International Journal of Advanced Engineering: Artificial Intelligence, Robotics and Automation
ISSN 3049-396X
Vol.:2 Issue: 1
Published online on: Dec 2025
PP. 25-34

Analysing Crime Datasets Using Hive and Pig: A Performance Perspective

Harsh Kumar Tomar [1], Sandeep K Tiwari[2], Shashank Swami[3], Shashi Pratap Tomar[4],
[1,2,3,4]Vikrant University, Gwalior, India

Abstract— Nowadays, as the population continues to grow, the incidence of crime and the crime rate also rise. To identify a crime pattern, it is essential to use an appropriate data mining technique, as superior data mining methods yield improved pattern outcomes, enabling us to manage the crime rate effectively. However, currently, the amount of data generated is extremely large, and traditional tools and techniques cannot manage the analysis of such vast and complex data. Thus, we require a robust instrument and method to manage significant volumes of data. This paper presents big data analytics through Pig and Hive, highlighting critical challenges that governments encounter in decision-making processes to lower crime rates. By analyzing extensive crime datasets with big data analytical tools, we can determine the crime rate categorized by year, district, and type of crime. Queries in Hive and scripts in Pig are run on the crime dataset. Considering factors such as execution duration and the count of map-reduce tasks, it has been analyzed that Hive is more efficient and superior to Pig.

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

I. INTRODUCTION

Each day, the rate is rising substantially. Crime cannot be predicted because it is neither orderly nor random. Moreover, electronic devices and advanced methods assist criminals in carrying out their illegal activities. According to the Crime Records Bureau, offenses such as theft and arson have decreased, while incidents like murder, sexual assault, and gang rape have increased. Although we often struggle to identify who the victims of crime might be, we can anticipate the locations where such incidents are likely to occur. Only in the last few decades has technology made specialized data processing a practical solution for a wide range of law enforcement personnel that is both affordable and accessible. Due to restrictions on the availability of criminal information or records, we tend to gather crime data from diverse sources such as websites, news platforms, blogs, social media, RSS feeds, etc. This vast amount of data is utilized as a reference for generating criminal record information. The primary challenge for researchers is to create a more efficient tool for detecting crime patterns effectively. The primary obstacles we typically encounter are:

- Increase in crime info that has got to be hold on and analyzed.
- Analysis of information is troublesome since data is incomplete and inconsistent.
- Limitation in obtaining crime information records from enforcement department.

The prevalence of cybercrimes is rising in this expanding realm of technology and is testing the capacity of investigators. These days, there is a growth in the creation of crime-related data, which is primarily digital in nature. These days, it is impossible to handle created data effectively using conventional analysis methods.

Corresponding Author's E-mail ID: soet_sandeep@vikrantuniversity.ac.in

Big Data

With each passing day, the rate is rising considerably. Crime cannot be predicted because it is not systematic or random. Moreover, the technologies of electronic equipment and sophisticated tactics aid criminals by accessing the vast amounts of data generated by social media sites and daily social media activities, such as news, news channels, new technologies, television, mobile devices, computers, and various industries, all of which constitute big data [11]. We can state that it exceeds thousands of data storage for the advancement of industries. Understanding if we possess information or history of prior data makes it significantly easier to implement upcoming changes within industries or businesses. In today’s competitive landscape, industries and businesses are rapidly expanding thanks to the utilization of historical data, commonly referred to as big data. Processing and analyzing big data is quite challenging and complex. However, with HADOOP's assistance, data can be processed and analyzed easily. Big data [12] is generated from various sources such as television, mobile devices, and other origins like industrial data records.
It is three characteristics of big data [13]:
l. Volume
2. Velocity
3. Verity.

A. Volume

Volume is primarily defined as data amount. Megabytes, gigabytes, zeta bytes, and peta bytes of data are increasing at an exponential rate. The volume of data is enormous, and processing and storing it is challenging. According to recent reports, Facebook processed one million photos every second in 2010. According to recent research, Twitter produces 12 terabytes of data per day. In 2012, Facebook reported that users registered 2.7 billion "likes," "comments," and massages every day.

.

B. Velocity

Social media is one the primary reason to offer the Data exponentially. Social media platforms consistently produce complicated, unstructured, and semi-structured data. Currently, 90% of the data was produced in the last two years. Increase the velocity of large data with use of mobile, televisions more sophisticated technology. Internet is main factor to collecting of huge data. based on the user's need, which is saved elsewhere. It's referred to as velocity.

C. Verity

There are two types of data: structured and unstructured. Unlike tabular data, ERP, and backup storage for massive amounts of data, structured data is always in a set format and cannot be changed. However, unstructured data—such as text, audio, video, photos, and data from several social media platforms like Facebook, Twitter, LinkedIn, logs, files, web conversations, etc.—never follows a set format. Up to 85% of the data in all businesses and sectors is in semi-structured, unstructured, and structured formats.

Hadoop

Components of Hadoop [10] Hdfs is used to store big datasets, and Map Reduce is used to process the data. Hadoop consists of two parts: Map Reduce for processing and HDFS for storing.

