CIS 5200 Term Project

Group 1

Abstract

Scope of Data Set:

This term project uses three data sets (Approx total file size:1.1GB) . The data sets are separately composed of vehicle and pedestrian stop data, arrest data, and crime data collected in Los Angeles from 2010 to present this month. The data sets were collected from government open data at the DataLA link and will include (1.482GB) :

<https://data.lacity.org/browse?tags=lapd>

<https://data.lacity.org/A-Safe-City/Vehicle-and-Pedestrian-Stop-Data-2010-to-Present/ci25-wgt7>

<https://data.lacity.org/A-Safe-City/Arrest-Data-from-2010-to-Present/yru6-6re4>

<https://data.lacity.org/A-Safe-City/Crime-Data-from-2010-to-Present/y8tr-7khq>

<https://controllerdata.lacity.org/Statistics/Demographic-and-Economic-Statistics/9z75-k9sy>

<https://data.lacity.org/dataset/2010-Census-Populations-by-Zip-Code/nxs9-385f>

<https://controllerdata.lacity.org/Statistics/Economy-Panel-LA-Data/4ndk-mmc8>

In 2014, Mayor Eric Garcetti launched DataLA portal through the Office of the Mayor’s Data team led by Chief Data Officer Lilian Coral to make the city more livable and equitable to all Angelenos by providing high value open data.

The term project hopes to accomplish things in terms of public safety and awareness, etc. Other topics are open transparency, accountability, address civic engagement, maximize public benefit, protect the public, ensure the integrity of services, and promote resilience.

Approaches:

Through the data compiled in this term paper, we will possibly follow similar goals and approaches done by the DataLA directive using Apache Hadoop, Hive, and other HDFS ecosystem supporting software.

Term project will take many approaches for collection, joining, analysis, and visualization. For example, table collection, analysis, deliverables will be done through class models.

Analysis of these datasets will be used for transforming the raw data into meaningful and crucial information which will help in discovering certain trends in Crime rate, Arrests and vehicles and/or pedestrians being stopped by the LA Police Department.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**--Load website log data and use file name to Hdfs**

**--Understand ' . ' specifies to current/home directory**

**--Vehicle and Pedestrian Stops**

**$ wget -O stops.csv https://data.lacity.org/api/views/ci25-wgt7/rows.csv?accessType=DOWNLOAD**

**$ hdfs dfs -put stops.csv**

**$ hdfs dfs -ls**

**--Arrests**

**$ wget -O arrests.csv https://data.lacity.org/api/views/yru6-6re4/rows.csv?accessType=DOWNLOAD**

**$ hdfs dfs -put arrests.csv**

**$ hdfs dfs -ls**

**--Crimes**

**$ wget -O crimes.csv https://data.lacity.org/api/views/y8tr-7khq/rows.csv?accessType=DOWNLOAD**

**$ hdfs dfs -put crimes.csv .**

**$ hdfs dfs -ls**

**—-Demographic and Economic Statistics**

**$ wget -O demo.csv https://controllerdata.lacity.org/api/views/9z75-k9sy/rows.csv?accessType=DOWNLOAD**

**$ hdfs dfs -put demo.csv .**

**$ hdfs dfs -ls**

**-—2010 Census**

**$ wget -O census.csv https://data.lacity.org/api/views/nxs9-385f/rows.csv?accessType=DOWNLOAD**

**$ hdfs dfs -put census.csv .**

**$ hdfs dfs -ls**

**—-Economy Panel**

**$ wget -O econ.csv https://controllerdata.lacity.org/api/views/4ndk-mmc8/rows.csv?accessType=DOWNLOAD**

**$ hdfs dfs -put econ.csv .**

**$ hdfs dfs -ls**

**--Execute PigQL**

**$ pig**

**--LOAD and then STORE stops**

**grunt> stops = LOAD '/user/actejas/stops.csv' USING PigStorage(',') AS (stop\_number:chararray,ref\_number:chararray,sex:chararray,descent:chararray,descent\_desc:chararray,stop\_date:chararray,stop\_time:chararray,off1\_number:chararray,off1\_divnum:chararray,div1\_desc:chararray,off2\_number:chararray,off2\_divnum:chararray,div2\_desc:chararray,report\_dist:chararray,stop\_type:chararray,post\_stop\_activity:chararray);**

