ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 8, Issue 4, July 2019

COMPARISON OF HIVE AND PIG PERFORMANCE ON CRIME DATA ANALYSIS

Poonam Harode, Nitesh Kumar Gupta

Abstract—***The problems faced by every country to reduce crime rate is not unique. But their large amount of crime record data and incomplete information creates a problem for analytics to analyze these data. So analyzing the complex data by using traditional tools and techniques is an expensive task. Instead of using traditional data analysis techniques it would be beneficial to use Big Data Analytics for that huge. In this paper we introduces bigdata analytics using pig and hive sheds light on significant issues faced by government for making decision to reduce the crime rate by analyzing the huge and large crime datasets with the help of bigdata analytical tools we can find the crime rate by year, district wise and the type of crimes.***

Keywords— **Big data, Hadoop, hive, pig, analysis, crime analysis.**

I. INTRODUCTION

With continually increasing population, crimes and crime rate analyzing related data is a huge issue for governments to make strategic decisions so as to maintain law and order. This is really necessary to keep the citizens of the country safe from crimes. In this growing field of technology, rate of cyber-crimes is increasing and are challenging the capabilities of investigation people. The data generation regarding crime is also increased nowadays which is mostly digital in nature.

Nowadays generated data cannot be handled efficiently with the use of traditional analysis techniques. Instead of using traditional data analysis techniques it would be beneficial to use Big Data Analytics for that huge data. Primarily collected data will be distributed over geographic location and based on that clusters will be created. In second phase the created clusters are analyzed using Big Data Analytics.

With continually increasing population, crimes and crime rate analyzing related data is a huge issue for governments to make strategic decisions so as to maintain law and order. The problems faced by every country to reduce crime rate is not unique. But their large amount of crime record data and incomplete information creates a problem for analytics to analyze these data. So analyzing the complex data by using traditional tools and techniques is an expensive task. Instead of using traditional data analysis techniques it would be beneficial to use Big Data Analytics for that huge. Typical attributes of crime data can be identified by three main attributes:

1. Volume – Crime data is huge and massive.
2. Velocity –Crime data is changes rapidly and arrives quickly so processing data in less time is very difficult.
3. Variety – Crime data have the different structure they are semi-structured or unstructured data.

II. LITERATURE REVIEW

In [1] present the log data analysis by using spark through sql type queries, the web server logs and unstructured in nature and these data are analyze by using spark and hadoop framework, and they also compare both the framework on the basis of various parameters. In [2] the last decade profitability change is mainly driven by input price change which exhibits a similar pattern to output price change. In presence of productivity growth, the output price increase is lower than the input price increase suggesting that part of productivity gains are transferred from airlines to consumers.

III. PROBLEM DEFINITION

Every country continuously trying to reduce their crime rate market but the problems faced by countries is not unique. But their large amount of unstructured data and incomplete information creates a problem for analytics to analyze these data. Traditional tools and techniques create the problem for storing and analyzing the huge amount of airlines data called big data [2] because of its nature so it’s the biggest challenges in big data to store and process such huge amount data [10].

IV. PROPOSED WORK

We are using Hadoop[5] for storing and processing a huge amount of data, For storing its uses HDFS (Hadoop Distributed File System) and for processing its uses MapReduce[8]. For analyzing we are using hive which uses hive QL statement which runs over MapReduce framework to analyze the data. And we can also analyze the data by using pig [4] which uses pig latin language and runs over map reduce framework.

