ISSN: 0976-3104
SPECIAL ISSUE: Emerging trends in Computer Engineering and Research (ECER)

IIOAB JOURNAL ISSN: 0976-3104

ARTICLE

ENHANCED HADOOP PERFORMANCE ANALYSIS USING HADOOP ECO-SYSTEM

Shivam Suryawanshi[1], Jabanjalin Hilda[2*]

[1]School of Computer Science and Engineering, Vellore Institute of Technology - Vellore, T.N, INDIA
[2]Faculty of School of Computer Science and Engineering, Vellore Institute of Technology - Vellore, T.N, INDIA

ABSTRACT

All over the World, the idea of Big Data analytics is constantly growing in the software domain. Hadoop as an open source has become a popular Java framework for large scale data processing in recent years. Day by day, the rate of data is also increasing through different sources, and to analyze such huge data it is required to be processed on Hadoop Distributed File System (HDFS) rather than processing using traditional methods. The Apache Hadoop eco-system supports many open source tools for analyzing substantial datasets. Three well-known tools for data analyzing in the HDFS are Hive, Pig, and MapReduce. In this project, we are analyzing crime dataset which is imported on HDFS using Sqoop tool and further, it is analyzed using Hive, Pig, and MapReduce program. The result shows that it is an enhanced Hadoop performance analysis though Hadoop eco-system components where Hive works more efficiently than Pig and MapReduce Program. Later, results are compared with MapReduce program. In this crime data analysis, Hive outperforms 2.23 times than MapReduce and 1.90 times than Pig.

THE IIOAB JOURNAL

KEY WORDS
HDFS, Hadoop, MapReduce, Hive, Pig, Sqoop

Received: 4 June 2017
Accepted: 30 Sept 2017
Published: 12 Oct 2017

***Corresponding Author**
Email: jabanjalin.hilda@vit.ac.in
Tel.: +91-7598193077

INTRODUCTION

These days a lot of information are coming from various sources like Social network profiles, advanced media - audio, photos, and web sources, so on. The successful storage, querying and analyzing of these information has turned into a challenging test to the business. When it comes to crime, size of crime data growing more day by day and storing, analyzing, processing such large data has dependably been a big concern in the field of database management domain. Nowadays big data has become a part of handling such issues related to large information. There are a few inquiries with regards to crime. Questions like - Is crime a serious issue where you live? What sorts of crimes happen frequently? Is the crime rate increasing or decreasing where you live? Do you think the world will be secure or riskier later on? [1] And many more. To solve this problem, it is required to analyze the crime related data firstly using big data technology. It is also important to characterize certain definitions that are identified with Big Data and Hadoop.

Big Data

Big Data are huge-volume, high-speed, as well as huge-variety data resources that require new types of handling to ensure upgraded process enhancement [2]. Increased processing speed, storage limit, and systems administration have made information to develop in every one of the 4 measurements. The four characterizing qualities of Big Data- volume, variety, value and velocity - are incorporated for the better performance of data handling. There are distinctive methods for characterizing and comparing Big Data with the customary information, for example, information size, content, collection and processing. Huge information has been characterized as expansive data sets that can't be prepared using conventional handling strategies, for example, Relational Database Management Systems, in an average preparing time. So for that new technology comes under big data is nothing but Hadoop [3], [4].