**grunt> a = LIMIT stops 5;**

**grunt> dump a;**

**grunt> STORE stops INTO 'PROJECT/STOPS';**

**--LOAD and then STORE arrests**

**grunt> arrests = LOAD '/user/actejas/arrests.csv' USING PigStorage(',') AS (report\_id:chararray,arrests\_date:chararray,arrests\_time:chararray,area\_id:chararray,area\_name:chararray,reporting\_dist:chararray,age:chararray,sex:chararray,descent\_code:chararray,charge\_group\_code:chararray,charge\_group\_desc:chararray,arrest\_type:chararray,charge:chararray,charge\_decs:chararray,address:chararray,cross\_street:chararray,location:chararray);**

**grunt> a = LIMIT arrests 5;**

**grunt> dump a;**

**grunt> STORE arrests INTO 'PROJECT/ARREST';**

**--LOAD and then STORE crimes**

**crimes = LOAD '/user/actejas/crimes.csv' USING PigStorage(',') AS (dr\_number:chararray,crimes\_date\_reported:chararray,crimes\_date:chararray,crimes\_time:chararray,area\_id:chararray,area\_name:chararray,reporting\_disct:chararray,crime\_code:chararray,mo\_code:chararray,victim\_age:chararray,victim\_sex:chararray,victim\_descent:chararray,premise\_code:chararray,premise\_desc:chararray,weapon\_code:chararray,weapon\_desc:chararray,status\_code:chararray,status\_desc:chararray,crime\_code1:chararray,crime\_code2:chararray,crime\_code3:chararray,crime\_code4:chararray,address:chararray,cross\_street:chararray,location:chararray);**

**grunt> STORE crimes INTO 'PROJECT/CRIMES';**

**--LOAD and then STORE demo**

**grunt> demo = LOAD '/user/actejas/demo.csv' USING PigStorage(',') AS (fiscal\_year:chararray,estimated\_population:chararray,income\_in\_thousands:chararray,income\_per\_capita:chararray,median\_age:chararray,public\_school\_enrollment:chararray,unemployment\_rate:chararray);**

**grunt> STORE demo INTO 'PROJECT/DEMO';**

**--LOAD and then STORE census**

**grunt> census = LOAD '/user/actejas/census.csv' USING PigStorage(',') AS (zip\_code:chararray,population:chararray,median\_age:chararray,total\_males:chararray,total\_females:chararray,total\_households:chararray,avg\_household\_size:chararray);**

**grunt> STORE census INTO 'PROJECT/CENSUS';**

**--LOAD and then STORE econ**

**grunt> econ = LOAD '/user/actejas/econ.csv' USING PigStorage(',') AS (category:chararray,indicator:chararray,sub\_indicator:chararray,locality:chararray,gender:chararray,council\_district:chararray,fiscal\_year:chararray,fy\_qtr:chararray,calender\_year:chararray,cy\_qtr:chararray,quarter\_start:chararray,as\_of\_date:chararray,unit\_of\_measure:chararray,unit\_text:chararray,value:chararray,source:chararray);**

**grunt> STORE econ INTO 'PROJECT/ECON';**

**1> --Which victim profile is most and least affected by LAPD stops?**

**grunt> a = GROUP stops by (descent);**

**grunt> b = FOREACH a GENERATE FLATTEN(group) as (descent), COUNT($1);**

**grunt> dump b;**

**2> --Which reporting district is most and least affected by LAPD stops?**

**grunt> a = GROUP stops by report\_dist;**

**grunt> b = FOREACH a GENERATE FLATTEN(group) as (report\_dist), COUNT($1);**

**grunt> c = ORDER b BY $1 DESC;**

**grunt> d = LIMIT c 5;**

**grunt> dump d;**

**3> —>which year has the most and least crime activity;**

crimes = LOAD '/user/actejas/crimes.csv' USING PigStorage(',') AS (dr\_number:chararray,crimes\_date\_reported:chararray,crimes\_date:chararray,crimes\_time:chararray,area\_id:chararray,area\_name:chararray,reporting\_disct:chararray,crime\_code:chararray,mo\_code:chararray,victim\_age:chararray,victim\_sex:chararray,victim\_descent:chararray,premise\_code:chararray,premise\_desc:chararray,weapon\_code:chararray,weapon\_desc:chararray,status\_code:chararray,status\_desc:chararray,crime\_code1:chararray,crime\_code2:chararray,crime\_code3:chararray,crime\_code4:chararray,address:chararray,cross\_street:chararray,location:chararray);

