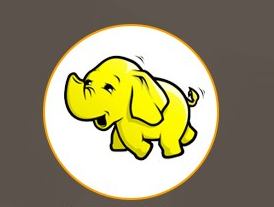
**ACADGILD**

**USA**

**CRIME**

**ANALYSIS**



**LEARN.DO.LEARN**

**acadgild.com**

**ABOUT ACADGILD**

ACADGILD is a technology education startup that aims to create an ecosystem for skilled development in which people can learn from mentors and from each other. We believe that soft­ware development requires highly specialized skills that are best learned with guidance from experienced practitioners. Online videos or classroom formats are poor substitutes for building real projects with help from a dedicated mentor. Our mission is to teach hands-on, job-ready soft­ware programming skills, globally, in small batches of 8 to 10 students, using industry experts.

**ACADGILD** offers courses in:



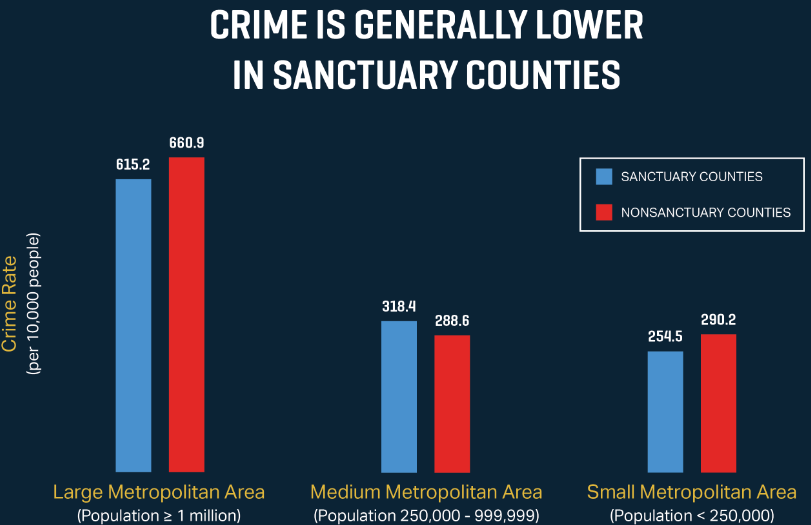
**acadgild.com**

**LEARN.DO.LEARN**

|  |
| --- |
| TABLE OF CONTENTS |
| 1.0 Project Description. |
| 2.0 Project Requirements |
| 3.0 Data Files |
| 4.0 Data Ingestion and Validation |
| 5.0 Data Analysis |
| 6.0 Project Implementation and Source Codes |
| 7.0 Conclusions |
| 8.0 Acknowledgement |
| 9.0 Bibliography |
|  |

**PROJECT DESCRIPTION**

A leading crime investigation company in USA is planning to analyze a large amount of data received from varieties of sources from website to track the behaviour of users, classify users, and calculate royalties associated with the number of investigated cases and make appropriate business strategies.



PROJECT REQUIREMENTS

HARDWARE REQUIREMENTS:

(i) Windows 8.1 with Linux (Intel CentOS) installed in Oracle Virtual Box.

(ii) Ram: 4GB and Linux Virtual Box Ram: 2GB.

(iii) Hard Disk Drive: 1TB.

(iv) Intel Core I3 Processor.

SOFTWARE REQUIREMENTS:

(i) Pig script in Linux for executing and implementing the data analysis part.

(ii) Microsoft Word 2010 for writing the project report.

|  |
| --- |
| DATA FILES |

ID, Case Number, Date, Block, IUCR, Primary Type, Description, Location Description, Arrest, Domestic, Beat, District, Ward, Community Area, FBICode, X Coordinate, Y Coordinate, Year, Updated On, Latitude, Longitude, Location

DATA FILEDS CONTAINS THE FOLLOWING :

|  |  |
| --- | --- |
| FIELD NAME | FIELD DESCRIPTION |
| Id | Unique identifier of every investigation |
| Case\_No | Provides the no of cases in each state |
| Date | This allows to keep the track of the crime information |
| Block | Provides the name of the block where the crime has happened |
| IUCR | Identifies the code of the crime investigation |
| Primary\_Type | Provides the type of crime which had happened |
| Description | Provides the description of crimes happened |
| Location Description | Provides the Location of where the crime had happened |
| Arrest | Provides that whether the crime which was happened was arrested or not |
| Domestic | Provides and evidence whether any domestic crimes happened or not. |
| Beat | Provides the number of polices in a police officer patrol |
| District | Provides the area where the police officers are present. |
| Ward | Provides the Ward number of police officer patrol. |
| Community\_Area | Defines the area |
| FBI\_Code | Uniquely identifies the investigation code. |
| X-coordinates | Examination of investigation in a crime scene |
| Y-coordinates | Examination of investigation in a crime scene |
| Year | Provides the year when the crime had happened |
| Updated\_on | Updated date of the crime investigation solved |
| Latitude |  |
| Longitude |  |
| Location |  |