Hadoop

Hadoop is the system which allows to store and process a huge amount of data across multiple computers connected in distributed environment. By using the Hadoop technology, we can scale up from single-node cluster to multi-node cluster of machines, each offering different storage capacity and computation. Hadoop can be used in a different application in order to process the data as the data is generating more and more information on a daily basis, and it is becoming very difficult to handle the data. The importance of big data technologies is providing more accurate analysis, which can be used for decision-making in any business process. There are different big data technologies such as operational Big Data which includes a system like MongoDB where data is primarily stored [5], [20]. These systems are supposed to take an advantage of new distributed computing systems that have created over the earlier decade which will run productively. This tends to information workloads significantly less demanding to manage, simple, quicker, and less costly to execute. Another Big data technology is, which uses the parallel environment such as MapReduce programming that provides analytical capabilities to review and complex examination analysis all of the type of the data. We can use MapReduce to scale up the single machine to multiple machines. It provides different method for mapping the information which is integral to the capacities gave by SQL.
Using mapping function, it takes the data from the huge dataset and distributes to multiple machines. [Fig. 1] Shows the Hadoop Framework which incorporates MapReduce layer and HDFS layer. A little Hadoop group incorporates a master part and different slave part. The framework consists of a DataNode, NameNode, Job Tracker and Task Tracker [6].

COMPUTER SCIENCE: Guest editor-Prof. S. Prabu & Prof. Swarnalatha P

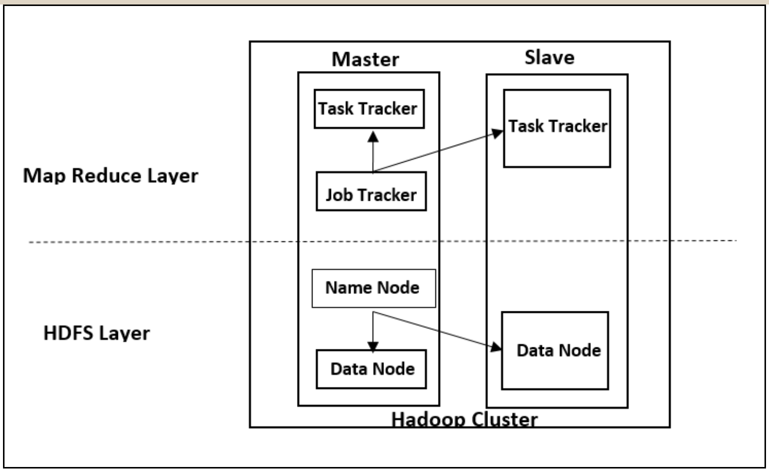

Fig. 1: Hadoop Framework [5]

- *Data Node:* Constantly inquire as to whether there is something for them to do, mainly used to track which data nodes are up or and which data nodes are down.
- *Name node:* Manages the record framework name space, it monitors where each block is present.
- *Job tracker:* Assigns the mapper job to undertaking tracker nodes that have the information or are near the information (same block)
- *Task tracker:* Keep the work as near the information as could be expected under the circumstances.

NameNode stores MetaData about the information being put away in DataNodes though the DataNode stores the real Data. JobTracker is an expert which makes and runs the jobs. JobTracker which can keep running on the NameNode distributes the job to TaskTrackers which keeps running on DataNodes; TaskTrackers run the assignments and report the status of the job to JobTracker. A slave part is responsible tracking job with the help of DataNode and TaskTracker. In a single-node cluster, both the NameNode and DataNode use the same machine for processing the data. In a multimode cluster, NameNode and DataNodes are ordinarily on various machines. There is one and only NameNode in a bunch and numerous DataNodes [6].

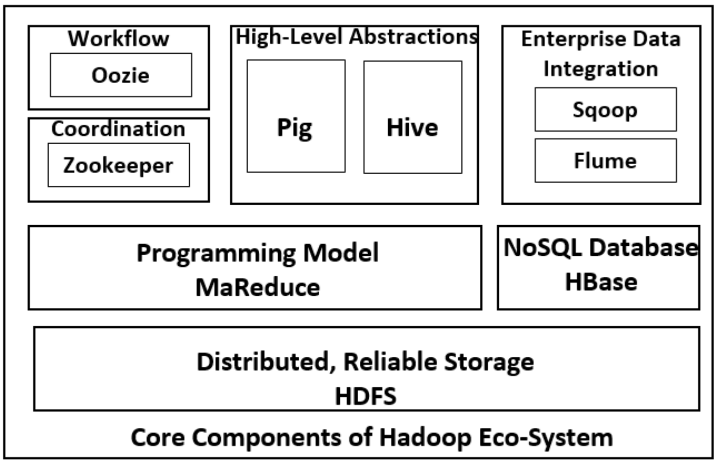

Fig. 2: Hadoop Eco-system components [8]