Apache Pig

Developed by Yahoo, Apache Pig [4] is a big data analytical tool that allows us to analyze massive datasets using the Hadoop architecture. It supports the Pig Latin language.Both organized and unstructured data are supported by Apache Pig, which can operate both with and without Hadoop. However, it cannot manage huge data in the absence of Hadoop (standalone mode).

.

Hive

Facebook created Apache Hive [5], a data warehouse analytical tool for processing and analyzing massive information on top of Hadoop.

II. LITERATURE REVIEW

According to [1], huge data is a vast and intricate collection of data that comes from a variety of sources, including sensors, social media posts, sales transactions, and more. It becomes difficult to manage this large amount of data with conventional processing programs. The industry offers a wide range of big data analytics tools and techniques. Because of the ongoing population growth, governments may find it difficult to make strategic decisions that maintain law and order by evaluating crime rates and related data. Protecting the country's voters from illicit activity is often crucial. . The vast amount of data generated every day from various sources through Big Data Analytics (BDA) is the ideal resource for pinpointing areas that require improvement. BDA helps analyze particular trends that must be found for efficient law enforcement and preserving a sense of security.

According to [2], mistreating cloud delivery models with enormous knowledge analytical skills may facilitate adoption for a number of businesses, and most importantly, it may alter useful insights that may give them entirely other kinds of competitive advantage. Many companies, including Amazon Huge Knowledge Analytics Platform, HIVE Internet-based Interface, SAP Huge Knowledge Analytics, IBM Info Sphere Big Insights, TERADATA Huge Knowledge Analytics, 1010data Huge Knowledge Platform, Cloud Era Huge Knowledge Resolution, and others, are offering online huge knowledge analytical tools. These companies use a variety of techniques to evaluate vast amounts of information and also provide an easy-to-use interface for data analysis.

This analytical article [7] shows how to apply Apache Pig on a global drawback. In order to provide innovative ways to present the news updates, this study report also offers analysis of data gathered from news websites. Data on representing news updates in various classes is provided in this work. The authors have presented study of the use of RSS as an input tool and the Apache Pig tool for relevant outcomes. This study uses three tools—HIVE, PIG, and Map—to analyze large amounts of data using Hadoop. It uses internet logs that record client behaviors, such as social media network searches, to determine the information's outcome. According to the authors, Map Cut Back takes longer to operate than HIVE and Pig Area. Data processing is given top priority in [6] information technology. A significant amount of both structured and unstructured data is generated from a variety of sources, including emails, online logs, social media platforms like Facebook and Twitter, and more. Hadoop is an open-source Java framework that facilitates distributed and parallel data processing and is used for dependable data storage. Based on past purchases, e-commerce corporations analyze website traffic or navigation patterns to identify likely opinions, interests, and dislikes of an individual or a group. We compare a few commonly used data analytics technologies in this study.[9]

III PROBLEM DEFINITION

Police complaints, newspaper reports, and articles were the primary sources of crime data in the past. These sources were either printed or handwritten. However, as technology has advanced, crime data is now available in both hard copy and soft copy formats. Previous situations varied because there was a lower crime rate and less data was produced on criminal activity. Mapping crime and forecasting potential crime hotspots depend heavily on historical data on criminal activity. Even though there was relatively little data, using typical data mining techniques to analyze it was an extremely laborious and time-consuming operation. . Data generation nowadays is vast due to increased crime rate which cannot be handled by traditional data analysis techniques.

IV PROPOSED WORK

We present Apache Hadoop, an open-source framework for processing and storing massive datasets, because we need a strong tool to store and analyze this kind of data [10].

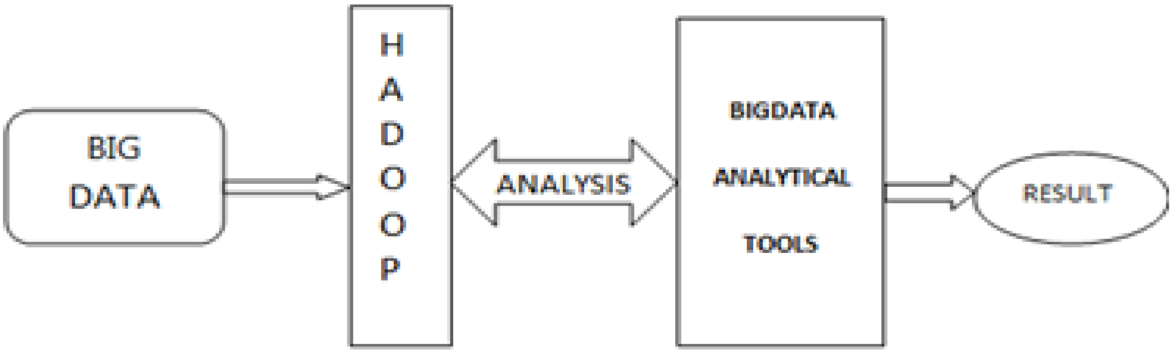


Figure 1 Workflow Diagram

In order to analyze consumer complaint databases, we require:

1. Data

We collected crime data records from 2001 to 2017 in order to gather the crime dataset, which together contains a huge number of crime records connected to which crimes occur in those years.