a = STREAM crimes THROUGH `tail -n +2` AS (dr\_number:chararray,crimes\_date\_reported:chararray,crimes\_date:chararray,crimes\_time:chararray,area\_id:chararray,area\_name:chararray,reporting\_disct:chararray,crime\_code:chararray,mo\_code:chararray,victim\_age:chararray,victim\_sex:chararray,victim\_descent:chararray,premise\_code:chararray,premise\_desc:chararray,weapon\_code:chararray,weapon\_desc:chararray,status\_code:chararray,status\_desc:chararray,crime\_code1:chararray,crime\_code2:chararray,crime\_code3:chararray,crime\_code4:chararray,address:chararray,cross\_street:chararray,location:chararray);

b = FOREACH a GENERATE dr\_number AS dr\_number, SUBSTRING(crimes\_date,6,10) AS date;

c = GROUP b by date;

d = FOREACH c GENERATE group as date,COUNT\_STAR(b.dr\_number);

dump d;

**4> → which year has the most and least arrests activity;**

arrests = LOAD '/user/actejas/arrests.csv' USING PigStorage(',') AS (report\_id:chararray,arrests\_date:chararray,arrests\_time:chararray,area\_id:chararray,area\_name:chararray,reporting\_dist:chararray,age:chararray,sex:chararray,descent\_code:chararray,charge\_group\_code:chararray,charge\_group\_desc:chararray,arrest\_type:chararray,charge:chararray,charge\_decs:chararray,address:chararray,cross\_street:chararray,location:chararray);

a = STREAM arrests THROUGH `tail -n +2` AS (report\_id:chararray,arrests\_date:chararray,arrests\_time:chararray,area\_id:chararray,area\_name:chararray,reporting\_dist:chararray,age:chararray,sex:chararray,descent\_code:chararray,charge\_group\_code:chararray,charge\_group\_desc:chararray,arrest\_type:chararray,charge:chararray,charge\_decs:chararray,address:chararray,cross\_street:chararray,location:chararray);

b = FOREACH a GENERATE report\_id AS report\_id, SUBSTRING(arrests\_date,6,10) AS date;

c = GROUP b by date;

d = FOREACH c GENERATE group as date,COUNT\_STAR(b.report\_id);

dump d;

**5> → which year has the most and least stops activity;**

**stops = LOAD '/user/actejas/stops.csv' USING PigStorage(',') AS (stop\_number:chararray,ref\_number:chararray,sex:chararray,descent:chararray,descent\_desc:chararray,stop\_date:chararray,stop\_time:chararray,off1\_number:chararray,off1\_divnum:chararray,div1\_desc:chararray,off2\_number:chararray,off2\_divnum:chararray,div2\_desc:chararray,report\_dist:chararray,stop\_type:chararray,post\_stop\_activity:chararray);**

a = STREAM stops THROUGH `tail -n +2` AS **(stop\_number:chararray,ref\_number:chararray,sex:chararray,descent:chararray,descent\_desc:chararray,stop\_date:chararray,stop\_time:chararray,off1\_number:chararray,off1\_divnum:chararray,div1\_desc:chararray,off2\_number:chararray,off2\_divnum:chararray,div2\_desc:chararray,report\_dist:chararray,stop\_type:chararray,post\_stop\_activity:chararray);**

b = FOREACH a GENERATE stop\_number AS stop\_number, SUBSTRING(stop\_date,6,10) AS date;

c = GROUP b by date;

d = FOREACH c GENERATE group as date,COUNT\_STAR(b.stop\_number);

dump d;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

-—GROUP crimes by victim descent and show which descents are more frequent victims

grunt> descent = GROUP crimes BY ‘victim\_descent’;

grunt> descent\_count = FOREACH descent GENERATE COUNT(crimes BY ‘victim descent’);

grunt> DUMP descent\_count;

