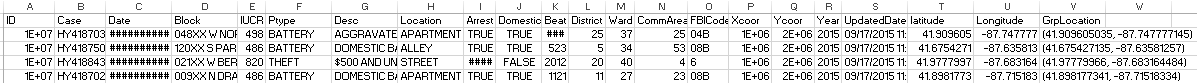
**Project 1: Crime Analysis**

Associated Data Files

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

Dataset Description:

ID,CaseNumber,Date,Block,IUCR,PrimaryType,Description,LocationDescription,Arrest,Domestic,Beat,District,Ward,Community Area,FBICode,X Coordinate,Y Coordinate,Year,Updated On,Latitude,Longitude,Location as GeoLocation



**pig -x local**

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

Note: FBICode is => $14

Data validation and cleansing process:

**Problem Statement**

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

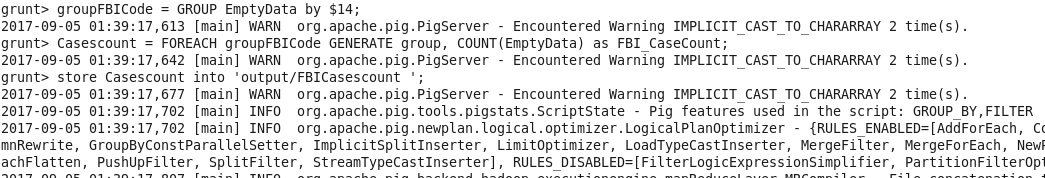
**Solution:**

CA = LOAD '/home/cloudera/Crimes\_Analysis.csv' USING PigStorage(',') as (ID:int,CaseNumber:chararray,Date:chararray,Block:chararray,IUCR:int,PrimaryType:chararray,Description:chararray,LocationDescription:chararray,Arrest:chararray,Domestic:boolean,Beat:int,District:int,Ward:int,CommunityArea:chararray,FBICode:chararray,XCoordinate:int,YCoordinate,Year:chararray,UpdatedOn:chararray,Latitude:double,Longitude:double,Location:chararray) ; --org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO\_MULTILINE')

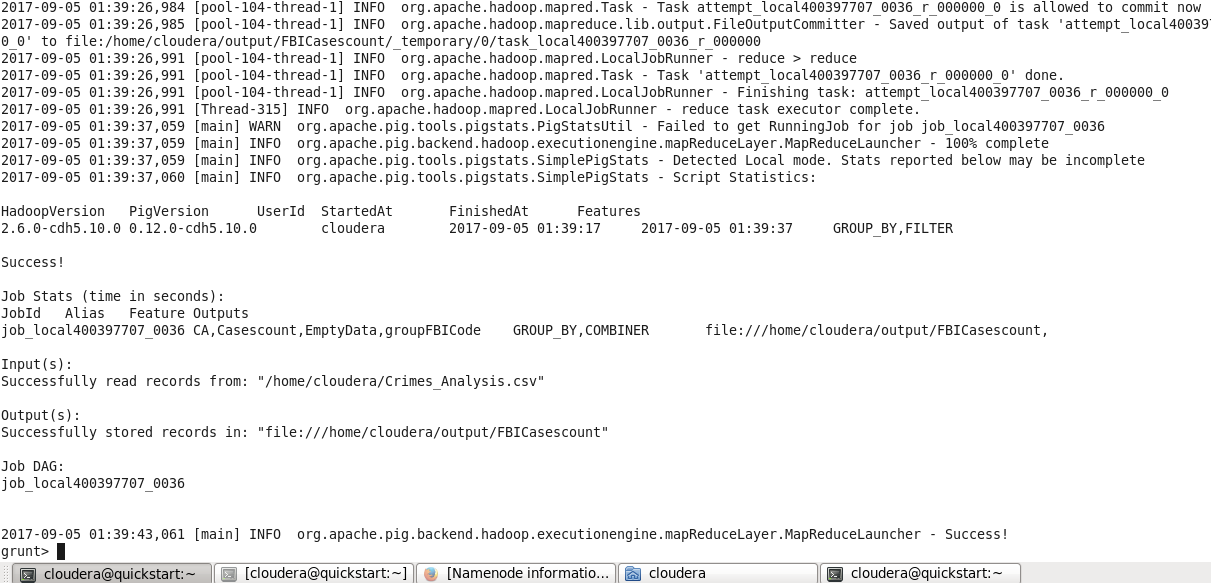
* EmptyData = FILTER CA BY $14 is not null;
* groupFBICode = GROUP EmptyData by $14;
* Casescount = FOREACH groupFBICode GENERATE group, COUNT(EmptyData) as FBI\_CaseCount;
* store Casescount into 'output/FBICasescount ';

