INTRODUCTION

3.1 Crime Analysis :-

Crime analysis is a law enforcement function that involves systematic analysis for identifying and analyzing patterns and trends in crime and disorder. Information on patterns can help law enforcement agencies deploy resources in a more effective manner, and assist detectives in identifying and apprehending suspects. Crime analysis also plays a role in devising solutions to crime problems, and formulating crime prevention strategies. Quantitative social science data analysis methods are part of the crime analysis process, though qualitative methods such as examining police report narratives also play a role.

3.1.1 Types of Crime Analysis

i) Intelligence Analysis :

Intelligence Analysis is the process of taking known information about situations and entities of strategic, operational, or tactical importance, characterizing the known, and, with appropriate statements of probability, the future actions in those situations and by those entities.

ii) Criminal Investigative –

Offender profiling, also known as criminal profiling, is an investigative tool used by law enforcement agencies to identify likely suspects (descriptive offender profiling) and analyze patterns that may predict future offenses and victims (predictive offender profiling).

iii) Tactical Crime :

The study of recent criminal incidents and potential criminal activity by examining characteristics such as how, when, and where the activity has occurred to assist in problem solving by developing patterns and trends, identifying investigative leads/suspects,. “Recent” can refer to the last few months or longer periods of time for specific ongoing problems. Tactical

crime analysis also focuses on specific information about each crime such as method of entry, point of entry, suspects actions, type of victim, type of weapon used, as well as the date, time,

location, and type of location. Field information such as suspicious activity calls for service,criminal trespass warnings, and persons with scars, marks, or tattoos collected by officers is also considered in the analysis. Although quantitative analysis is often conducted once a pattern has been identified, qualitative analysis, (i.e., critical thinking and content analysis) is used to identify patterns and trends initially. Three purposes of tactical crime analysis are 1) linking cases together and identifying the notable characteristics of the patterns and trends, 2) identifying potential suspects of a crime or crime pattern, and 3) clearing cases. The focus of tactical crime analysis is examining data daily in order to identify patterns, trends, and investigative leads for recent criminal and potential criminal activity. Once a crime pattern, suspect, or investigative lead is identified, the information is compiled and disseminated to patrol officers and detectives.

iv) Strategic Crime Analysis :

Strategic analysis is a law enforcement function that involves systematic analysis for identifying and analyzing patterns and trends in crime and disorder. Information on patterns can help law enforcement agencies deploy resources in a more effective manner, and assist detectives in identifying and apprehending suspects.....

v) Administrative Crime Analysis :-

Administrative crime analysis is different from the previous types of analysis in that it refers to presentation of findings rather than to statistical analysis or research. The decision of what and how to present information is the primary focus of administrative crime analysis. Often, the type of information that is presented represents the “tip of the iceberg” of all the work and analysis that has previously been done, for example, an executive summary of a report. The purpose and the audience of the information determine “what” is presented along with legal (e.g., privacy and confidentiality), political (e.g., union issues, election concerns), andpractical concerns (e.g., complexity of the information presented). The primary purpose of administrative crime analysis is to inform audiences. These audiences may vary from one situation to the next,

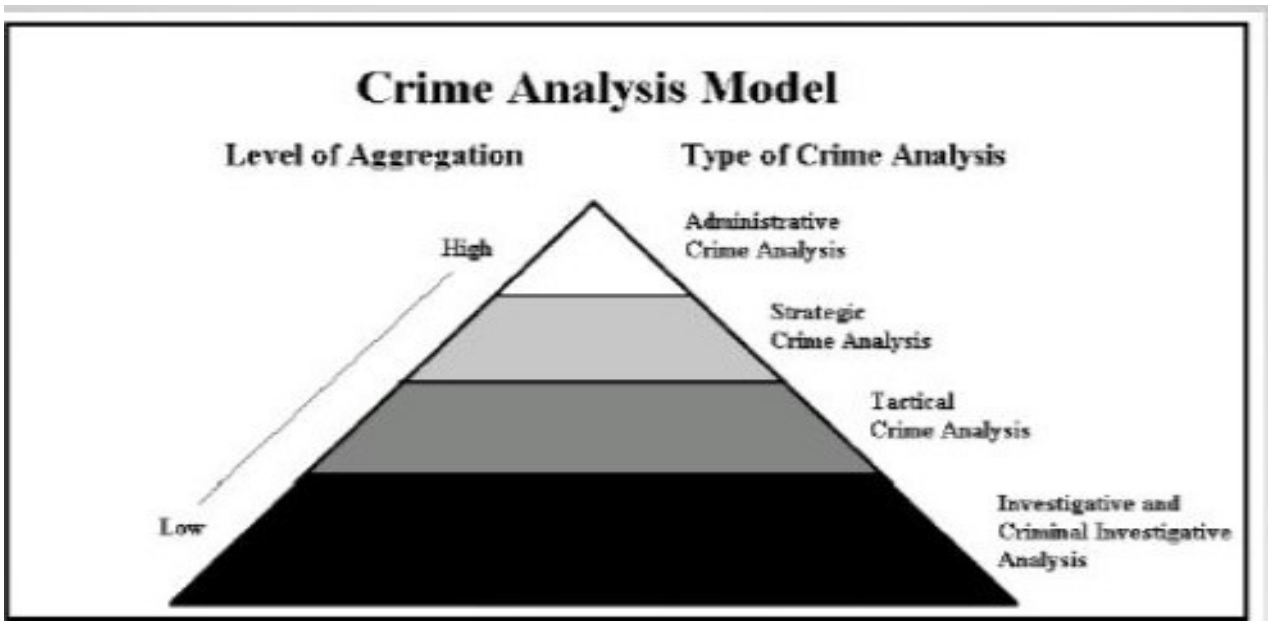
which is why the type and quantity of information should vary as well. Audiences can be police executives, city council, media, citizens, and neighborhood groups or a combination.

4

An excellent example of administrative crime analysis is the use of the Internet to provide information to the general public. Audiences of a police Internet site include citizens, police personnel, businesses, victims, criminals, and media—essentially everyone; therefore, the type of information published should be appropriate for an array of diverse customers. The information provided should be simple, clear, and concise and should not disclose sensitive information. One rule of thumb would be to only publish information that one would be comfortable seeing on the evening news.