—-LOAD projects/demo

grunt> demo = LOAD ‘demo’ USING PigStorage(',') AS (fiscal year: int, unemployment rate:int);

grunt> DESCRIBE demo;

grunt> unemployment= GROUP demo ALL;

grunt> DUMP unemployment;

—-GROUP crimes by date occurred and how which years have most crime. Then show if increased years of unemployment like in 2008 correlates to increased crime

grunt> date = GROUP crimes BY date occurred;

grunt> date\_year = FOREACH date GENERATE GetYear AS ‘year’;

grunt> DUMP date\_year;

grunt> date\_count = FOREACH date\_year GENERATE COUNT(‘year’);

grunt> DUMP date\_count;

--Nesting function example

top\_salaries = FOREACH title\_group {  
 sorted = ORDER employees BY salary DESC;  
 highest\_paid = LIMIT sorted 3;  
 GENERATE group, highest\_paid;

--Reporting District Legend

<https://www.arcgis.com/home/item.html?id=c4f83909b81d4786aa8ba8a74a4b4db1#data>

--Map

<http://geohub.lacity.org/datasets/c4f83909b81d4786aa8ba8a74a4b4db1_4?uiTab=table&geometry=-119.993%2C33.679%2C-117.702%2C34.362>

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**LAPD ACTIVITY AND GEO-SPATIAL ANALYSIS**

**Which reporting district is most and least affected by LAPD stops**

a = GROUP stops by report\_dist;

b = FOREACH a GENERATE FLATTEN(group) as (report\_dist), COUNT($1);

c = ORDER b BY $1 DESC;

d = LIMIT C 5;

**Need to relate stops crimes and arrests using Report Districts to fetch the number of LAPD activity in each Area(Geo-Spatial Insight).**

**Link for zipcodes of EACH Reporting DIstrict**

[**http://www.lapdonline.org/our\_communities/content\_basic\_view/6279**](http://www.lapdonline.org/our_communities/content_basic_view/6279)

**--> Which reporting district is most and least affected by LAPD crimes**

**--> Which reporting district is most and least affected by LAPD arrests**

**which year has the most and least stops activity;**

**stops = LOAD '/user/actejas/stops.csv' USING PigStorage(',') AS (stop\_number:chararray,ref\_number:chararray,sex:chararray,descent:chararray,descent\_desc:chararray,stop\_date:chararray,stop\_time:chararray,off1\_number:chararray,off1\_divnum:chararray,div1\_desc:chararray,off2\_number:chararray,off2\_divnum:chararray,div2\_desc:chararray,report\_dist:chararray,stop\_type:chararray,post\_stop\_activity:chararray);**

a = STREAM stops THROUGH `tail -n +2` AS **(stop\_number:chararray,ref\_number:chararray,sex:chararray,descent:chararray,descent\_desc:chararray,stop\_date:chararray,stop\_time:chararray,off1\_number:chararray,off1\_divnum:chararray,div1\_desc:chararray,off2\_number:chararray,off2\_divnum:chararray,div2\_desc:chararray,report\_dist:chararray,stop\_type:chararray,post\_stop\_activity:chararray);**

b = FOREACH a GENERATE stop\_number AS stop\_number, SUBSTRING(stop\_date,6,10) AS date;

c = GROUP b by date;

d = FOREACH c GENERATE group as date,COUNT\_STAR(b.stop\_number);

dump d;

**which year has the most and least crime activity;**

crimes = LOAD '/user/actejas/crimes.csv' USING PigStorage(',') AS (dr\_number:chararray,crimes\_date\_reported:chararray,crimes\_date:chararray,crimes\_time:chararray,area\_id:chararray,area\_name:chararray,reporting\_disct:chararray,crime\_code:chararray,crime\_code\_desc:chararray,mo\_code:chararray,victim\_age:chararray,victim\_sex:chararray,victim\_descent:chararray,premise\_code:chararray,premise\_desc:chararray,weapon\_code:chararray,weapon\_desc:chararray,status\_code:chararray,status\_desc:chararray,crime\_code1:chararray,crime\_code2:chararray,crime\_code3:chararray,crime\_code4:chararray,address:chararray,cross\_street:chararray,location:chararray);