2. Hadoop

Components of Hadoop Hdfs is used to store big datasets, and Map Reduce is used to process the data. Hadoop consists of two parts: Map Reduce for processing and HDFS for storing.

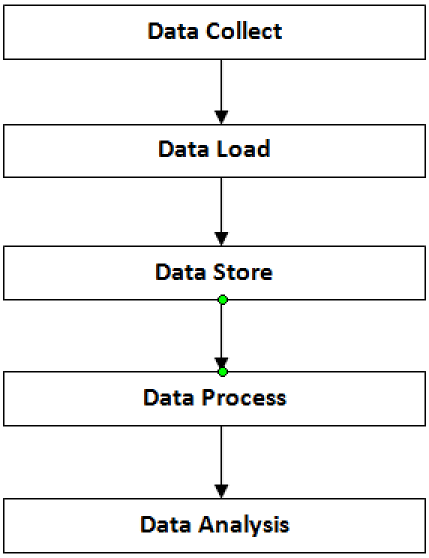
3. Analytical Tools for Big Data

We require effective analytical tools [6] that operate on top of Hadoop, Apache Hive, and Apache Pig so that we can examine the criminal datasets in order to understand these vast amounts of data.

V PROPOSED METHODOLOGY

Algorithm Steps are:

Step 1: We start by gathering crime datasets from online sources.
Step 2: Using the Hadoop command line, we may load the customer complaint datasets following collection.
Step 3: The datasets are saved in HDFS, a very dependable system for storing large or complicated data sets.
Step 4: Map Reduce, a processing engine within the Hadoop framework, processes the crime datasets.
Step 5: Using big data analytical tools like Apache Hive and Apache Pig, which can operate on top of Hadoop and process the information in the backend, we can analyze these crimes.

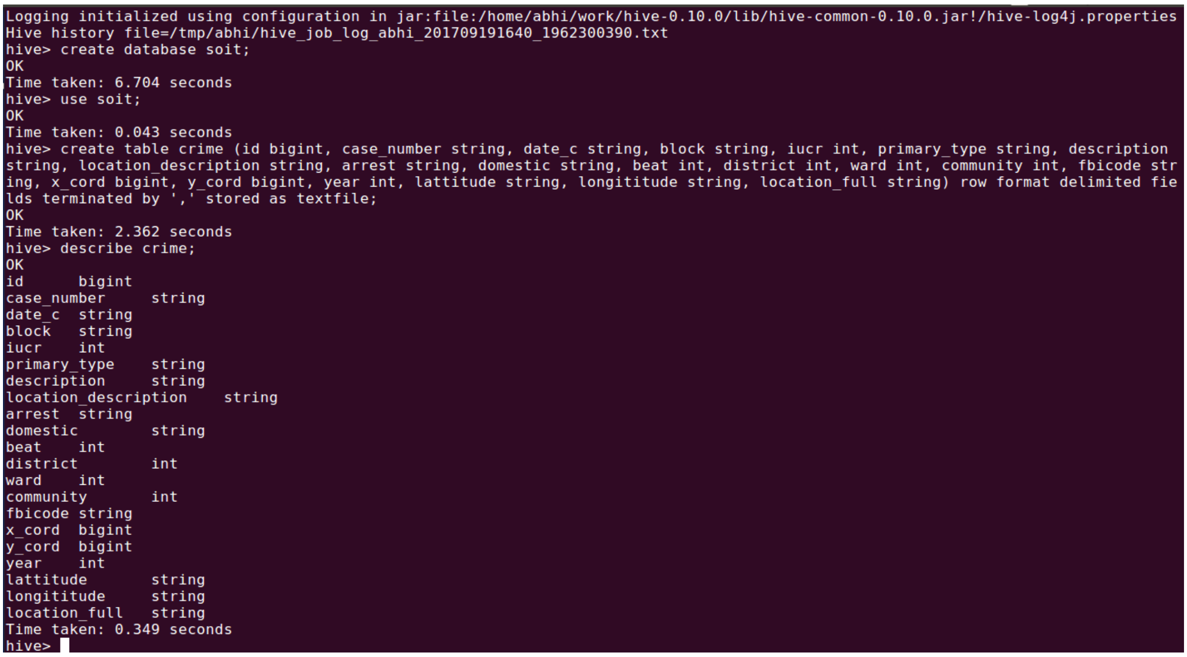


Analysis Steps

VI 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



```
Logging initialized using configuration in jar:file:/home/abhi/work/hive-0.10.0/lib/hive-common-0.10.0.jar!/hive-log4j.properties
Hive history file=/tmp/abhi/hive_job_log_abhi_201709191640_1962300390.txt
hive> create database soit;
OK
Time taken: 6.704 seconds
hive> use soit;
OK
Time taken: 0.043 seconds
hive> create table crime (id bigint, case_number string, date_c string, block string, iucr int, primary_type string, description
string, location_description string, arrest string, domestic string, beat int, district int, ward int, community int, fbicode str
ing, x_cord bigint, y_cord bigint, year int, lattitude string, longititude string, location_full string) row format delimited fie
lds terminated by ',' stored as textfile;
OK
Time taken: 2.362 seconds
hive> describe crime;
OK
id      bigint
case_number     string
date_c  string
block   string
iucr    int
primary_type    string
description     string
location_description    string
arrest  string
domestic        string
beat    int
district        int
ward    int
community       int
fbicode string
x_cord  bigint
y_cord  bigint
year    int
lattitude       string
longititude     string
location_full   string
Time taken: 0.349 seconds
hive> 
```