or

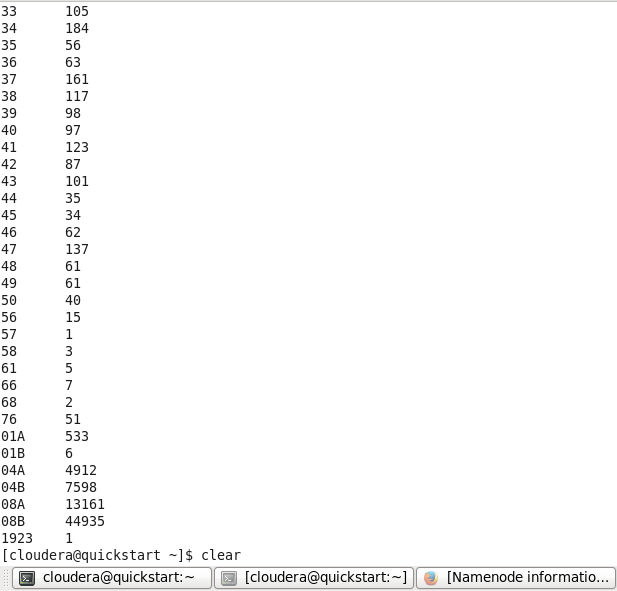
* DUMP Casescount;

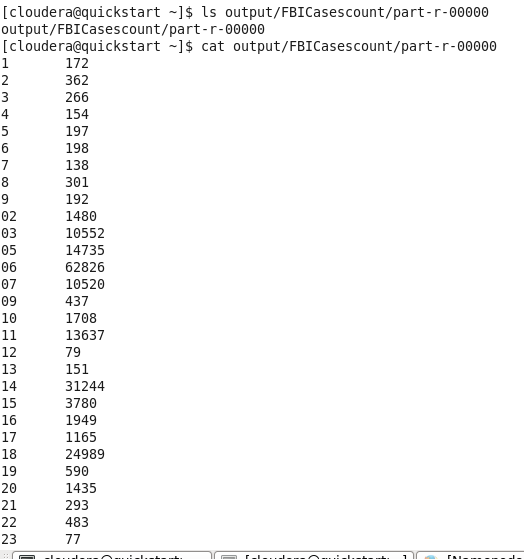


**Job Successfully executed:**



Output:



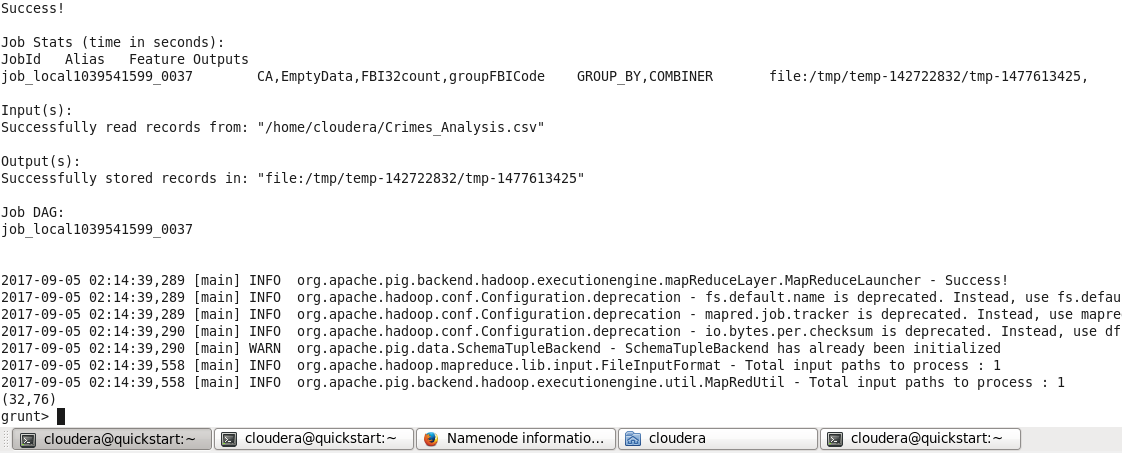


**2. Write a MapReduce/Pig program to calculate the number of cases investigated under FBI code 32.**

**Solution:**

* EmptyData = FILTER CA BY $14 is not null;
* FBI\_32 = FILTER EmptyData BY ($14=='32');
* groupFBICode = GROUP FBI\_32 by $14;
* FBI32count = FOREACH groupFBICode GENERATE group, COUNT(FBI\_32) as FBI\_32Count;
* DUMP FBI32count;