a = STREAM crimes THROUGH `tail -n +2` AS (dr\_number:chararray,crimes\_date\_reported:chararray,crimes\_date:chararray,crimes\_time:chararray,area\_id:chararray,area\_name:chararray,reporting\_disct:chararray,crime\_code:chararray,crime\_code\_desc:chararray,mo\_code:chararray,victim\_age:chararray,victim\_sex:chararray,victim\_descent:chararray,premise\_code:chararray,premise\_desc:chararray,weapon\_code:chararray,weapon\_desc:chararray,status\_code:chararray,status\_desc:chararray,crime\_code1:chararray,crime\_code2:chararray,crime\_code3:chararray,crime\_code4:chararray,address:chararray,cross\_street:chararray,location:chararray);

b = FOREACH a GENERATE dr\_number AS dr\_number, SUBSTRING(crimes\_date,6,10) AS date;

c = GROUP b by date;

d = FOREACH c GENERATE group as date,COUNT\_STAR(b.dr\_number);

dump d;

**which year has the most and least arrests activity;**

arrests = LOAD '/user/actejas/arrests.csv' USING PigStorage(',') AS (report\_id:chararray,arrests\_date:chararray,arrests\_time:chararray,area\_id:chararray,area\_name:chararray,reporting\_dist:chararray,age:chararray,sex:chararray,descent\_code:chararray,charge\_group\_code:chararray,charge\_group\_desc:chararray,arrest\_type:chararray,charge:chararray,charge\_desc:chararray,address:chararray,cross\_street:chararray,location:chararray);

a = STREAM arrests THROUGH `tail -n +2` AS (report\_id:chararray,arrests\_date:chararray,arrests\_time:chararray,area\_id:chararray,area\_name:chararray,reporting\_dist:chararray,age:chararray,sex:chararray,descent\_code:chararray,charge\_group\_code:chararray,charge\_group\_desc:chararray,arrest\_type:chararray,charge:chararray,charge\_desc:chararray,address:chararray,cross\_street:chararray,location:chararray);

b = FOREACH a GENERATE report\_id AS report\_id, SUBSTRING(arrests\_date,6,10) AS date;

c = GROUP b by date;

d = FOREACH c GENERATE group as date,COUNT\_STAR(b.report\_id);

dump d;

**--> Which month has the most and least stops activity**

**stops = LOAD '/user/actejas/stops.csv' USING PigStorage(',') AS (stop\_number:chararray,ref\_number:chararray,sex:chararray,descent:chararray,descent\_desc:chararray,stop\_date:chararray,stop\_time:chararray,off1\_number:chararray,off1\_divnum:chararray,div1\_desc:chararray,off2\_number:chararray,off2\_divnum:chararray,div2\_desc:chararray,report\_dist:chararray,stop\_type:chararray,post\_stop\_activity:chararray);**

a = STREAM stops THROUGH `tail -n +2` AS **(stop\_number:chararray,ref\_number:chararray,sex:chararray,descent:chararray,descent\_desc:chararray,stop\_date:chararray,stop\_time:chararray,off1\_number:chararray,off1\_divnum:chararray,div1\_desc:chararray,off2\_number:chararray,off2\_divnum:chararray,div2\_desc:chararray,report\_dist:chararray,stop\_type:chararray,post\_stop\_activity:chararray);**

b = FOREACH a GENERATE stop\_number AS stop\_number, SUBSTRING(stop\_date,0,2) AS date;

c = GROUP b by date;

d = FOREACH c GENERATE group as date,COUNT\_STAR(b.stop\_number);

dump d;