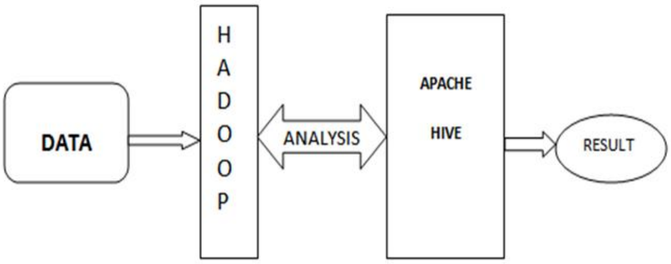


Fig 1. Workflow Diagram

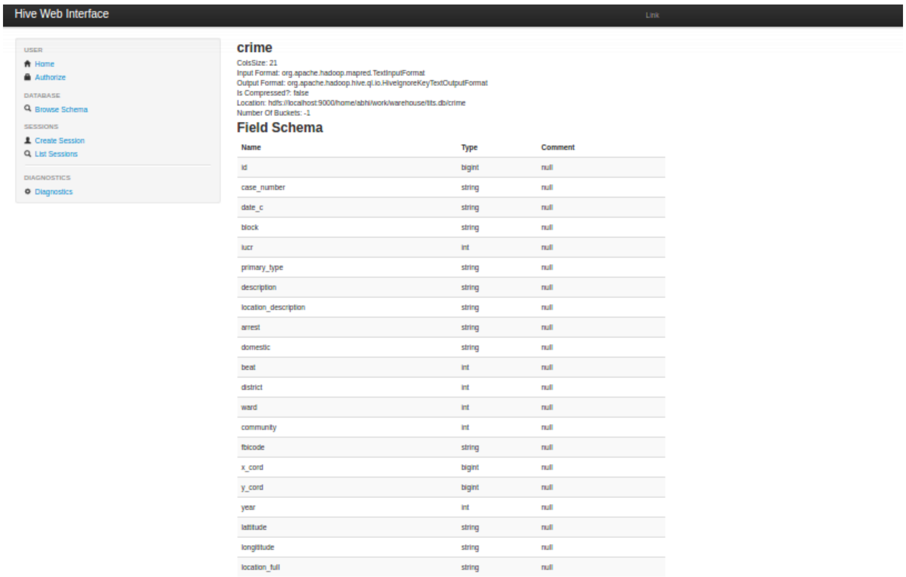
V. RESEARCH QUESTIONS

Some of the problem statements along which the analysis has been done in this paper:-

- Total Number of crimes in a year.
- Total number of crime types per year
- Total number of crime in each district
- Total number of ‘NARCOTICS’ cases filed in each year
- Total number of ‘THEFT’ cases filed in each year
- Total number of ‘GAMBLING’ cases filed in each year
- Performance comparison of hive and pig

Data Set: The data was in .csv format, that is each line represents data record and each record has one or more field separated by commas.

Data Set Description: The data set includes the following fields



crime

ColsSize: 21
Input Format: org.apache.hadoop.mapred.TextInputFormat
Output Format: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Is Compressed?: false
Location: hdfs://localhost:9000/home/abhi/work/warehouse/tits.db/crime
Number Of Buckets: -1

Field Schema

Name	Type	Comment
id	bigint	null
case_number	string	null
date_c	string	null
block	string	null
iucr	int	null
primary_type	string	null
description	string	null
location_description	string	null
arrest	string	null
domestic	string	null
beat	int	null
district	int	null
ward	int	null
community	int	null
fbicode	string	null
x_cord	bigint	null
y_cord	bigint	null
year	int	null
lattitude	string	null
longitiude	string	null
location_full	string	null

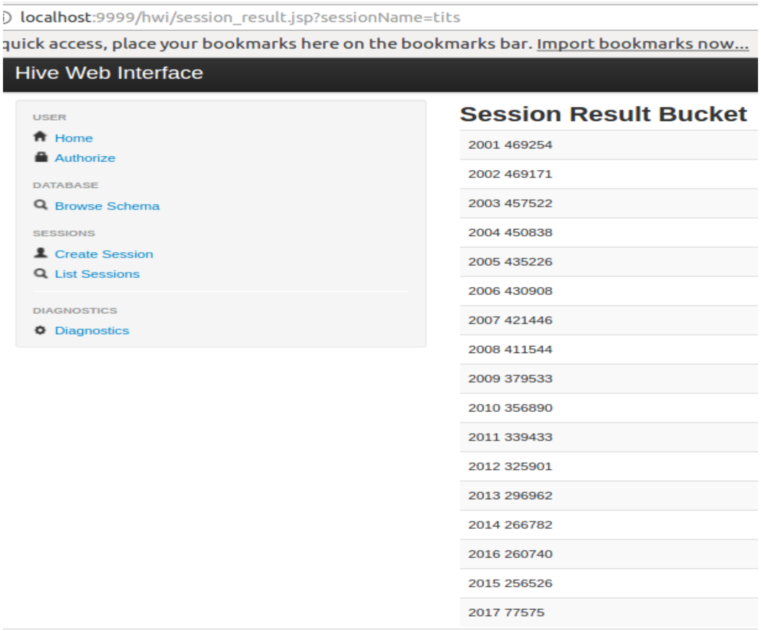
Fig 2. Dataset Description

Tools and Technologies Used:
1. Hadoop
2. Hive Web Interface

VI. EXPERIMENTAL FINDINGS

We can configure hadoop-1.1.2 on ubuntu and along with hadoop we integrate hive and pig on top of the hadoop, from with the help of hive and pig we can analyze the crime datasets. After loading datasets into hdfs we can analyze the dataset by hive and pig , and both the analytical tools produces accurate result but both uses different programming paradigm to analyze the crime datasets. Hive Web Interface supports the Hive query language (HQL) and also support SQL Pig supports pig latin language which is a scripting language.