**Ans : Code 32 =76 cases**



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

**Arrest $8**

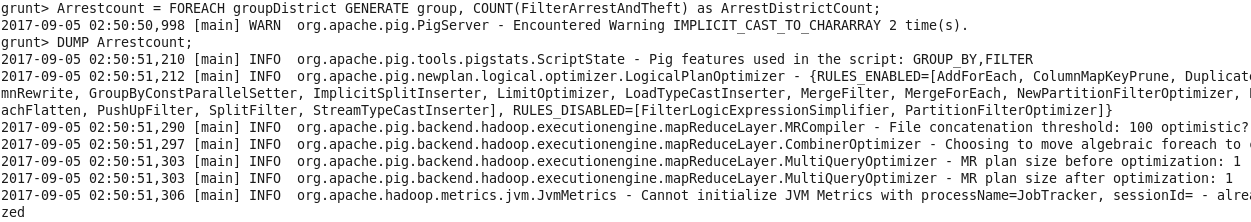
**Theft $5**

**District $11**

**Solution:**

* FilterArrest = FILTER CA BY $8 is not null;
* FilterArrestAndTheft = FILTER FilterArrest BY ($8 == 'true' and $5 == 'THEFT');
* groupDistrict = GROUP FilterArrestAndTheft by $11;
* Arrestcount = FOREACH groupDistrict GENERATE group, COUNT(FilterArrestAndTheft) as ArrestDistrictCount;
* DUMP Arrestcount;

**Ans:**



**Arrest Count District Wise:**

**(1,1119)**

**(2,220)**

**(3,157)**

**(4,221)**

**(5,273)**

**(6,649)**

**(7,172)**

**(8,458)**

**(9,318)**

**(10,166)**

**(11,174)**

**(12,353)**

**(14,227)**

**(15,111)**

**(16,171)**

**(17,227)**

**(18,732)**

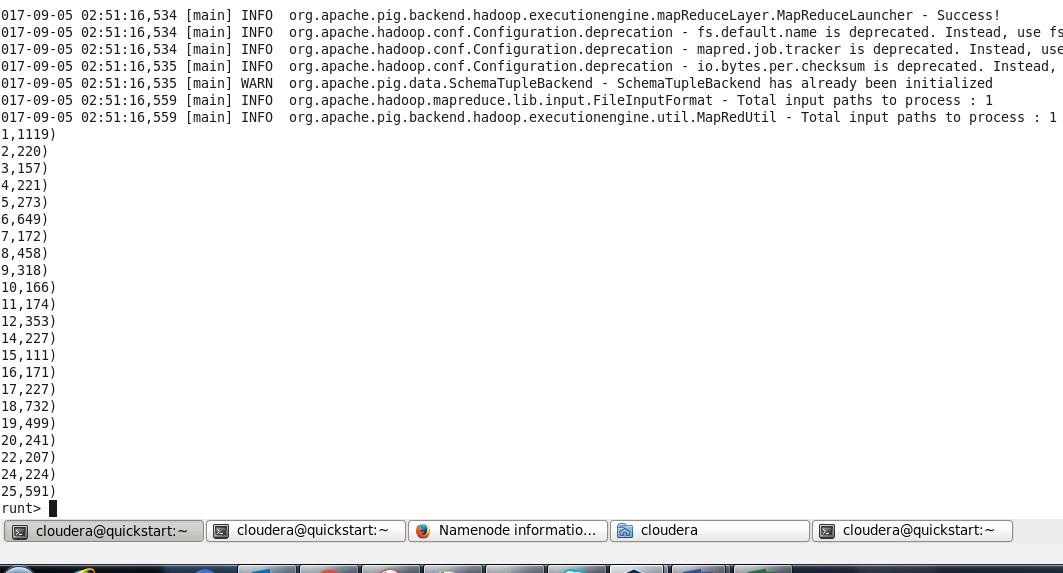
**(19,499)**

**(20,241)**

**(22,207)**

**(24,224)**

**(25,591)**



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

**Solution:**

CA = LOAD '/home/cloudera/Crimes\_Analysis.csv' USING PigStorage(',') as (ID:int,CaseNumber:chararray,Date:chararray,Block:chararray,IUCR:int,PrimaryType:chararray,Description:chararray,LocationDescription:chararray,Arrest:chararray,Domestic:boolean,Beat:int,District:int,Ward:int,CommunityArea:chararray,FBICode:chararray,XCoordinate:int,YCoordinate,Year:chararray,UpdatedOn:chararray,Latitude:double,Longitude:double,Location:chararray) ; --org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO\_MULTILINE')

* GetArrest = FILTER CA BY ($8 != '' or $8 is not null) and $8 == 'true';
* DateFormat = FOREACH GetArrest GENERATE $8 as Arrest, CONCAT((CHARARRAY) SUBSTRING($2,6,10), CONCAT('-',CONCAT((CHARARRAY) SUBSTRING($2,0,2), CONCAT('-', (CHARARRAY) SUBSTRING($2,3,5))))) as Date;
* DataRange = FILTER DateFormat BY (Date >='2014-10-01' and Date <='2015-10-31');
* GetArrestGroup = GROUP DataRange by Arrest;
* countArrest = FOREACH GetArrestGroup GENERATE group, COUNT(DataRange) as ArrestCount;
* DUMP countArrest;

ANS : 63173 Arrest cases during