3.1.3 Crime Analysis Model:

The following figure displays how all of these types of crime analysis relate to one another in terms of the level of aggregation of the information. That is, types with low levels of aggregation focus on individual cases and used qualitative data and analysis techniques and those with high levels of aggregation focus on a limited scope of larger amounts of data and information. At the top of the figure, criminal investigative analysis and intelligence analysis utilize the least aggregated and most qualitative data. The data consist of information about informal networks of criminals and their non-criminal acquaintances and relatives as well as where individuals live, work, and "play." The focus here is on the specifics of criminals, the nature of their crimes, their relationships, and their lives in general.



5

Tactical crime analysis utilizes only crimes and activity reported to the police so the data are more aggregate and somewhat less abundant than those used for criminal investigative and intelligence analysis. Tactical crime analysis is primarily qualitative in nature but depending on the data, quantitative techniques can be used to describe characteristics of a given pattern such as the most common time the crimes occur (time series) or where the crimes are located in relationship to one another. Strategic crime analysis utilizes large amounts of data that are even more aggregated than tactical and investigative data. For example, information used in tactical crime analysis is primarily made up of crime incidents but includes such information as date, time, location, methods of the crime, and detailed description of the crime. Strategic crime analysis focuses only on those variables that can be easily quantified, such as date, time, location, type of location, type of crime, and priority.

3.2 BIG DATA & HADOOP:-

3.2.1. BIG DATA-

Big Data is a phrase used to mean a massive volume of both structured and unstructured data that is so large it is difficult to process using traditional database and software techniques. In most enterprise scenarios the volume of data is too big or it moves too fast or it exceeds current processing capacity.

The three V's" – "volume, velocity and variety," concepts in big data originally coined by Doug Laney in 2001 (PDF) to refer to the challenge of data management. In short, it's a lot of data produced very quickly in many different forms. This could involve customer transactional histories, production databases, web traffic logs, online videos, social media interactions, and so forth.

3.2.1. HADOOP(HDFS)-

Hadoop is an open source, Java-based programming framework that supports the processing and storage of extremely large data sets in a distributed computing environment

The Hadoop Distributed File System (HDFS) is designed to **store** very large **data sets** reliably, and to stream those **data sets** at high bandwidth to user applications. In a large **cluster**, thousands of servers both host directly attached storage and execute user application tasks.

3.2.1.1 Characteristics of Hadoop Eco-System:-

1. Flexible:

As it is a known fact that only 20% of data in organizations is structured, and the rest is all unstructured, it is very crucial to manage unstructured data which goes unattended. Hadoop manages different types of Big Data, whether structured or unstructured, encoded or formatted, or any other type of data and makes it useful for decision making process. Moreover, Hadoop is simple, relevant and schema-less! Though Hadoop generally supports **Java Programming**, any programming language can be used in Hadoop with the help of the **MapReduce technique**. Though Hadoop works best on **Windows** and **Linux**, it can also work on other operating systems like **BSD** and **OS X**.

2. Scalable

Hadoop is a scalable platform, in the sense that new nodes can be easily added in the system as and when required without altering the data formats, how data is loaded, how programs are written, or even without modifying the existing applications. Hadoop is an open source platform and runs on industry-standard hardware. Moreover, Hadoop is also fault tolerant – this means, even if a node gets lost or goes out of service, the system automatically reallocates work to another location of the data and continues processing as if nothing had happened!

3. Building more efficient data economy:

Hadoop has revolutionized the processing and analysis of big data world across. Till now, organizations were worrying about how to manage the non-stop data overflowing in their systems. Hadoop is more like a "Dam", which is harnessing the flow of unlimited amount of data and generating a lot of power in the form of relevant information. Hadoop has changed the economics of storing and evaluating data entirely!

4. Robust Ecosystem:

Hadoop has a very robust and a rich ecosystem that is well suited to meet the analytical needs of developers, web start-ups and other organizations. Hadoop Ecosystem consists of various related projects such as MapReduce, Hive, HBase, Zookeeper, HCatalog, Apache Pig, which make Hadoop very competent to deliver a broad spectrum of services.

5. Hadoop is getting more "Real-Time"!

Did you ever wonder how to stream information into a cluster and analyze it in real time? Hadoop has the answer for it. Yes, Hadoop's competencies are getting more and more real-time. Hadoop also provides a standard approach to a wide set of APIs for big data analytics comprising MapReduce, query languages and database access, and so on.

6. Cost Effective:

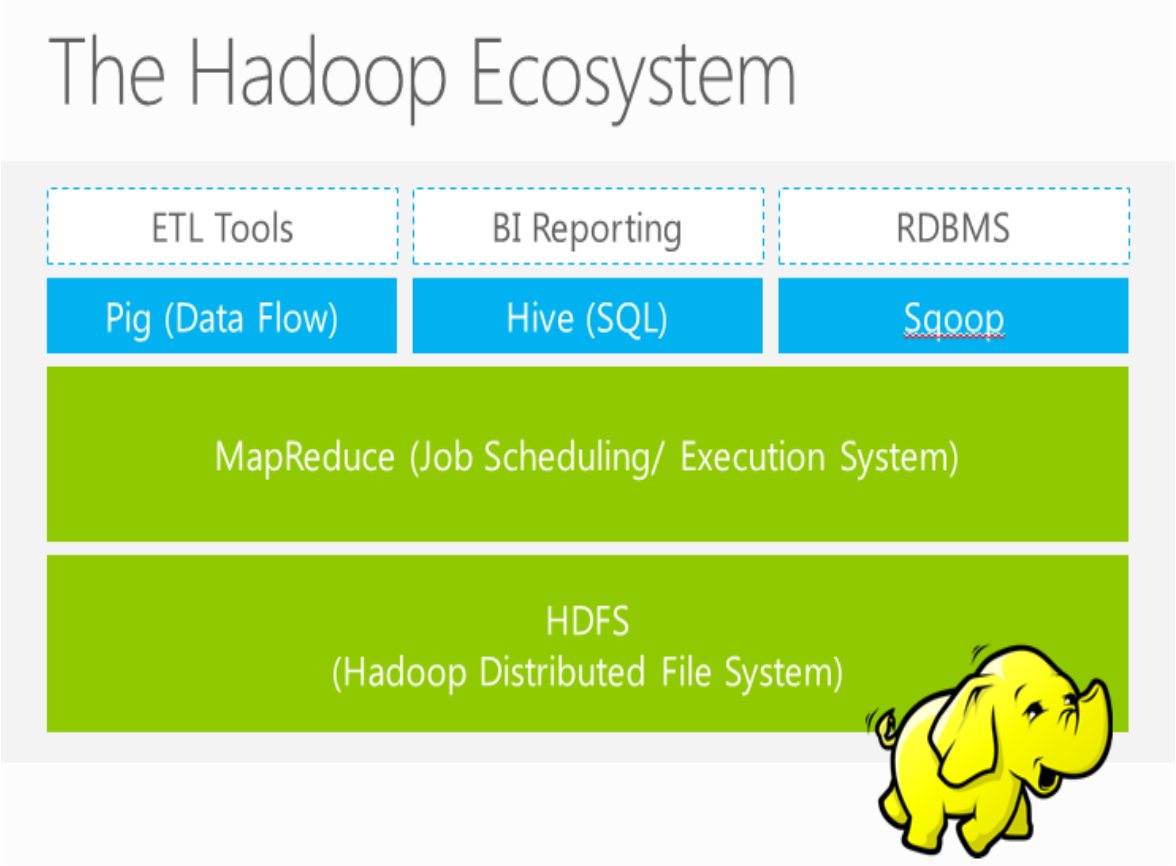
Loaded with such great features, the icing on the cake is that Hadoop generates cost benefits by bringing massively parallel computing to commodity servers, resulting in a substantial reduction in the cost per terabyte of storage, which in turn makes it reasonable to model all your data. The basic idea behind Hadoop is to perform cost-effective data analysis present across world wide web!