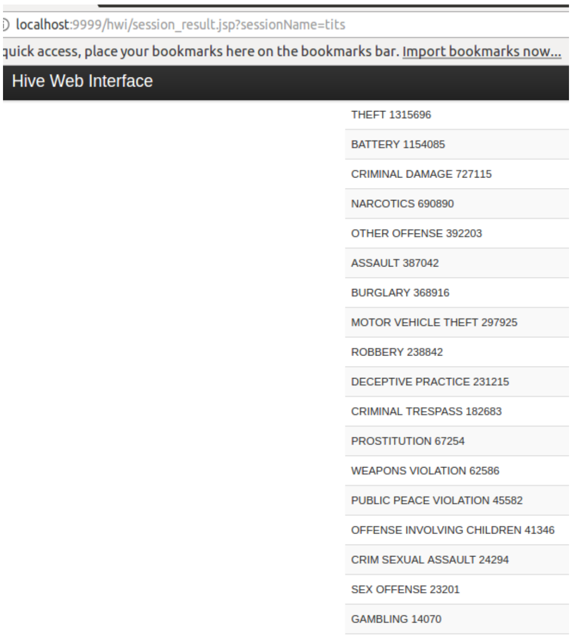
Both the big data analytical tools runs on top of the hadoop means both the tools have a capabilities to launched a map reduce job on background through which the crime data is processed and we can get the outcomes of that result. In this paper we collect 17 year crime data and by analyzing these datasets we find the number of crime per year, number of crime in each district, number of crime types along with its frequency.


localhost:9999/hwi/session_result.jsp?sessionName=tits
quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Hive Web Interface
USER
Home
Authorize
DATABASE
Browse Schema
SESSIONS
Create Session
List Sessions
DIAGNOSTICS
Diagnostics

Session Result Bucket

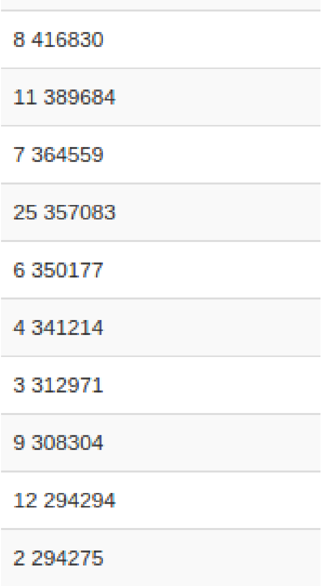
```
2001 469254
2002 469171
2003 457522
2004 450838
2005 435226
2006 430908
2007 421446
2008 411544
2009 379533
2010 356890
2011 339433
2012 325901
2013 296962
2014 266782
2016 260740
2015 256526
2017 77575
```

Fig 3. Hive Session result


localhost:9999/hwi/session_result.jsp?sessionName=tits
quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Hive Web Interface

```
THEFT 1315696
BATTERY 1154085
CRIMINAL DAMAGE 727115
NARCOTICS 690890
OTHER OFFENSE 392203
ASSAULT 387042
BURGLARY 368916
MOTOR VEHICLE THEFT 297925
ROBBERY 238842
DECEPTIVE PRACTICE 231215
CRIMINAL TRESPASS 182683
PROSTITUTION 67254
WEAPONS VIOLATION 62586
PUBLIC PEACE VIOLATION 45582
OFFENSE INVOLVING CHILDREN 41346
CRIM SEXUAL ASSAULT 24294
SEX OFFENSE 23201
GAMBLING 14070
```

Fig 4. Number of crime types



8	416830
11	389684
7	364559
25	357083
6	350177
4	341214
3	312971
9	308304
12	294294
2	294275