**--> Which month has the most and least crimes activity**

crimes = LOAD '/user/actejas/crimes.csv' USING PigStorage(',') AS (dr\_number:chararray,crimes\_date\_reported:chararray,crimes\_date:chararray,crimes\_time:chararray,area\_id:chararray,area\_name:chararray,reporting\_disct:chararray,crime\_code:chararray,crime\_code\_desc:chararray,mo\_code:chararray,victim\_age:chararray,victim\_sex:chararray,victim\_descent:chararray,premise\_code:chararray,premise\_desc:chararray,weapon\_code:chararray,weapon\_desc:chararray,status\_code:chararray,status\_desc:chararray,crime\_code1:chararray,crime\_code2:chararray,crime\_code3:chararray,crime\_code4:chararray,address:chararray,cross\_street:chararray,location:chararray);

a = STREAM crimes THROUGH `tail -n +2` AS (dr\_number:chararray,crimes\_date\_reported:chararray,crimes\_date:chararray,crimes\_time:chararray,area\_id:chararray,area\_name:chararray,reporting\_disct:chararray,crime\_code:chararray,crime\_code\_desc:chararray,mo\_code:chararray,victim\_age:chararray,victim\_sex:chararray,victim\_descent:chararray,premise\_code:chararray,premise\_desc:chararray,weapon\_code:chararray,weapon\_desc:chararray,status\_code:chararray,status\_desc:chararray,crime\_code1:chararray,crime\_code2:chararray,crime\_code3:chararray,crime\_code4:chararray,address:chararray,cross\_street:chararray,location:chararray);

b = FOREACH a GENERATE dr\_number AS dr\_number, SUBSTRING(crimes\_date,0,2) AS date;

c = GROUP b by date;

d = FOREACH c GENERATE group as date,COUNT\_STAR(b.dr\_number);

dump d;

**--> Which month has the most and least arrests activity**

arrests = LOAD '/user/actejas/arrests.csv' USING PigStorage(',') AS (report\_id:chararray,arrests\_date:chararray,arrests\_time:chararray,area\_id:chararray,area\_name:chararray,reporting\_dist:chararray,age:chararray,sex:chararray,descent\_code:chararray,charge\_group\_code:chararray,charge\_group\_desc:chararray,arrest\_type:chararray,charge:chararray,charge\_desc:chararray,address:chararray,cross\_street:chararray,location:chararray);

a = STREAM arrests THROUGH `tail -n +2` AS (report\_id:chararray,arrests\_date:chararray,arrests\_time:chararray,area\_id:chararray,area\_name:chararray,reporting\_dist:chararray,age:chararray,sex:chararray,descent\_code:chararray,charge\_group\_code:chararray,charge\_group\_desc:chararray,arrest\_type:chararray,charge:chararray,charge\_desc:chararray,address:chararray,cross\_street:chararray,location:chararray);

b = FOREACH a GENERATE report\_id AS report\_id, SUBSTRING(arrests\_date,0,2) AS date;

c = GROUP b by date;

d = FOREACH c GENERATE group as date,COUNT\_STAR(b.report\_id);

dump d;

**-->Find the % relation of crimes that led or did not lead to arrests**

**LAPD ACTIVITY AND DEMOGRAPHIC/ECONOMIC ANALYSIS**

**which victim profile is most and least affected by stops;**

a = GROUP stops by (descent);

b = FOREACH a GENERATE FLATTEN(group) as (descent), COUNT($1);

dump b;

**—> which victim profile is most and least affected by crimes**

a = GROUP crimes by (victim\_descent);

b = FOREACH a GENERATE FLATTEN(group) as (victim\_descent), COUNT($1);

dump b;

a = GROUP crimes by (victim\_age);

b = FOREACH a GENERATE FLATTEN(group) as (victim\_age), COUNT($1);

dump b;

**—> which victim profile is most and least affected by arrests**

a = GROUP arrests by (descent\_code);

b = FOREACH a GENERATE FLATTEN(group) as (descent\_code), COUNT($1);

dump b;

**→ which age is most and least affected by crimes**

a = GROUP crimes by (descent\_code);

b = FOREACH a GENERATE FLATTEN(group) as (descent\_code), COUNT($1);

dump b;

**—> Crime years correlate to unemployment rates**

demo = LOAD '/user/actejas/demo.csv' USING PigStorage(',') AS (fiscal\_year:chararray,estimated\_population:chararray,income\_in\_thousands:chararray,income\_per\_capita:chararray,median\_age:chararray,public\_school\_enrollment:chararray,unemployment\_rate:chararray);

a = STREAM demo THROUGH `tail -n +2` AS (fiscal\_year:chararray,estimated\_population:chararray,income\_in\_thousands:chararray,income\_per\_capita:chararray,median\_age:chararray,public\_school\_enrollment:chararray,unemployment\_rate:chararray);

b = FOREACH a GENERATE fiscal\_year AS fiscal\_year,unemployment\_rate AS unemployment\_rate;

c = GROUP b by fiscal\_year;

d = FOREACH c GENERATE group as fiscal\_year,b.unemployment\_rate;

dump d;