7. Upcoming Technologies using Hadoop:

With reinforcing its capabilities, Hadoop is leading to phenomenal technical advancements. For instance, HBase will soon become a vital Platform for Blob Stores (Binary Large Objects) and for Lightweight OLTP (Online Transaction Processing). Hadoop has also begun serving as a strong foundation for new-school graph and NoSQL databases, and better versions of relational databases.

8. Hadoop is getting cloudy!

Hadoop is getting cloudier! In fact, cloud computing and Hadoop are synchronizing in several organizations to manage Big Data. Hadoop will become one of the most required apps for cloud computing. This is evident from the number of Hadoop clusters offered by cloud vendors in various businesses. Thus, Hadoop will reside in the cloud soon!



PIG SCRIPTING AND MAPREDUCE

4.1. PIG:-

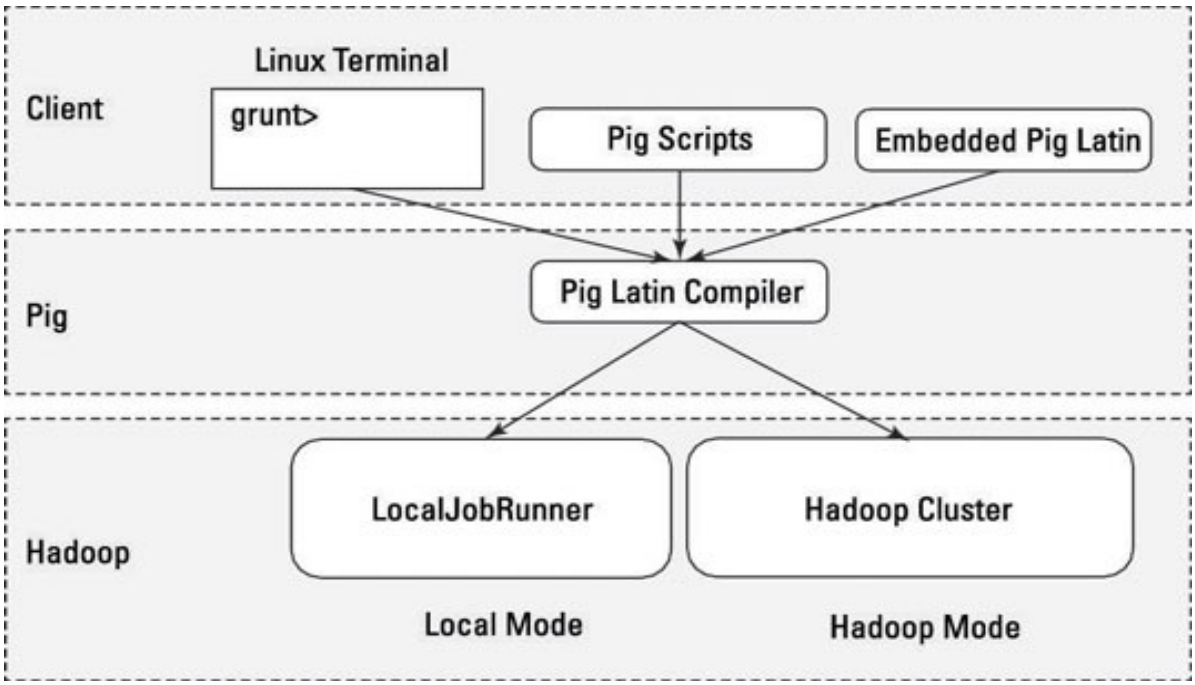
Pig is a high level scripting language that is used with Apache Hadoop. Pig enables data workers to write complex data transformations without knowing Java. Pig's simple SQL-like scripting language is called Pig Latin, and appeals to developers already familiar with scripting languages and SQL.

Pig is complete, so you can do all required data manipulations in Apache Hadoop with Pig. Through the User Defined Functions(UDF) facility in Pig, Pig can invoke code in many languages like JRuby, Jython and Java. You can also embed Pig scripts in other languages. The

result is that you can use Pig as a component to build larger and more complex applications that tackle real business problems.

Pig works with data from many sources, including structured and unstructured data, and store the results into the Hadoop Data File System.

Pig scripts are translated into a series of MapReduce jobs that are run on the Apache Hadoop cluster.



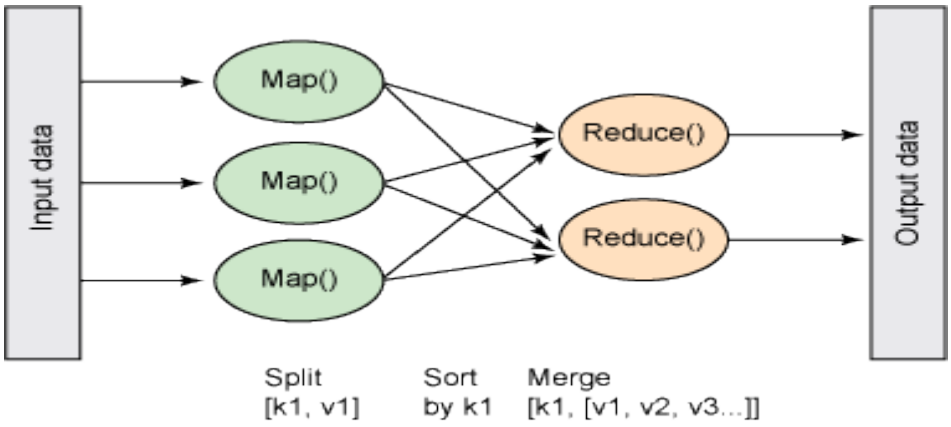
Local mode	MapReduce mode
$ pig -x local **Sample_script.pig**	$ pig -x mapreduce **Sample_script.pig**

4.2 MAP REDUCE:-

MapReduce is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster.