In addition, by enhancing the Hadoop we can reduce processing time, data size to read and other parameters in Hadoop MapReduce environment [7], [8]. This paper focuses on conceptual technologies, tools about huge information analysis and results shown through a couple of situations how it's useful for associations inside different segments if examination are conducted effectively. There are core components of the Hadoop eco-system, they are Sqoop, Pig, Hive, and Map Reduce. [Fig. 2] shows the architecture of Hadoop Eco-system components [9].

LITERATURE SURVEY

At current situation, each organization is confronting common difficulties which should be adapted up rapidly and effectively. Hadoop was incorporated with various segments that can be utilized for placing, executing and analyzing information significantly as well as proficiently. The decision of a specific utility relies on upon the necessities of the data analysis, the user technical knowledge, and the tradeoff between development time and execution time. Big data environment exhibits extraordinary turn for organizations inside different parts to compete an upper hand. There are irregularities and difficulties inside Big Data analysis: adequate calculations to cover raw information or investigation, dependability and integrities of Big Data, information storage issues and MapReduce paradigm [5].

Fuad et al. [9], presented a conference paper which introduces the execution time of Hive, Pig, and MySQL Cluster on a basic information system with basic queries while the information is developing. In this, they have framed MySql database issue related to storing and processing information. The issue with MySQL Cluster is that as the information becomes bigger, amount of time to prepare the information increments and for that extra sources might be required. With Hadoop - Hive and Pig, handling time can be speedier than MySQL Cluster. Hence, three data analyzers with similar information model will run basic queries and to discover at what number of columns Hive or Pig is speedier than MySQL Cluster. They have worked on Group-Lens data set where an outcome shows that Hive is the most proper for this information model in a minimal effort hardware condition.

Prabhu et al. [10], presented a paper and they have taken web log information for the experiment and probed on Native Hadoop attributes that is a benchmark framework where they examined from the outcome i.e. when they streamline Hadoop framework attributes then they can enhance the framework execution. In this way they worked on enhancing the parameter in view of the framework assets and application and they also talked about why Hadoop setup must be transformed from its default to particular framework. After executing the experiment, they noticed that native execution has enhanced by 32.97%. For that they have considered couple of parameters.

Sathyadevan et al. [11], presented a paper where they explained about, crime data analysis and its prevention methods which can be a precise way of recognizing the common crime patterns. In their approach, they are predicting the areas where crimes are happening for more number of time and they can imagine crime inclined zones. In this paper they have used the idea of data mining where they are extracting the unknown existing features, valuable data from an unstructured node. Here they gone through an approach between software engineering and criminal equity to build up an information mining methodology that can help tackle crimes speedier. Rather than concentrating on reasons for crime event like criminal foundation of wrongdoer, political hatred etc., they are concentrating chiefly on crime variables of every day.

Alshammari et al. [12], presented a paper where they displayed Enhanced Hadoop (H2Hadoop) that permits a Name-Node to distinguish the blocks inside the cluster where particular data is present. In H2hadoop, input data is less, also input functions are lessened by the quantity of Data-Nodes conveying the main data-blocks that are again distinguished by sending a job to Task-Tracker. They have added some control feature on Name-Node whose purpose is to assign a task of particular data to a Data-Node without sending it to a whole cluster. They talked about the proposed work of H2Hadoop and showed the execution time of H2Hadoop which is efficient with respect to native Hadoop.