**—> Crime years correlate to public school enrollment**

demo = LOAD '/user/actejas/demo.csv' USING PigStorage(',') AS (fiscal\_year:chararray,estimated\_population:chararray,income\_in\_thousands:chararray,income\_per\_capita:chararray,median\_age:chararray,public\_school\_enrollment:chararray,unemployment\_rate:chararray);

a = STREAM demo THROUGH `tail -n +2` AS (fiscal\_year:chararray,estimated\_population:chararray,income\_in\_thousands:chararray,income\_per\_capita:chararray,median\_age:chararray,public\_school\_enrollment:chararray,unemployment\_rate:chararray);

b = FOREACH a GENERATE fiscal\_year AS fiscal\_year,public\_school\_enrollment AS public\_school\_enrollment;

c = GROUP b by fiscal\_year;

d = FOREACH c GENERATE group as fiscal\_year,b.public\_school\_enrollment;

dump d;

**—> Crime years correlate to personal income per capita**

demo = LOAD '/user/actejas/demo.csv' USING PigStorage(',') AS (fiscal\_year:chararray,estimated\_population:chararray,income\_in\_thousands:chararray,income\_per\_capita:chararray,median\_age:chararray,public\_school\_enrollment:chararray,unemployment\_rate:chararray);

a = STREAM demo THROUGH `tail -n +2` AS (fiscal\_year:chararray,estimated\_population:chararray,income\_in\_thousands:chararray,income\_per\_capita:chararray,median\_age:chararray,public\_school\_enrollment:chararray,unemployment\_rate:chararray);

b = FOREACH a GENERATE fiscal\_year AS fiscal\_year,income\_per\_capita AS income\_per\_capita;

c = GROUP b by fiscal\_year;

d = FOREACH c GENERATE group as fiscal\_year,b.income\_per\_capita;

dump d;

**→ Total number of Crimes and arrests per year.**

crimes = LOAD '/user/actejas/crimes.csv' USING PigStorage(',') AS (dr\_number:chararray,crimes\_date\_reported:chararray,crimes\_date:chararray,crimes\_time:chararray,area\_id:chararray,area\_name:chararray,reporting\_disct:chararray,crime\_code:chararray,crime\_code\_desc:chararray,mo\_code:chararray,victim\_age:chararray,victim\_sex:chararray,victim\_descent:chararray,premise\_code:chararray,premise\_desc:chararray,weapon\_code:chararray,weapon\_desc:chararray,status\_code:chararray,status\_desc:chararray,crime\_code1:chararray,crime\_code2:chararray,crime\_code3:chararray,crime\_code4:chararray,address:chararray,cross\_street:chararray,location:chararray);

a = STREAM crimes THROUGH `tail -n +2` AS (dr\_number:chararray,crimes\_date\_reported:chararray,crimes\_date:chararray,crimes\_time:chararray,area\_id:chararray,area\_name:chararray,reporting\_disct:chararray,crime\_code:chararray,crime\_code\_desc:chararray,mo\_code:chararray,victim\_age:chararray,victim\_sex:chararray,victim\_descent:chararray,premise\_code:chararray,premise\_desc:chararray,weapon\_code:chararray,weapon\_desc:chararray,status\_code:chararray,status\_desc:chararray,crime\_code1:chararray,crime\_code2:chararray,crime\_code3:chararray,crime\_code4:chararray,address:chararray,cross\_street:chararray,location:chararray);

b = FOREACH a GENERATE dr\_number AS Drnumber , SUBSTRING(crimes\_date,6,10) AS crime\_date;

c = GROUP b by crime\_date;

d = FOREACH c GENERATE group as crime\_date, COUNT\_STAR(b.Drnumber);

dump d;