A MapReduce program is composed of a Map() procedure (method) that performs filtering and sorting (such as sorting students by first name into queues, one queue for each name) and a Reduce() method that performs a summary operation (such as counting the number of students in each queue, yielding name frequencies). The "MapReduce System" (also called "infrastructure" or "framework") orchestrates the processing by marshalling the distributed servers, running the various tasks in parallel, managing all communications and data transfers between the various parts of the system, and providing for redundancy and fault tolerance.

The model is a specialization of the split-apply-combine strategy for data analysis.It is inspired by the map and reduce functions commonly used in functional programming,although their purpose in the MapReduce framework is not the same as in their original forms.. As such, a single-threaded implementation of MapReduce will usually not be faster than a traditional (non-MapReduce) implementation; any gains are usually only seen with multi-threaded implementations.The use of this model is beneficial only when the optimized distributed shuffle operation (which reduces network communication cost) and fault tolerance features of the MapReduce framework come into play. Optimizing the communication cost is essential to a good MapReduce algorithm.



PIG VS HIVE VS HBASE

PIG - It is a workflow language and it has its own scripting language called Pig Latin. Pig is one of the alternatives for MapReduce but NOT the exact replacement. It works good with both

structured and unstructured data. Script is written in a special shell called grunt. Default mode is mapreduce and have actually two modes(local and mapreduce) for writing the scripts.

HIVE- Another alternative to avoid tedious MapReduce codes. It's a datawarehouse language, more like SQL. Works better with structured data and has a metadata table on top of HDFS. Supports queries with some awesome features like partitioning and bucketing which helps in query performance. Easy to write and query/process data. External and internal tables can be created which is not supported by Pig. Pig will store data temporarily until it is 'stored' or saved to HCatalog.

HBASE - It's a NoSQL database (no relation with Hive and Pig) on top of HDFS. It is column oriented, good for OLAP operations as row level update is not possible. Has a special architecture consisting of Master Server,Region Server, Hfile, HMaster,WAL, etc. Supports random read across data which normal HDFS can't provide

SOURCE OF DATA

Here we have used Govt. Data Set to perfrom the Crime Data Analysis Project.

(Here is the link of used data Set)

https://edureka.wistia.com/medias/s6rcbu3pgk/download?media_file_id=67589933

6.1 Data Set Desciption

Then data represents the Crime locations and Crime age details which includes the below fields:

» State

»Crime head

»Year

» Male below 18 years »

Female below 18 years. »

Male Between 18-30 Years

» Female Between 18-30 Years

» Male Between 30-45 Years

» Female Between 30-45 Years

» Male Between 45-60 Years

» Female Between 45-60 Years

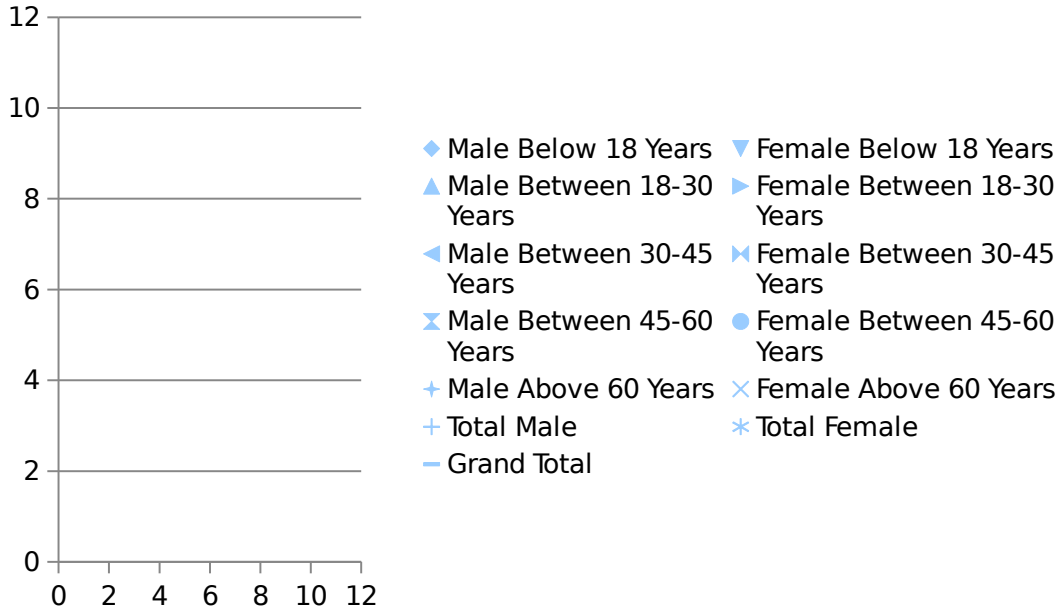
» Male Above 60 Years

» Female Above 60 Years

» Total Male

» Total Female

6.2 Graphical Representation(crime data set)



SPECIFICATIONS & USAGE OF SOFTWARES:-

** * Oracle VM Virtual Box*

Specification

Processor Type -64 bit

Minimum Processor Speed -1.83 GHz*2

Minimum Memory - 8.0 GB

Minimum Swap Space - 2.1 GB

Minimum Disk Space - 5.5 GB in /u01; 3 GB in /tmp; 400MB in /var; 300 MB in /usr

Operating System - Oracle Linux 5 Update 5 64-bit or later; Oracle Linux 6 64-bit or later;Oracle Linux 7 64-bit or later RHEL 5 Update 5 64-bit or later; RHEL 6 64-bit or later; RHEL 7 64-bit or later.