Fig 5. Number of crimes in each district

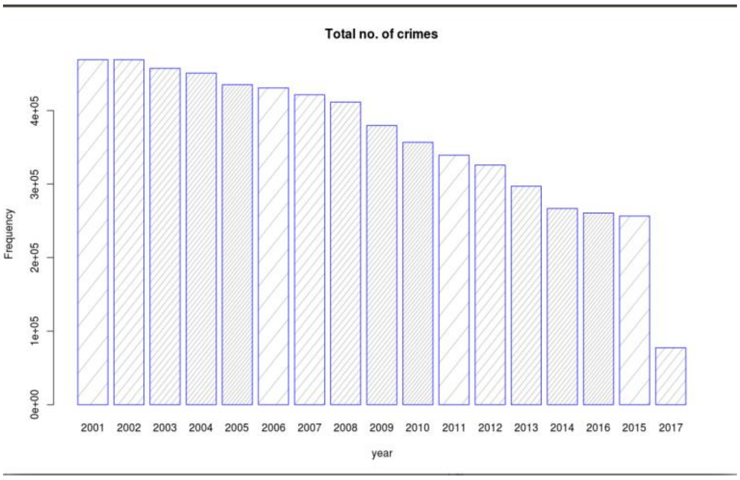


Fig 6. Number of crime in each year

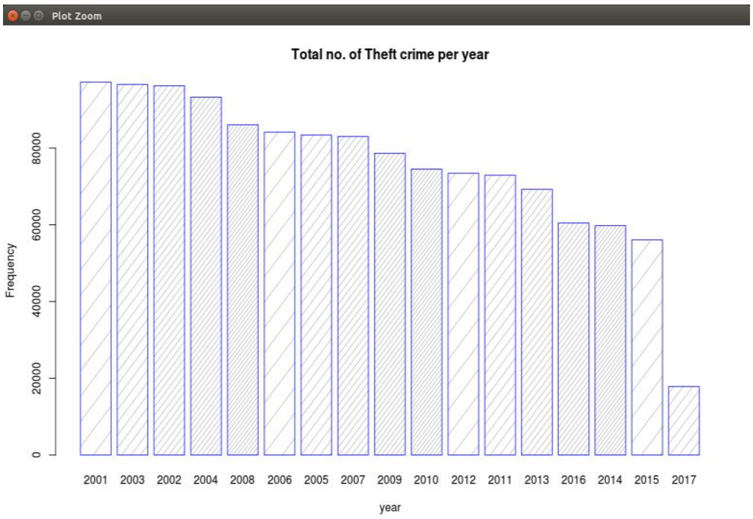


Fig 7. Total number of ‘THEFT’ cases filed in each year

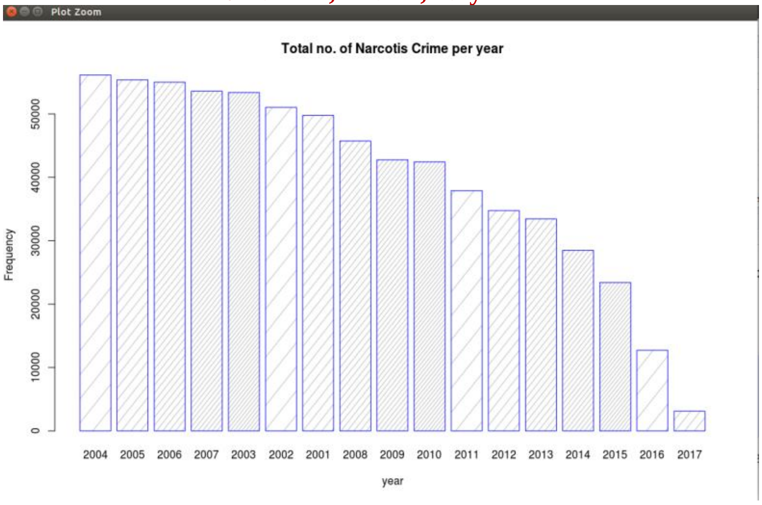


Fig 8. Total number of 'NARCOTIS' cases filed in each year

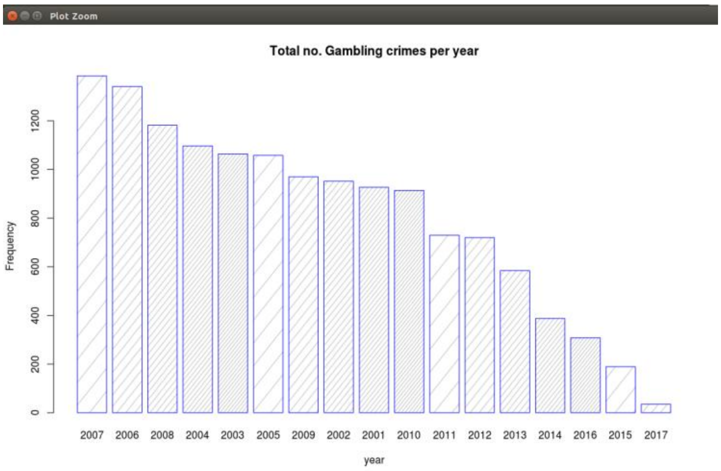


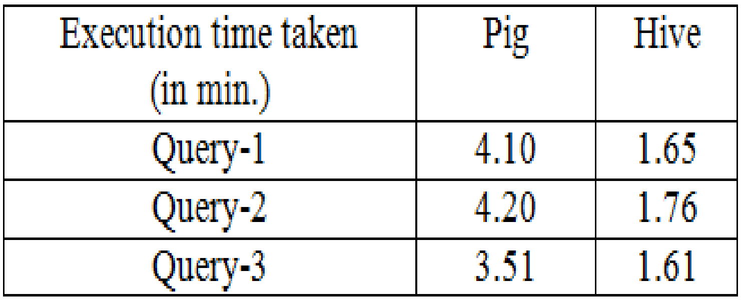
Fig 9. Total number of 'GAMBLING' cases filed in each year