arrests = LOAD '/user/actejas/arrests.csv' USING PigStorage(',') AS (report\_id:chararray,arrests\_date:chararray,arrests\_time:chararray,area\_id:chararray,area\_name:chararray,reporting\_dist:chararray,age:chararray,sex:chararray,descent\_code:chararray,charge\_group\_code:chararray,charge\_group\_desc:chararray,arrest\_type:chararray,charge:chararray,charge\_decs:chararray,address:chararray,cross\_street:chararray,location:chararray);

a = STREAM arrests THROUGH `tail -n +2` AS (report\_id:chararray,arrests\_date:chararray,arrests\_time:chararray,area\_id:chararray,area\_name:chararray,reporting\_dist:chararray,age:chararray,sex:chararray,descent\_code:chararray,charge\_group\_code:chararray,charge\_group\_desc:chararray,arrest\_type:chararray,charge:chararray,charge\_decs:chararray,address:chararray,cross\_street:chararray,location:chararray);

b = FOREACH a GENERATE report\_id AS reportid , SUBSTRING(arrests\_date,6,10) AS arrests\_date;

c = GROUP b by arrests\_date;

d = FOREACH c GENERATE group as arrests\_date, COUNT\_STAR(b.reportid);

dump d;

**-->Which LAPD Reporting District Has the Most Reports of LAPD Activity**

crimes = LOAD '/user/actejas/crimes.csv' USING PigStorage(',') AS (dr\_number:chararray,crimes\_date\_reported:chararray,crimes\_date:chararray,crimes\_time:chararray,area\_id:chararray,area\_name:chararray,reporting\_dist:chararray,crime\_code:chararray,crime\_code\_desc:chararray,mo\_code:chararray,victim\_age:chararray,victim\_sex:chararray,victim\_descent:chararray,premise\_code:chararray,premise\_desc:chararray,weapon\_code:chararray,weapon\_desc:chararray,status\_code:chararray,status\_desc:chararray,crime\_code1:chararray,crime\_code2:chararray,crime\_code3:chararray,crime\_code4:chararray,address:chararray,cross\_street:chararray,location:chararray);

arrests = LOAD '/user/actejas/arrests.csv' USING PigStorage(',') AS (report\_id:chararray,arrests\_date:chararray,arrests\_time:chararray,area\_id:chararray,area\_name:chararray,reporting\_dist:chararray,age:chararray,sex:chararray,descent\_code:chararray,charge\_group\_code:chararray,charge\_group\_desc:chararray,arrest\_type:chararray,charge:chararray,charge\_decs:chararray,address:chararray,cross\_street:chararray,location:chararray);

stops = LOAD '/user/actejas/stops.csv' USING PigStorage(',') AS (stop\_number:chararray,ref\_number:chararray,sex:chararray,descent:chararray,descent\_desc:chararray,stop\_date:chararray,stop\_time:chararray,off1\_number:chararray,off1\_divnum:chararray,div1\_desc:chararray,off2\_number:chararray,off2\_divnum:chararray,div2\_desc:chararray,report\_dist:chararray,stop\_type:chararray,post\_stop\_activity:chararray);

zipcode = LOAD '/user/actejas/zipcode.csv' USING PigStorage(',') AS (reporting\_dist:chararray,area\_name:chararray,zip\_code:chararray);

a = JOIN arrests by reporting\_dist,zipcodes by reporting\_dist;

b = FOREACH a GENERATE arrests::reporting\_dist AS reporting\_dist , arrests::arrests\_date AS arrests\_date, zipcode::zip\_code AS zipcode;

dump b;

STORE b INTO 'OUTPUT/ARRESTDATAFORMAP';

a = JOIN crimes by reporting\_dist,zipcode by reporting\_dist;

b = FOREACH a GENERATE crimes::reporting\_dist AS reporting\_dist , crime::crimes\_date AS crimes\_date, zipcode::zip\_code AS zipcode;

dump b;

STORE b INTO 'OUTPUT/CRIMESDATAFORMAP';

a = JOIN stops by reporting\_dist,zipcode by reporting\_dist;

b = FOREACH a GENERATE stops::report\_dist AS reporting\_dist , stops::stop\_date AS crimes\_date, zipcode::zip\_code AS zipcode;

dump b;

STORE b INTO 'OUTPUT/STOPSFORMAP';