Fig 2. Dataset Description

Tools and Technologies Used:

1. Hadoop
2. Hive
3. Pig

VII. EXPERIMENTAL FINDINGS

To analyze criminal datasets, we can install Hadoop-1.1.2 on Ubuntu and add Hive and Pig on top of Hadoop. We can use Hive and Pig to evaluate datasets once they have been loaded into hdfs. Although they use different programming paradigms to analyze crime datasets, both analytical methods produce accurate results. The Hive Web Interface supports both SQL and the Hive query language (HQL). Pig is compatible with Pig Latin, a scripting language.

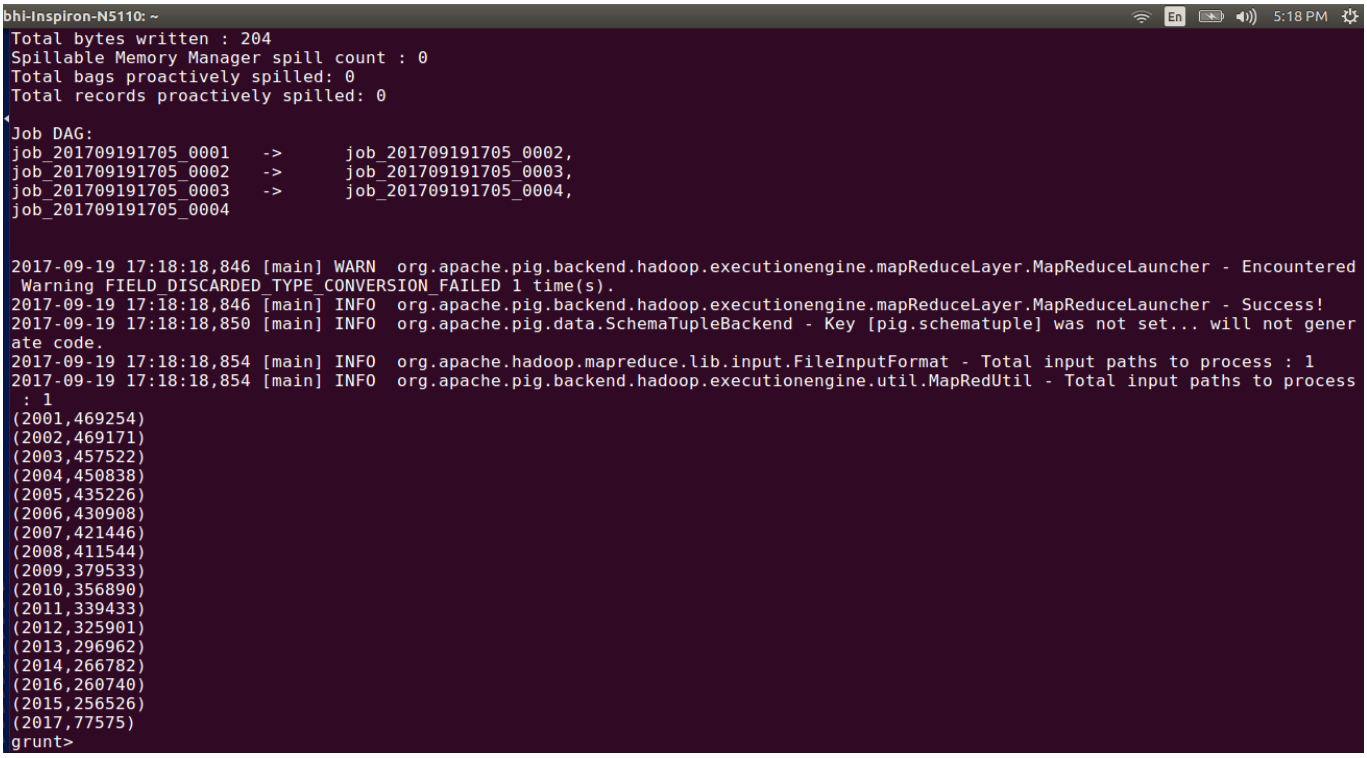
Question 1 total amount of crimes committed annually?