HADOOP ECOSYSTEM AND ITS METHODOLOGY

Problem Evaluation

With persistently expanding population, crimes and its rate dissecting related information is a tremendous issue for governments to settle on vital choices to keep up the peace. This is truly important to guard the residents of the nation from violations. The best place to admire opportunity to get better is the voluminous raw information that is created all the time from different sources by applying Big Data Analytics which breaks down to specific patterns that must be found, so that law can be kept up legitimately and there is a faith of security and prosperity among the people of the nation. In this paper, the three methodologies are differentiated with the help of a use case for Hadoop: Crime data investigation [13], [14].

The dataset examined in these tests were produced by a MapReduce program, using these crime data set as info, it is possible to check the output of the executed programs for precision. The issue is characterized further in the following segment, trailed by areas on the Hive, Pig and MapReduce solutions, and then the outcomes. Three well known tools for data analyzing occupant in the HDFS are Hive, Pig, and MapReduce. Hive gives a SQL like front end with a database foundation. Pig gives high level programming language to perform information processing that additionally empowers the users to misuse the parallelism innate in a Hadoop Cluster. MapReduce needs a PC program (frequently Java Programming) for inserting, handling and showing the output information. Hive and Pig produce MapReduce code to do the genuine performance analysis [15].

Sqoop

Sqoop is a command line tool used to transfer data from RDBMS to HDFS and vice versa [16]. Firstly, user can import data via sqoop from RDBMS (either MySql, SQL Server, PostgreSQL, etc.). After importing data

ISSN: 0976-3104

from RDBMS it will sink to HDFS using Hadoop MapReduce functionality. Following workflow shows Sqoop Architecture, [Fig. 3].

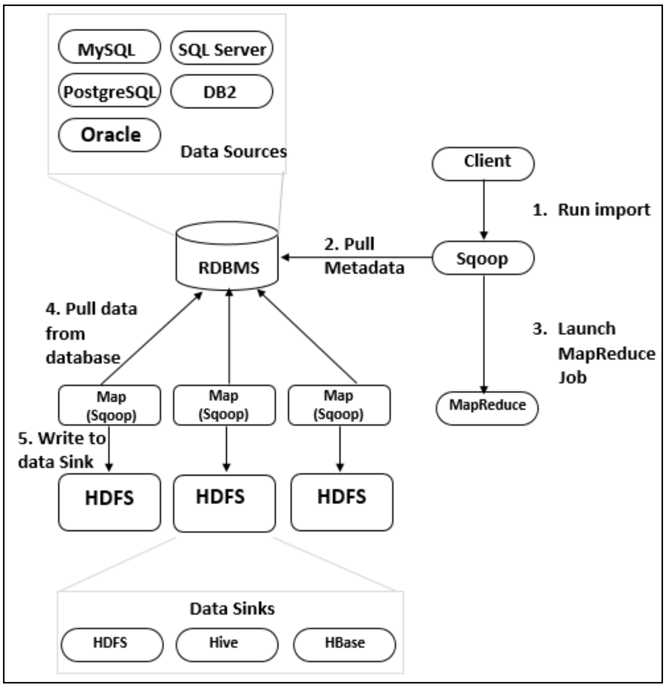

Fig. 3: Sqoop workflow architecture. [16]

HIVE

The Apache Hive device is not a RDBMS tool, it is a part of the Hadoop eco-system, which works on the data which is stored in HDFS using HiveQL (HQL), is a Structured Query Language (SQL) interface, to solve or execute the query based on the available data [17]. This SQL based language in Hadoop domain gives good platform to view the information present in tables. Hive makes a query plan that implements the HQL in a progression of MapReduce projects, produces the code for these projects, after that executes the code, gives appropriate results. Following structure [Fig. 4] is the HIVE Workflow architecture. It shows that, how hive and Hadoop works together when it comes to MapReduce task.