VII.EXPERIMENTAL RESULT ANALYSIS

After performing operations on the dataset using pig and hive, we can find the total number of crime, crime type in each year and total crime in each district, from the analysis result we can clearly examine the crime rate and by analyzing these datasets we make a decision to reduce the crime rate by making some decision.

In our experiment we also introduced hive which is more useful as compared to pig on analysis of .csv datasets. We can say that hive performs fast as compared to pig on the basis of various parameters, also the above query results demonstrate that the execution time taken by hive is very less as compared to pig. And the map reduce job generated by hive is less as compared to pig whereby the execution time is less in hive. The experimental results are shown below.

Table 1. Execution time taken by hive and pig



Execution time taken (in min.)	Pig	Hive
Query-1	4.10	1.65
Query-2	4.20	1.76
Query-3	3.51	1.61

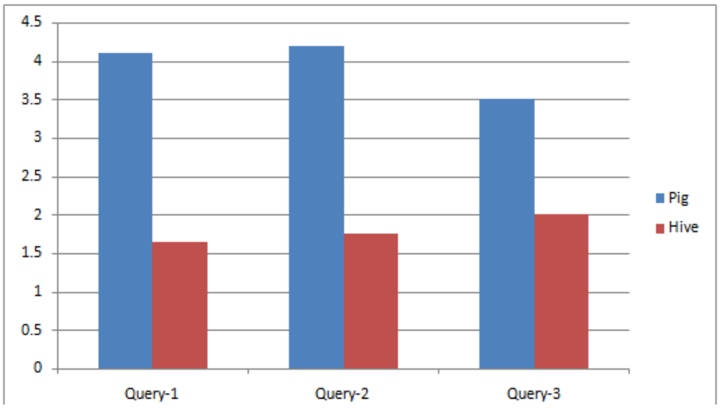


Fig 10. Execution time taken by hive and pig

VIII. CONCLUSION

On analyzing complete scenario regarding the analysis of big data we say that using the traditional analytical tool we cannot perform analysis on such huge and complex data, so we use a new powerful tool which is designed for deep analysis called Hadoop. HDFS is used for storing huge amount of crime data, Hive queries and pig script are executed on the crime dataset. Based on the parameters like execution time, number of map reduce jobs, it has been examined that hive holds better and efficient than pig.

REFERENCES

[1] Arushi Jain, Vishal Bhatnagar, “Crime Data Analysis Using Pig with Hadoop” in International Conference on Information Security & Privacy (ICISP2015), 11-12 December 2015, Nagpur, INDIA, in ELSEVIER 2015.

[2] Rahul Kumar Chawda, Dr. Ghanshyam Thakur, “Big Data and Advanced Analytics Tools”, 2016 Symposium on Colossal Data Analysis and Networking (CDAN), IEEE 2016, ISSN: 978-1-5090-0669-4/16.

[3] http://hadoop.apache.org/ .

[4] https://pig.apache.org/.

[5] https://hive.apache.org/.

[6] Mrunal Sogodekar, Shikha Pandey, Isha Tupkari, Amit Manekar, “BIG DATA ANALYTICS: HADOOP AND TOOLS”, in 978-1-5090-2730-9/16, 2016 IEEE.

[7] Shiju Sathyadevan, Devan M.S, Surya Gangadharan. S, “Crime Analysis and Prediction Using Data Mining”, in IEEE 2014.

[8] Hadoop Wiki Website, Apache, http://wiki.apache.org/hadoop.

[9] Pulkit Sharma, Komal Mahajan, Dr. Vishal Bhatnagar, “Analyzing Click stream Data using Hadoop” in IEEE 2016.

[10] Jyoti Nandimath, Ankur Patil, Ekata Banerjee, Pratima Kakade, Saumitra Vaidya, “Big Data Analysis Using Apache Hadoop” in IEEE IRI 2013, August 14-16, 2013, San Francisco, California, USA.

[11] Shankar Ganes h Manikandan, Siddart h Ravi, “Big Data Analysis using Apache Hadoop” in IEEE 2014.