Web Browser - Mozilla Firefox 8 or later; Microsoft Internet Explorer 10 or later; Apple Safari 6 and above; Google Chrome 15 and above

Usages:

VirtualBox is a general-purpose full virtualizer for x86 hardware, targeted at server, desktop and embedded use. VirtualBox may be installed on a number of host operating systems, including: Linux, macOS, Windows, Solaris, and OpenSolaris. It supports the creation and management of guest virtual machines running versions and derivations of Windows, Linux, BSD, OS/2, Solaris, Haiku, OSx86 and others,[7] and limited virtualization of macOS guests on Apple hardware

PROBLEM STATEMENTS:-

1. Show the states Involved in crimes

2. Total number of crimes occurred in each state.

3. Type and number of crimes occurred in each state.

4. Calculate the number of females (age between 18-30 years) who were victims in different crimes in different states.

EXECUTION PROCESS

Step 1:

>Register the jar file.

Grunt>REGISTER '/home/edureka/Downloads/piggybank.jar';

>Load the govt crime data set.

a = LOAD '/home/edureka/Downloads/crimedata.csv' USING org.apache.pig.piggybank.storage.CSVLoader ( ) AS (STATE: chararray, Crimehead: bytearray, MaleBelow18Years: bytearray,FemaleBelow18Years:bytearray,MaleBetween18_30Years:bytearray,FemaleBetween18_30Years:bytearray,MaleBetween30_45Years:bytearray,FemaleBetween30_45Years:bytearray,MaleBetween45_60Years:bytearray,FemaleBetween45_60Years:bytearray,MaleAbove60Years:bytearray,FemaleAbove60Years:bytearray,TotalMale:bytearray,TotalFemale:bytearray,GrandTotal:bytearray);

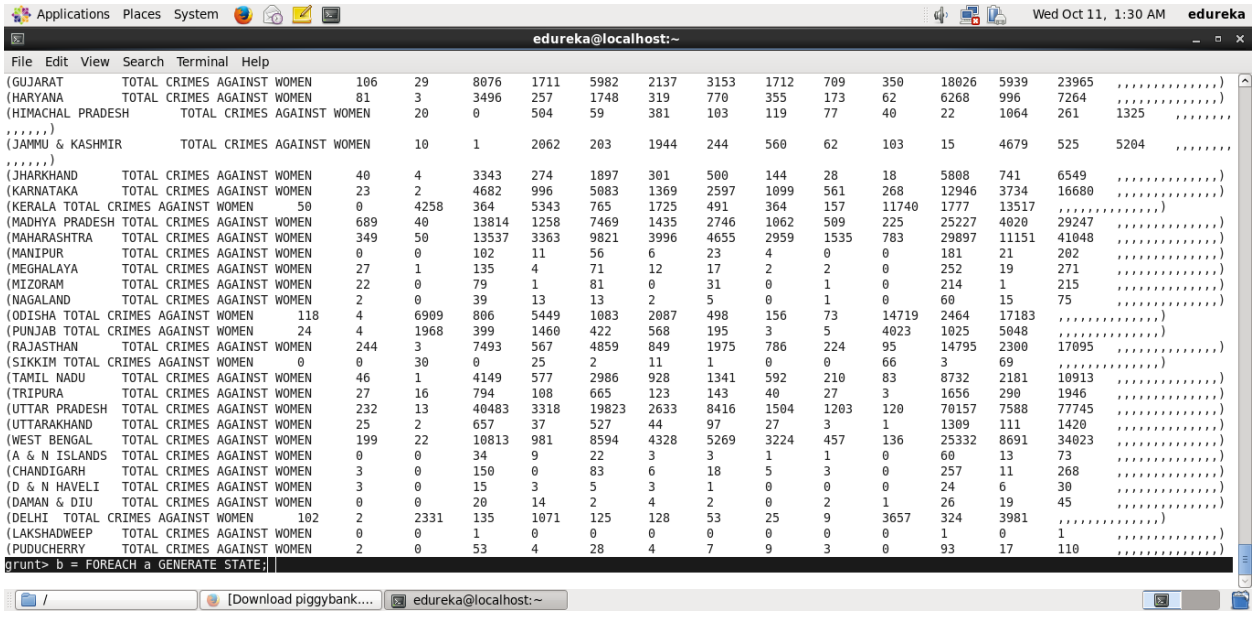
Grunt>dump a;

```
(HARYANA       TOTAL CRIMES AGAINST WOMEN      81      3       3496    257     1748    319     770     355     173     82      6268    996     7
(HIMACHAL PRADESH       TOTAL CRIMES AGAINST WOMEN      20      0       504     59      381     103     119     77      40      22      1064    2
,,,,,,)
(JAMMU & KASHMIR        TOTAL CRIMES AGAINST WOMEN      10      1       2062    203     1944    244     560     62      103     15      4679    5
,,,,,,)
(JHARKHAND      TOTAL CRIMES AGAINST WOMEN      40      4       3343    274     1897    301     500     144     28      18      5808    741     6
(KARNATAKA      TOTAL CRIMES AGAINST WOMEN      23      2       4682    996     5083    1369    2597    1099    561     268     12946   3734    1
(KERALA TOTAL CRIMES AGAINST WOMEN      50      0       4258    364     5343    765     1725    491     364     157     11740   1777    13517   ,
(MADHYA PRADESH TOTAL CRIMES AGAINST WOMEN      689     40      13814   1258    7469    1435    2746    1062    509     225     25227   4020    2
(MAHARASHTRA    TOTAL CRIMES AGAINST WOMEN      349     50      13537   3363    9821    3996    4655    2959    1535    783     29897   11151   4
(MANIPUR        TOTAL CRIMES AGAINST WOMEN      0       0       102     11      56      6       23      4       0       0       181     21      2
(MEGHALAYA      TOTAL CRIMES AGAINST WOMEN      27      1       135     4       71      12      17      2       2       0       252     19      2
(MIZORAM        TOTAL CRIMES AGAINST WOMEN      22      0       79      1       81      0       31      0       1       0       214     1       2
(NAGALAND       TOTAL CRIMES AGAINST WOMEN      2       0       39      13      13      2       5       0       1       0       60      15      7
(ODISHA TOTAL CRIMES AGAINST WOMEN      118     4       6909    806     5449    1083    2087    498     156     73      14719   2464    17183   ,
(PUNJAB TOTAL CRIMES AGAINST WOMEN      24      4       1968    399     1460    422     568     195     3       5       4023    1025    5048    ,
(RAJASTHAN      TOTAL CRIMES AGAINST WOMEN      244     3       7493    567     4859    849     1975    786     224     95      14795   2300    1
(SIKKIM TOTAL CRIMES AGAINST WOMEN      0       0       30      0       25      2       11      1       0       0       66      3       69      ,
(TAMIL NADU     TOTAL CRIMES AGAINST WOMEN      46      1       4149    577     2986    928     1341    592     210     83      8732    2181    1
(TRIPURA        TOTAL CRIMES AGAINST WOMEN      27      16      794     108     665     123     143     40      27      3       1656    290     1
(UTTAR PRADESH  TOTAL CRIMES AGAINST WOMEN      232     13      40483   3318    19823   2633    8416    1504    1203    120     70157   7588    7
(UTTARAKHAND    TOTAL CRIMES AGAINST WOMEN      25      2       657     37      527     44      97      27      3       1       1309    111     1
(WEST BENGAL    TOTAL CRIMES AGAINST WOMEN      199     22      10813   981     8594    4328    5269    3224    457     136     25332   8691    3
(A & N ISLANDS  TOTAL CRIMES AGAINST WOMEN      0       0       34      9       22      3       3       1       1       0       60      13      7
(CHANDIGARH     TOTAL CRIMES AGAINST WOMEN      3       0       150     0       83      6       18      5       3       0       257     11      2
(D & N HAVELI   TOTAL CRIMES AGAINST WOMEN      3       0       15      3       5       3       1       0       0       0       24      6       3
(DAMAN & DIU    TOTAL CRIMES AGAINST WOMEN      0       0       20      14      2       4       2       0       2       1       26      19      4
(DELHI  TOTAL CRIMES AGAINST WOMEN      102     2       2331    135     1071    125     128     53      25      9       3657    324     3981    ,
(LAKSHADWEEP    TOTAL CRIMES AGAINST WOMEN      0       0       1       0       0       0       0       0       0       0       1       0       1
```