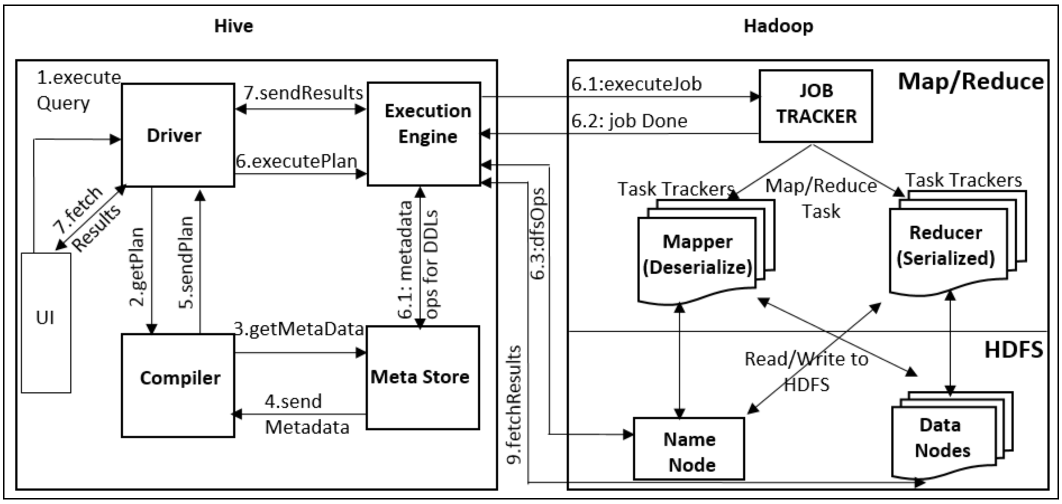

Fig. 4: Hive workflow diagram. [17]

PIG

Apache Pig tool is another information analysis device in the Hadoop Eco-system [18]. Pig is a data flow language, Pig Latin, which allows the client to determine joins, and different calculations without the need to compose an entire MapReduce program. Like Hive, Pig creates a flow of MapReduce projects to solve the data analysis steps. Following flow chart [Fig. 5] will describe PIG architecture.

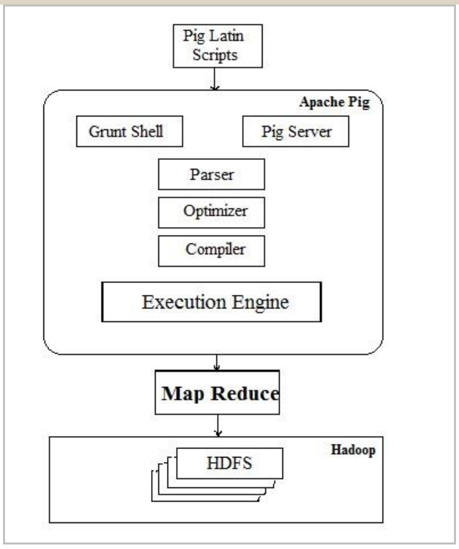

Fig. 5: Pig architecture. [18]

Map reduce

MapReduce is a system utilized by Google for handling large measures of information in a distributed domain and Hadoop is Apache's open source execution of the MapReduce structure. Hadoop is helpful for putting vast volume of information into Hadoop Distributed File System and that information get prepared by MapReduce paradigm in parallel. MapReduce is a versatile and effective programming model to perform substantial scale information. At the point when handling this monstrous information asset has been constrained to single PCs, computational force and capacity rapidly get to be bottlenecks. This massive amount of information can be handled in distributed environment by processing each task one by one. The Hadoop MapReduce structure gives a stage to such parallelization of tasks [19].