Using pig:-

We can now use Pig to analyze the criminal datasets after setting up Hadoop, Hive, and Pig. To do this, we must first launch Pig using the command pig. It allows us to enter the grunt shell.Once we are in the Grunt shell, we can write the pig script that will allow us to retrieve the query-1 result. The query-1 problem is solved by the following pig script:

```
A = Load the data set using Pig Storage;
B = for each A generate year as year;
C = filter B by year is not null;
D = group C by year;
E = for each D generate group, COUNT(C.year);
F = order E by $1 DESC;
Result = LIMIT F 17;
dump Result;
```

After executing these pig scripts into grunt shell the output of the scripts are shown in fig.3 in which the result show the total number of crime happend in each year.



```
Total bytes written : 204
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_201709191705_0001   ->      job_201709191705_0002,
job_201709191705_0002   ->      job_201709191705_0003,
job_201709191705_0003   ->      job_201709191705_0004,
job_201709191705_0004


2017-09-19 17:18:18,846 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered
 Warning FIELD_DISCARDED_TYPE_CONVERSION_FAILED 1 time(s).
2017-09-19 17:18:18,846 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-09-19 17:18:18,850 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not gener
ate code.
2017-09-19 17:18:18,854 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-09-19 17:18:18,854 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process
 : 1
(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)
grunt>
```

Figure 3. Query 1 result using pig

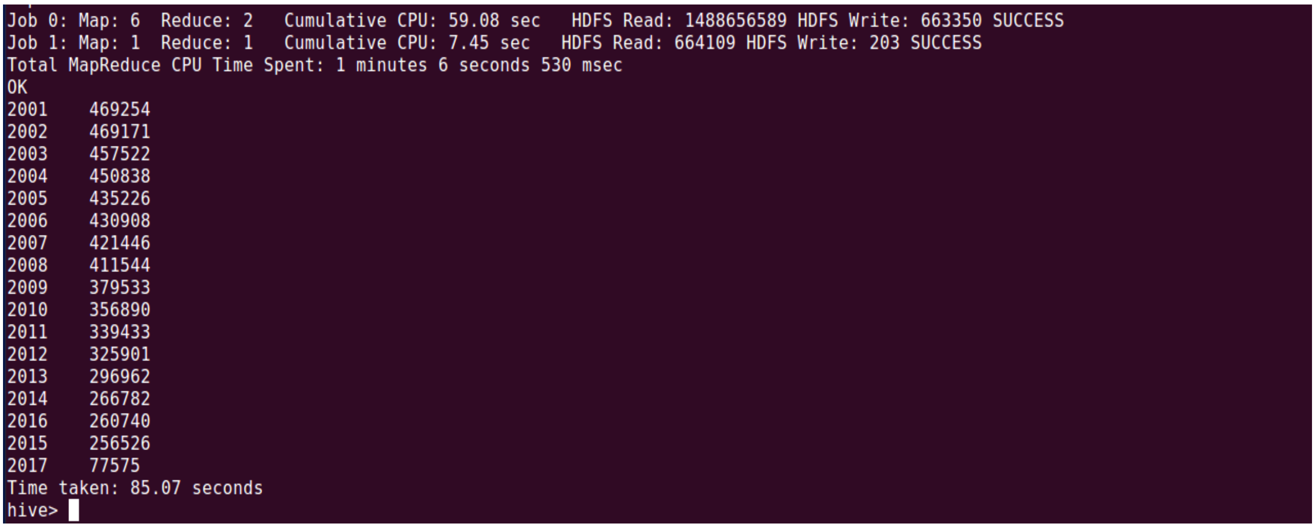
Using Hive:-

Now we can perform same query written and executed by hive, by which we can start the hive by typing hive command to enter into hive shell.

After entering into hive shell, we can write the hive query to execute the query-1 problem. Hive support all the sql queries and we can write the query is:-

Select year, count (*) as count from crime group by year order by count desc limit 17;

The query which is written above and for these query a map reduce job is launched by hive to solve a problem. After finishing the execution of map reduce job for hive query, the results are shown in fig. 4, in which the maximum numbers of crimes per year are shown and also the time taken by the hive query is also shown in figure 4.