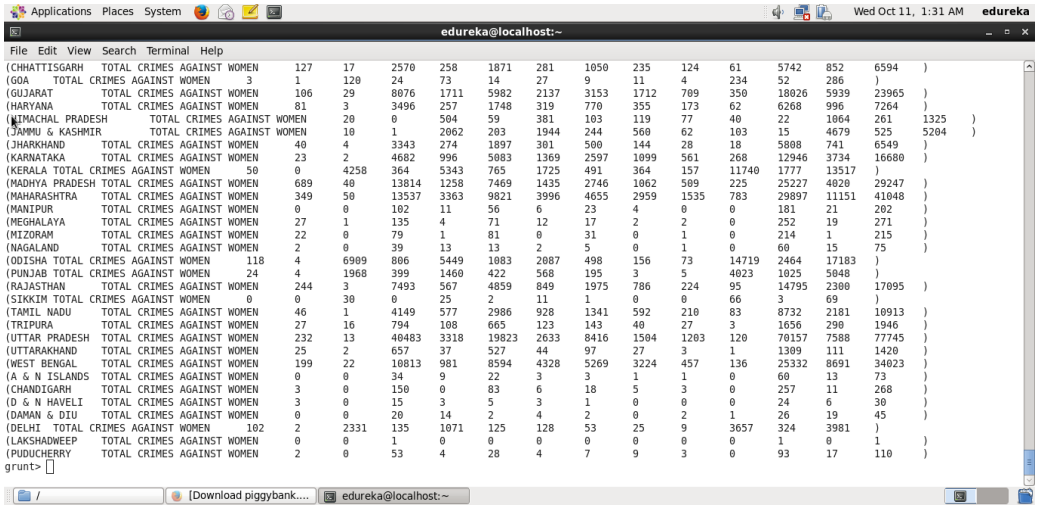
Step2

1. **Show the states Involved in crimes**

Grunt> b = FOREACH a GENERATE STATE;



Grunt>dump b;

