Date, (boolean)$8 AS arrest, (int)$0 as ID;*

*B = foreach A generate ToDate(Date, 'MM/dd/yyyy HH:mm:ss a', 'Europe/London') AS Date\_Time, arrest, ID;*

*C = filter B By (arrest==true) AND (DaysBetween(Date\_Time, (datetime)ToDate('10/01/2014', 'MM/dd/yyyy')) >(long)0) AND (DaysBetween(Date\_Time, (datetime)ToDate('10/01/2015', 'MM/dd/yyyy')) <(long)0);*

*D = GROUP C BY arrest;*

*ArrestsOct14toOct15 = FOREACH D GENERATE COUNT(C.arrest);*

*DUMP Arrests;*

**Result**

*(64675)*

**Explanation**

Loaded data from crime.csv into “CrimeData” relation.

For performing operations on Date, a DateTime object is required. So, with the help of ToDate() function we generated a DateTime object from the input uscrime. Here, since dataset is from USA\_Crimes, I have specified local timezone to be “Europe/London” and input date format is (MM/dd/yyyy HH:mm:ss a) observed from dataset.

Now, then filter the records based on Arrest==’true’ condition and using the DaysBetween() function which accepts two date-time objects(dt1 and dt2) and calculates the number of days(positive days if dt1 occurred after dt2 and negative days if dt1 occurred before dt2 ) between the two given date-time objects, It has been filtered such that the arrest cases are between 1st October 2014 and 1st October 2015 The Date should be after '10/01/2014' and before ‘10/01/2015’.

Now, for counting the number of arrests happened, we group all the above filtered records and then using COUNT function, we calculate the number of arrests done in between October 2014 and October 2015.

Outputting the total number of arrest cases using the dump command.