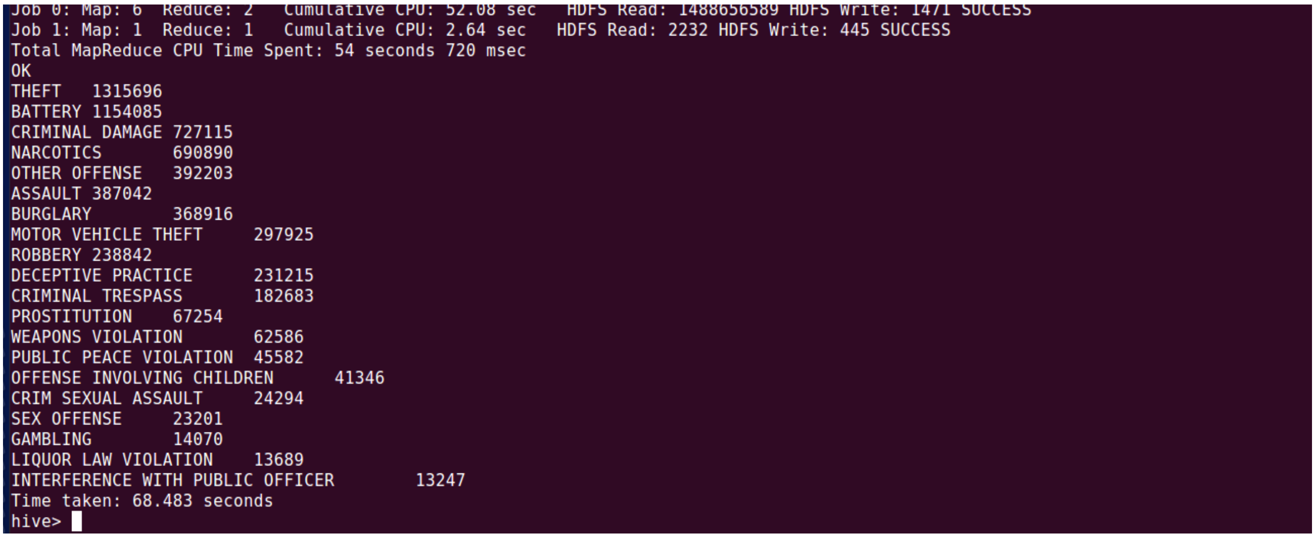


```
Job 0: Map: 6  Reduce: 2   Cumulative CPU: 59.08 sec   HDFS Read: 1488656589 HDFS Write: 663350 SUCCESS
Job 1: Map: 1  Reduce: 1   Cumulative CPU: 7.45 sec   HDFS Read: 664109 HDFS Write: 203 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 6 seconds 530 msec
OK
2001    469254
2002    469171
2003    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
Time taken: 85.07 seconds
hive>
```

Figure 4. Query-1 execution result & time taken by hive

Similarly, by using hive and pig we can find the result of query 2 and query 3 and the outcomes are shown below. Figure 5 & figure 6 shows the result of query 2 and result of query 3.

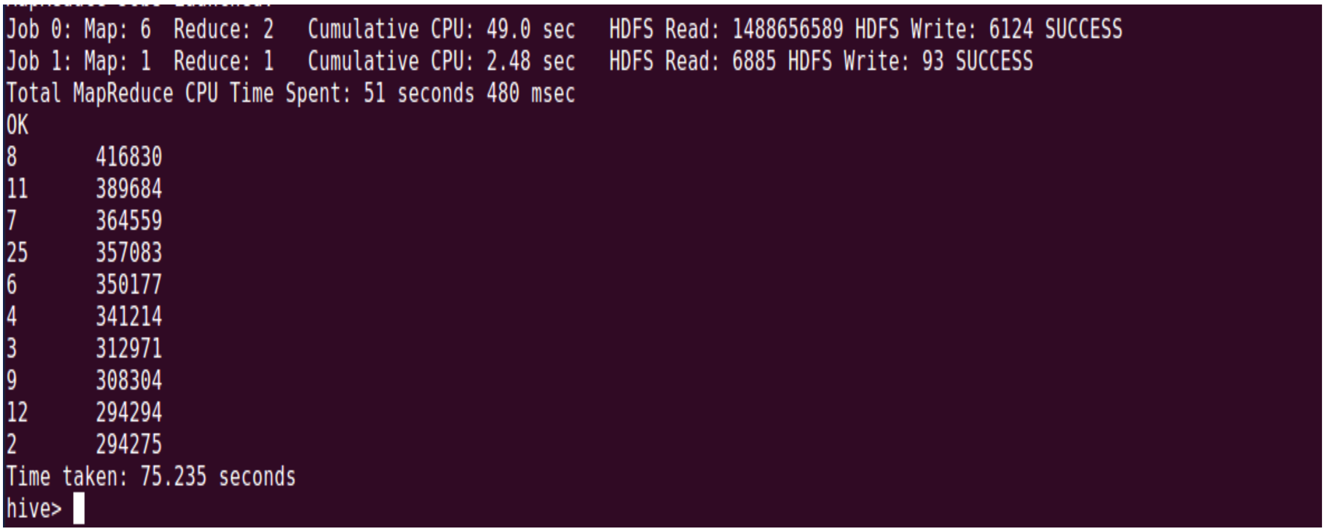
Query-2 Total number of crime in each type?



```
Job 0: Map: 6  Reduce: 2   Cumulative CPU: 52.08 sec   HDFS Read: 1488656589 HDFS Write: 1471 SUCCESS
Job 1: Map: 1  Reduce: 1   Cumulative CPU: 2.64 sec   HDFS Read: 2232 HDFS Write: 445 SUCCESS
Total MapReduce CPU Time Spent: 54 seconds 720 msec
OK
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
LIQUOR LAW VIOLATION    13689
INTERFERENCE WITH PUBLIC OFFICER        13247
Time taken: 68.483 seconds
hive>
```

Figure 5. Query 2 result

Query 3:- Total number of crime in each district?