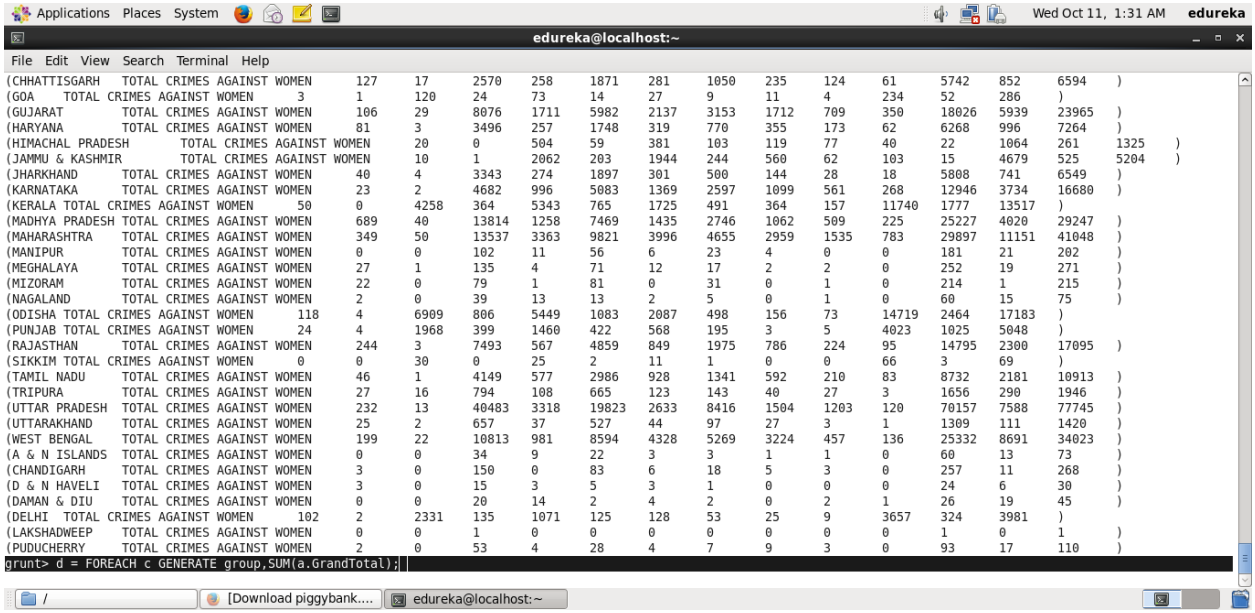


18

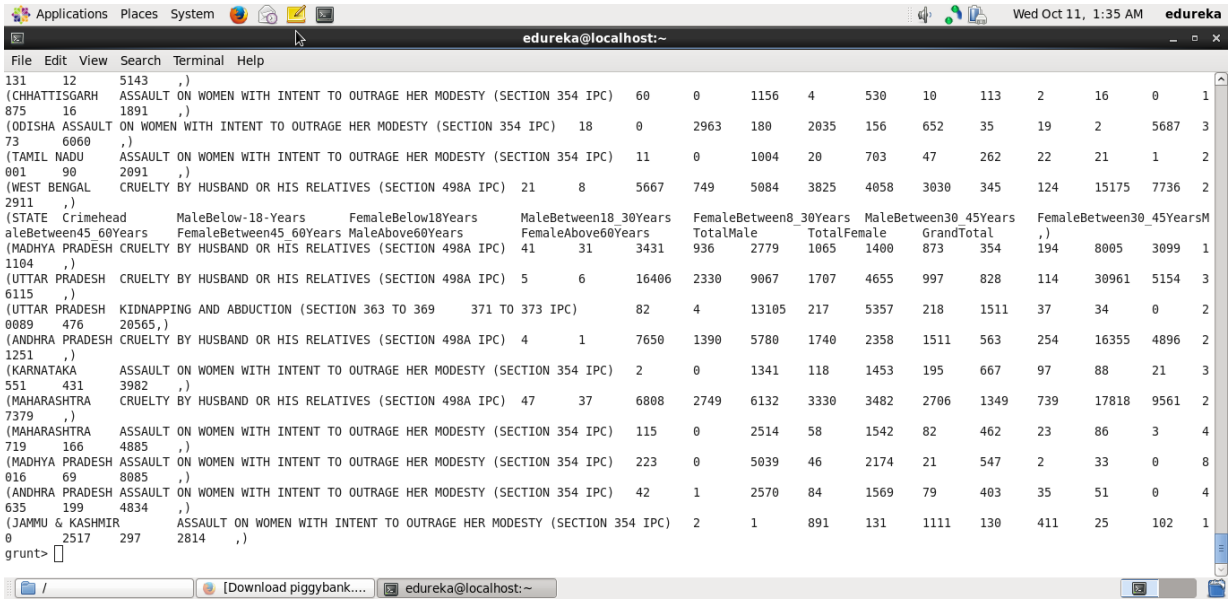
Step 3.

Total number of crimes occurred in each state.

Grunt>d = FOREACH c GENERATE group,SUM(a.GrandTotal);



Grunt>dump d;

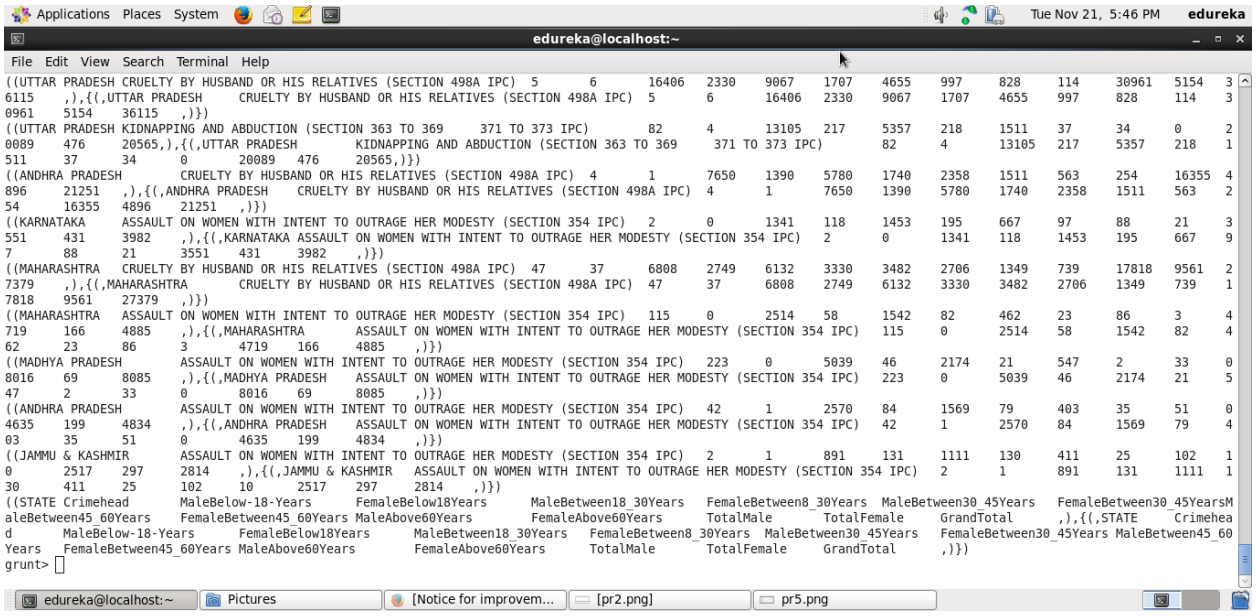


step 4:

Types and number of crimes occurred in each state.

c2 = GROUP c1 BY (STATE,Crimehead);

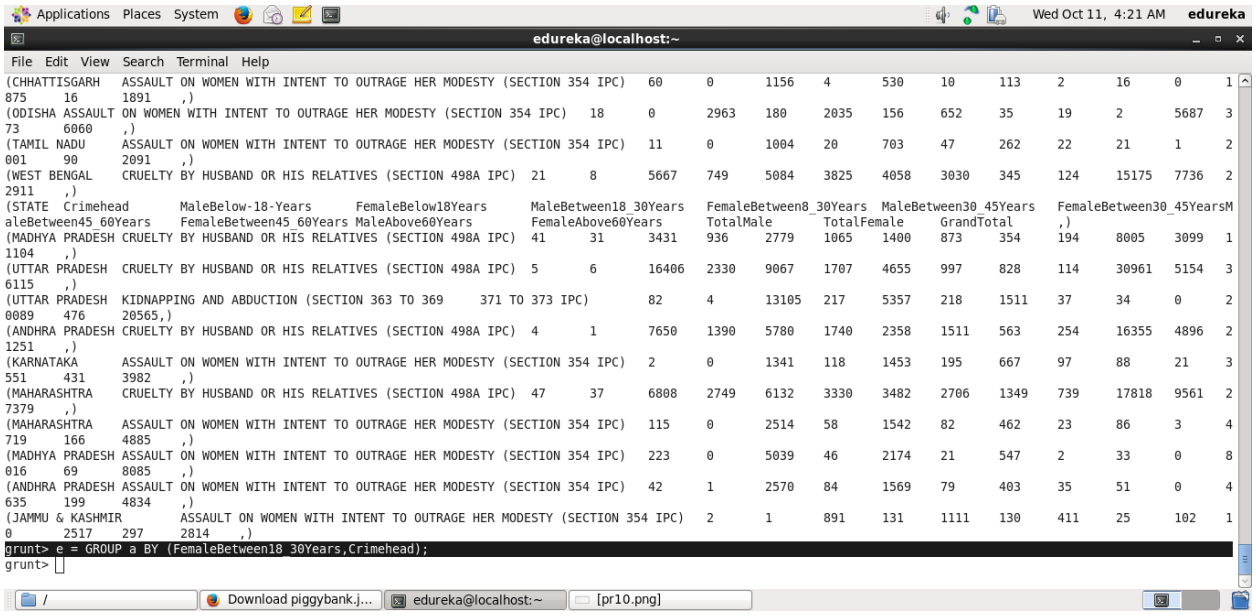
Grunt>dump c2;



