Project 1.1

crimes\_data = LOAD '/home/acadgild/Crimes.csv' USING PigStorage(',')

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:chararray,

Longitude:chararray,

Location:chararray

);

grunt> describe crimes\_data;

crimes\_data: {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: chararray,Longitude: chararray,Location: chararray}

Problem Statement

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

FBI code

**grunt> B = FOREACH crimes\_data GENERATE Case\_Number,FBICode;**

**grunt>C = GROUP B BY FBICode;**

**grunt> case\_count = FOREACH C GENERATE group,COUNT(B.FBICode);**

**grunt> dump case\_count;**

Input(s):

Successfully read 291268 records from: "/home/acadgild/Crimes.csv"

Output(s):

Successfully stored 71 records in: "file:/tmp/temp-502653654/tmp-526408957"

Counters:

Total records written : 71

Total bytes written : 0

Spillable Memory Manager spill count : 0

Total bags proactively spilled: 0

Total records proactively spilled: 0

(1,172)

(2,362)

(3,266)

(4,154)

(5,197)

(6,198)

(7,138)

(8,301)

(9,192)

(02,1480)

(03,10552)

(05,14735)

(06,62826)

(07,10520)

(09,437)

(10,1708)

(11,13637)

(12,79)

(13,151)

(14,31244)

(15,3780)

(16,1949)

(17,1165)

(18,24989)

(19,590)

(20,1435)

(21,293)

(22,483)

(23,77)

(24,4114)

(25,142)

(26,29009)

(27,175)

(28,385)

(29,196)

(30,115)

(31,93)

(32,76)

(33,105)

(34,184)

(35,56)

(36,63)

(37,161)

(38,117)

(39,98)

(40,97)

(41,123)

(42,87)

(43,101)

(44,35)

(45,34)

(46,62)

(47,137)

(48,61)

(49,61)

(50,40)

(56,15)

(57,1)

(58,3)

(61,5)

(66,7)

(68,2)

(76,51)

(01A,533)

(01B,6)

(04A,4912)

(04B,7598)

(08A,13161)

(08B,44935)

(1923,1)

(,0)

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

code 32.

**grunt> D = FILTER B BY FBICode == '32';**

grunt> describe D;

D: {Case\_Number: chararray,FBICode: chararray}

**grunt> E = GROUP D BY FBICode;**

**grunt> case\_count\_32 = FOREACH E GENERATE group,COUNT(D.FBICode);**

**grunt>DUMP case\_count\_32**

Input(s):

Successfully read 291268 records from: "/home/acadgild/Crimes.csv"

Output(s):

Successfully stored 1 records in: "file:/tmp/temp-502653654/tmp-1704657042"

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

(32,76)

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

**grunt> F = FILTER crimes\_data BY Primary\_Type == 'THEFT';**

**grunt> G = FOREACH F GENERATE Arrest,District;**

**grunt> K = FILTER G BY Arrest == 'true';**

**grunt> I = GROUP K BY District;**

**grunt> arrests\_per\_District = FOREACH I GENERATE group,COUNT(K.District);**

**grunt>DUMP arrests\_per\_District;**

Input(s):

Successfully read 291268 records from: "/home/acadgild/Crimes.csv"

Output(s):

Successfully stored 22 records in: "file:/tmp/temp-502653654/tmp-983251727"

Counters:

Total records written : 22

Total bytes written : 0

Spillable Memory Manager spill count : 0

Total bags proactively spilled: 0

Total records proactively spilled: 0

Job

DAG:

job\_local110395704\_0022

(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.

**grunt> yeardata = FOREACH crimes\_data GENERATE Primary\_Type,Arrest,SUBSTRING(Date,0,2) as Month;**

**grunt> W = FILTER yeardata BY Primary\_Type == 'THEFT' and Month == '10' AND Arrest == 'true';**

**grunt> O = GROUP W ALL;**

**grunt> No\_of\_arrests = FOREACH O GENERATE COUNT(W.Primary\_Type);**

**grunt> dump No\_of\_arrests;**

Input(s):

Successfully read 291268 records from: "/home/acadgild/Crimes.csv"

Output(s):

Successfully stored 1 records in: "file:/tmp/temp-502653654/tmp522732013"

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\_local265893963\_0048

2017-11-04 22:26:10,349 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized

2017-11-04 22:26:10,361 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized

2017-11-04 22:26:10,363 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized

2017-11-04 22:26:10,366 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!

2017-11-04 22:26:10,366 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum

2017-11-04 22:26:10,367 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS

2017-11-04 22:26:10,367 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapreduce.job.counters.limit is deprecated. Instead, use mapreduce.job.counters.max

2017-11-04 22:26:10,367 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized

2017-11-04 22:26:10,416 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1

2017-11-04 22:26:10,416 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1

(649)