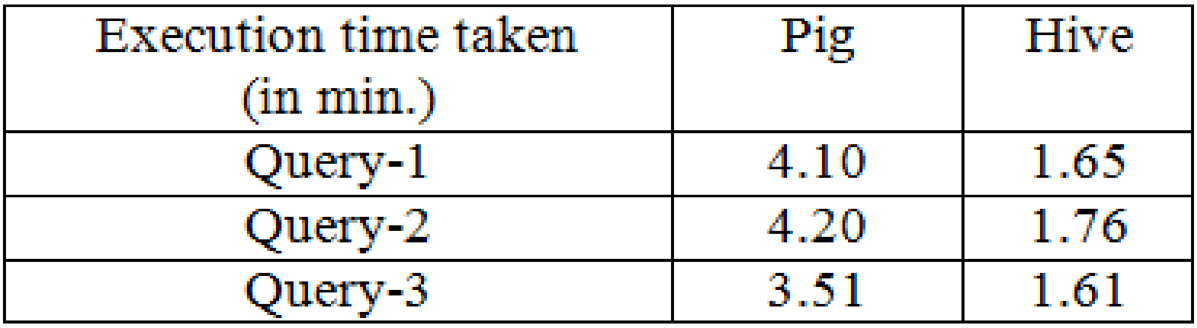


```
Job 0: Map: 6  Reduce: 2   Cumulative CPU: 49.0 sec   HDFS Read: 1488656589 HDFS Write: 6124 SUCCESS
Job 1: Map: 1  Reduce: 1   Cumulative CPU: 2.48 sec   HDFS Read: 6885 HDFS Write: 93 SUCCESS
Total MapReduce CPU Time Spent: 51 seconds 480 msec
OK
8       416830
11      389684
7       364559
25      357083
6       350177
4       341214
3       312971
9       308304
12      294294
2       294275
Time taken: 75.235 seconds
hive> 
```

Figure 6. Result of query 3

VIII. EXPERIMENTAL RESULT ANALYSIS

We can determine the total number of crimes, the type of crimes committed each year, and the total number of crimes committed in each district after using pig and hive to perform operations on the dataset. Based on the analysis results, we can clearly examine the crime rate, and by analyzing these datasets, we make a decision to lower the crime rate. We also included hive in our testing, which is more helpful for analyzing.csv datasets than pig. Based on a number of factors, we may conclude that hive outperforms pig. Additionally, the query results above show that hive takes significantly less time to execute than pig. Additionally, the map shows that the hive generates less jobs than the pig, which results in a shorter execution time. The outcomes of the experiment are displayed below.