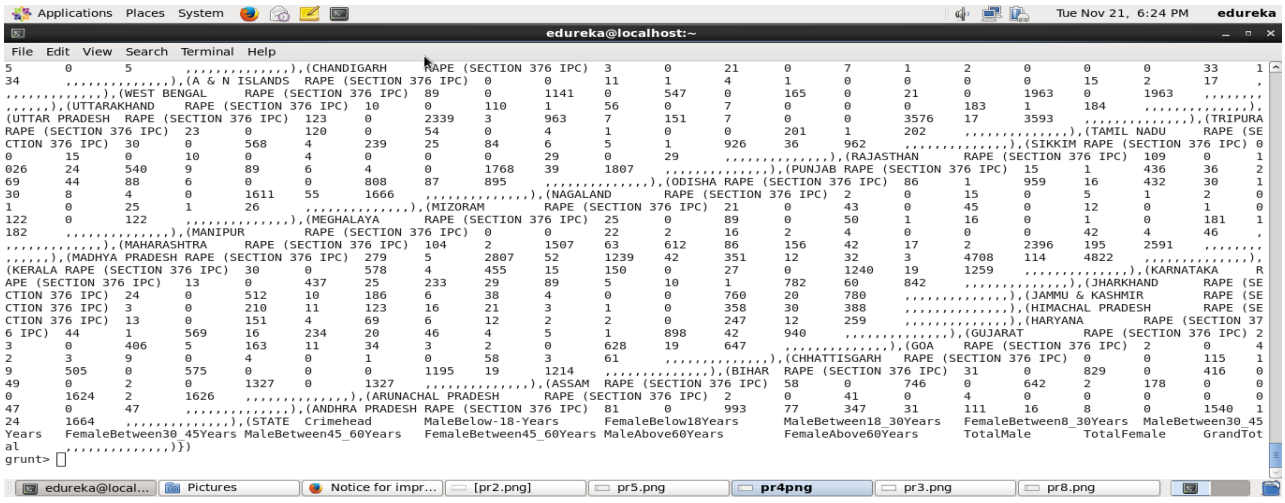
step 5:

Calculate the number of females (age between 18-30 years) who were victims in different crimes in different states.

Grunt>e = GROUP a BY (FemaleBetween18_30Years,Crimehead);


Applications Places System
Wed Oct 11, 4:21 AM edureka
edureka@localhost:~
File Edit View Search Terminal Help
(CHHATTISGARH ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY (SECTION 354 IPC) 60 0 1156 4 530 10 113 2 16 0 1
875 16 1891 ,)
(ODISHA ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY (SECTION 354 IPC) 18 0 2963 180 2035 156 652 35 19 2 5687 3
73 6060 ,)
(TAMIL NADU ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY (SECTION 354 IPC) 11 0 1004 20 703 47 262 22 21 1 2
001 90 2091 ,)
(WEST BENGAL CRUELTY BY HUSBAND OR HIS RELATIVES (SECTION 498A IPC) 21 8 5667 749 5084 3825 4058 3030 345 124 15175 7736 2
2911 ,)
(STATE Crimehead MaleBelow-18-Years FemaleBelow18Years MaleBetween18_30Years FemaleBetween8_30Years MaleBetween30_45Years FemaleBetween30_45YearsM
aleBetween45_60Years FemaleBetween45_60Years MaleAbove60Years FemaleAbove60Years TotalMale TotalFemale GrandTotal ,)
(MADHYA PRADESH CRUELTY BY HUSBAND OR HIS RELATIVES (SECTION 498A IPC) 41 31 3431 936 2779 1065 1400 873 354 194 8005 3099 1
1104 ,)
(UTTAR PRADESH CRUELTY BY HUSBAND OR HIS RELATIVES (SECTION 498A IPC) 5 6 16406 2330 9067 1707 4655 997 828 114 30961 5154 3
6115 ,)
(UTTAR PRADESH KIDNAPPING AND ABDUCTION (SECTION 363 TO 369 371 TO 373 IPC) 82 4 13105 217 5357 218 1511 37 34 0 2
0089 476 20565,)
(ANDHRA PRADESH CRUELTY BY HUSBAND OR HIS RELATIVES (SECTION 498A IPC) 4 1 7650 1390 5780 1740 2358 1511 563 254 16355 4896 2
1251 ,)
(KARNATAKA ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY (SECTION 354 IPC) 2 0 1341 118 1453 195 667 97 88 21 3
551 431 3982 ,)
(MAHARASHTRA CRUELTY BY HUSBAND OR HIS RELATIVES (SECTION 498A IPC) 47 37 6808 2749 6132 3330 3482 2706 1349 739 17818 9561 2
7379 ,)
(MAHARASHTRA ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY (SECTION 354 IPC) 115 0 2514 58 1542 82 462 23 86 3 4
719 166 4885 ,)
(MADHYA PRADESH ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY (SECTION 354 IPC) 223 0 5039 46 2174 21 547 2 33 0 8
016 69 8085 ,)
(ANDHRA PRADESH ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY (SECTION 354 IPC) 42 1 2570 84 1569 79 403 35 51 0 4
635 199 4834 ,)
(JAMMU & KASHMIR ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY (SECTION 354 IPC) 2 1 891 131 1111 130 411 25 102 1
0 2517 297 2814 ,)
grunt> e = GROUP a BY (FemaleBetween18_30Years,Crimehead);
grunt>
/
Download piggybank.j...
edureka@localhost:~
[pr10.png]

Grunt>dump e;


Applications Places System
Tue Nov 21, 6:24 PM edureka
edureka@localhost:~
File Edit View Search Terminal Help
grunt>
edureka@local...
Pictures
Notice for impr...
[pr2.png]
pr5.png
pr4png
pr3.png
pr8.png