**DATA INGESTION AND VALIDATION**

1. Data coming from web applications has a csv format.
2. All the Date fields coming from web applications is of the format DD-MM-YY.
3. If any of the fields Id, Case\_No, Date, Block, IUCR, Beat, District, Ward, Community\_Area, FBI\_Code is NULL or absent, consider that record is invalid.

**DATA ANALYSIS**

It is not only the data which is important, rather it is the insight it can be used to generate important. Once we have made the data ready for analysis, we have to perform below analysis on a daily basis.

(i) Write a Pig program to calculate the number of cases investigated under each FBI code.

(ii) Write a Pig program to calculate the number of cases investigated under Ward 32.

(iii) Write a Pig program to calculate the number of arrests in theft district wise.

(iii) Write a Pig program to calculate the number of arrests occurs between October 2014 and October 2015.

**IMPLEMENTATION AND SOURCE CODES**

Using Pig script we are able to execute the data analysis problems:

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

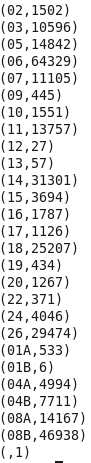
(i) **Write a Pig program to calculate the number of cases investigated under each FBI code.**

Solution:



Output:

This program displays the number of cases investigated under FBIcode



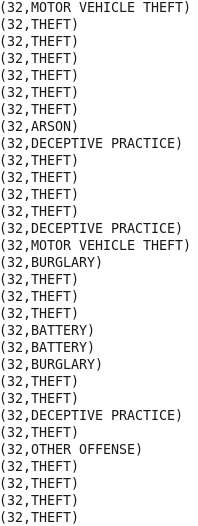
(ii) **Write a Pig program to calculate the number of cases investigated under Ward 32.**

Solution:



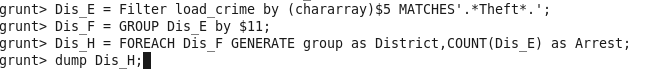
Output:

This program displays the number of cases under Ward number 32



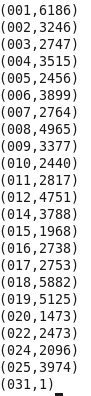
(iii) **Write a Pig program to calculate the number of arrests in theft district wise.**

Solution:



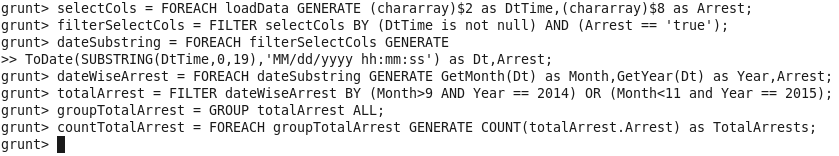
Output:

This program displays the number of arrest done by theft and in which District.



(iv) **Write a Pig program to calculate the number of arrests done between October 2014 and October 2015.**

Solution:



Output:

This program displays the sum total of number of arrests happened in between October 2014 and October 2015.



CONCLUSIONS

ONCLUSION

A study on Hadoop component apache pig is made in order to write MapReduce code to analyze the Big Data in a minimum of time with simple coding format. An experimental analysis is made using a library data set with 3.40 lakh record made an analysis of frequent library users, authors preferred by the students, date and time accessed by the students frequently are extracted .The above results are extracted within five steps of coding using Pig script and the results are obtained on average of 35 sec of time duration. So from the above analyze it is concluded that using pig script it is possible to handle Big Database in an easy and efficient manner with minimum of time duration with simple coding.

I would like to express my special gratitude to all the acadgild technical support team for helping me to complete the project within a span of time and also helping me to clear my doubts and clarification in the project.

**ACKNOWLEDGEMENT**

Secondly I would like to express my special thanks to my acadgild mentor ‘Mr. Ravi Kshirsagar’ for clearing all the concepts during the entire session of ‘Big Data Hadoop & Spark’ training course.

**BIBLIOGRAPHY**

The references that I have used to complete my project are:

(i) Study material given by acadgild.

(ii) Need the help from internet.



**BY**

**SOHAM NEOGY**

**ACADGILD**