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

Table 1. Execution time taken by hive and pig

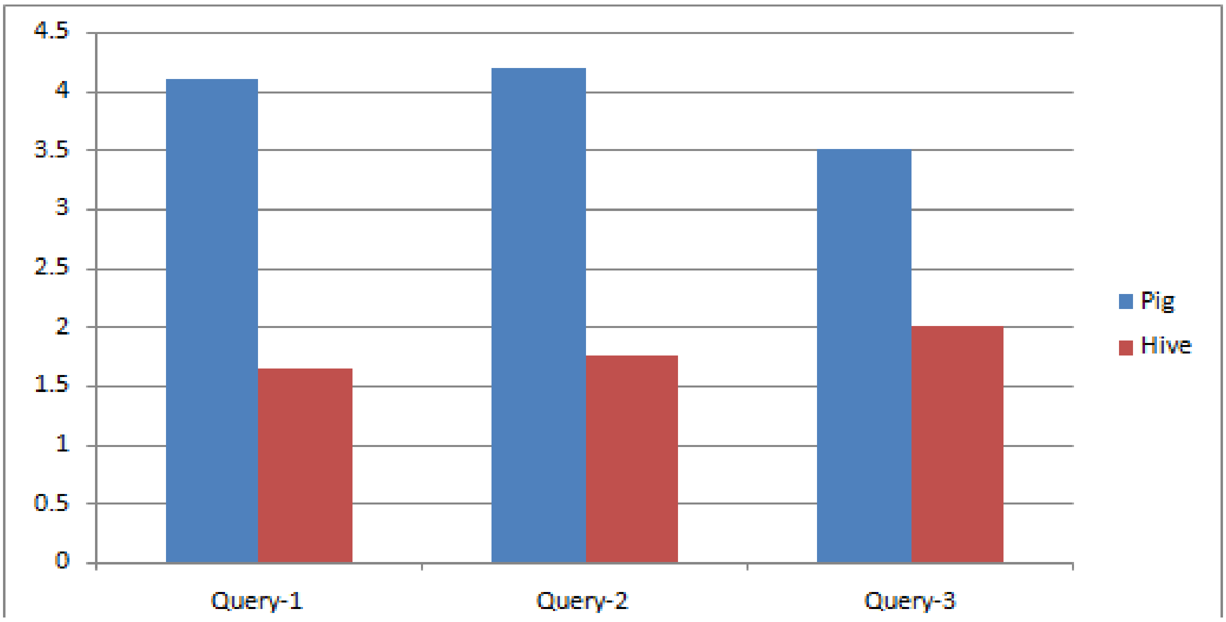


Fig 7. Execution time taken by hive and pig

No. of jobs launched	Pig	Hive
Query-1	4	2
Query-2	4	2
Query-3	4	2

Table 2. No. of job launched by pig and hive

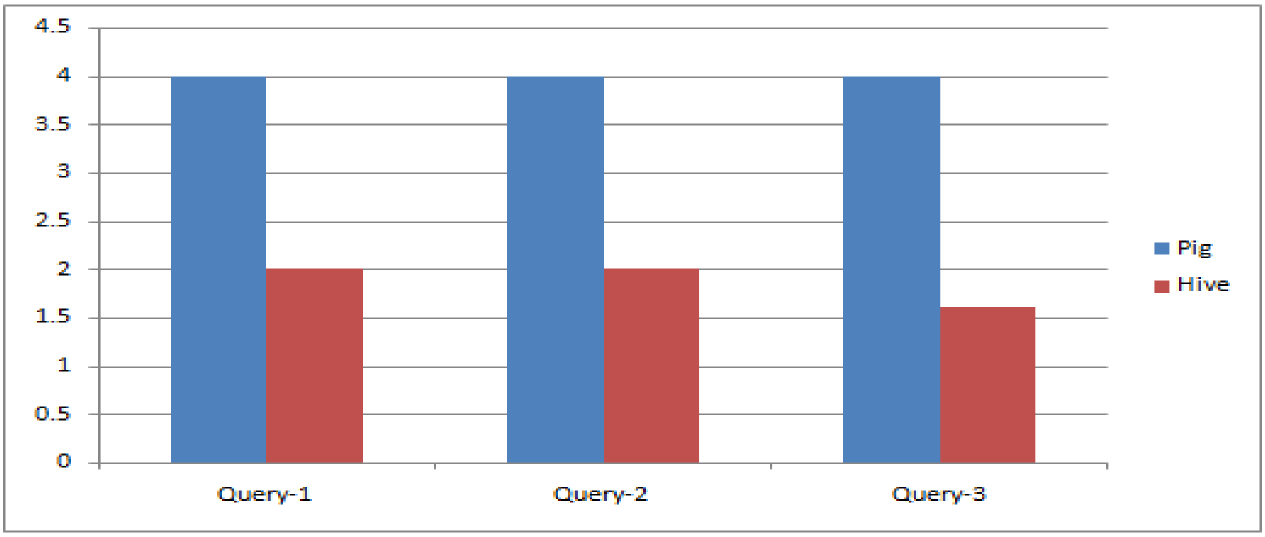
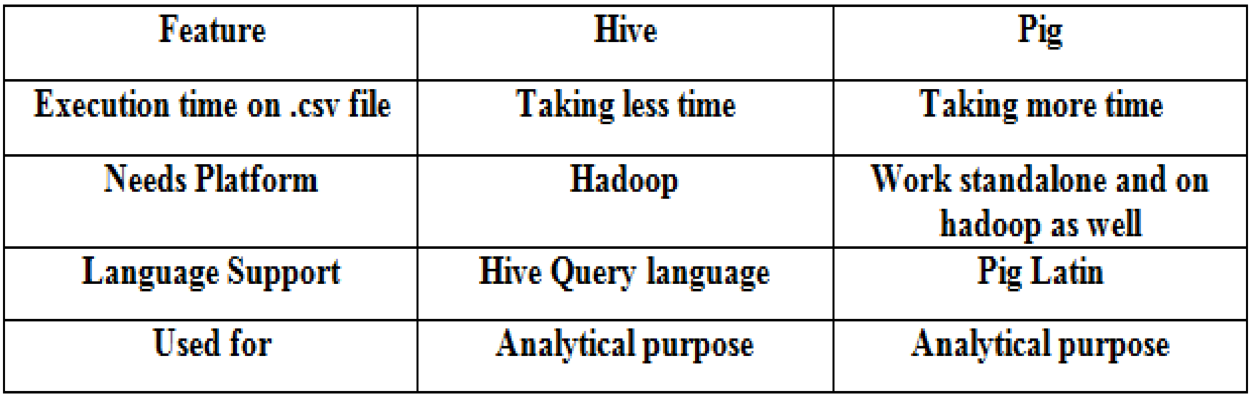


Fig 8. No. of job launched by pig and hive



Feature	Hive	Pig
Execution time on .csv file	Taking less time	Taking more time
Needs Platform	Hadoop	Work standalone and on hadoop as well
Language Support	Hive Query language	Pig Latin
Used for	Analytical purpose	Analytical purpose

Table 3. Difference between hive & Pig

VI CONCLUSION

Nowadays, a common option for carrying out extensive data analytics is Hadoop Map Reduce. By examining the massive and extensive crime datasets, big data analytical techniques allow us to determine the crime rate by year, district, and kind of crime, which illuminates important challenges that the government faces when making decisions to lower the crime rate. Pig scripts and hive queries are run on the crime dataset. It has been determined that hives are more effective and efficient than pigs based on factors like execution time and the quantity of map reduction jobs.

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

[4] [07] Jurmo Mehine, Satish Srirama, Pelle Jakovits “Large Scale Data Analysis Using Apache Pig”

[5] [08] Dave Jaffe “Three Approaches to Data Analysis with Hadoop”

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

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

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

[9] 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.

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