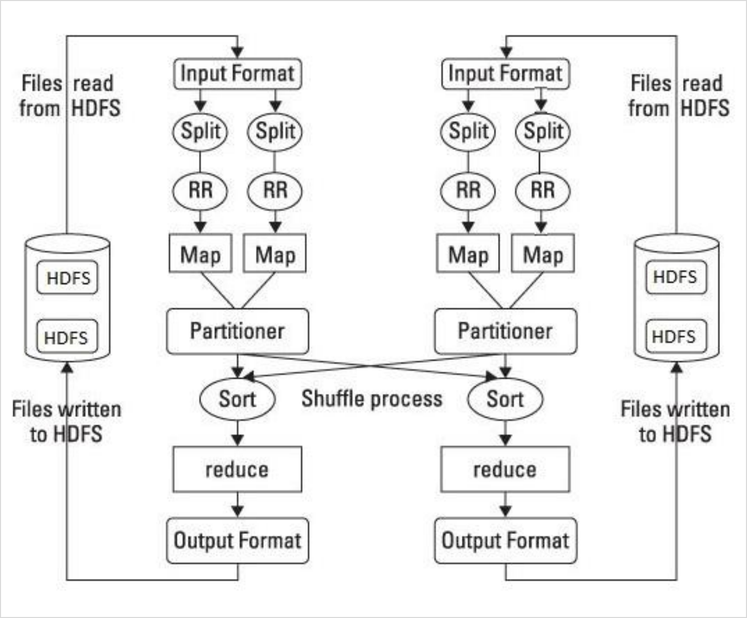

Fig. 6: Mapreduce paradigm. [19]

MapReduce Code Snippet

```
map(LongWritable k,Text v,Context c)
String s = v.toString();
            String s1[]=s.split(",");
            String type_crime=s1[6];
            c.write(new Text(type_crime), new IntWritable(1));
String report=s1[2];
            c.write(new Text(report), new IntWritable(1));
```

```
reduce(Text k,Iterable<IntWritable> v,Context c)
int count=0;
        while(v.iterator().hasNext())
                IntWritable i=v.iterator().next();
                count+=i.get();
        c.write(k,new IntWritable(count));
```

SIMULATION AND EXPERIMENTAL RESULTS

We have setup an experiment on Hadoop Cluster in Linux based Cent OS with single node running and configuration as: Minimum 2 GB RAM and Minimum 100GB Hard Disk Space. In this Crime Data Analysis, we are focusing on following three problem statements.

Table 1: Problem Definition

Problem #	Problem Statement	Result
Problem 1	Finding total number of crimes reported by	Fig. 7.1
Problem 2	Finding total number of Each Crime in last 6 Years (2011-2016)	Fig. 7.2
Problem 3	0.29 sec	Fig. 7.3

Using Sqoop Tool we are storing MySql data to HDFS. It can be done via following command. We can also revert the process i.e. HDFS to MySql [16].

Sqoop *->sqoop import --connect jdbc:mysql://localhost/training --username training --password training --table sr --m 1 --target-dir Crime_1;*

By using above statement, we can fetch the data present on the HDFS, Stored data further can be analyzed by hive and pig, results of each problem is summarized in the following section.

Hive analysis

Hive information is placed in HDFS, which is additionally sent to different nodes. This information is placed in a plain document with CSV, as provided by Sqoop. Hive will read entire file using indexing which results to faster query output. Hive won't execute a MapReduce Task if it does not include either of join, group by, order by aggregate operation. By querying this operation, hive can promptly begin the MapReduce task, which may requires 5-10 seconds to begin the MapReduce. [Table 2] shows Hive analysis. Solution using Hive for -

- ***Problem 1:*** *hive> select report_by,count(*)as tot from cdata group by report_by order by tot desc;*
- ***Problem 2:*** *hive> select" TOTAL CRIME FOR LAST 8 YEARS ", crime_type,count(*)as tot from cdata group by crime_type order by tot desc ;*
- ***Problem 3:*** *hive> hive> select year,count(*)as tot from cdata group by year order by tot desc;*

Table 2: Hive result analysis

Tool Hive:	Starting Time	Finishing TIme	Finished In (Sec)	# of Mappers	# of Reducers	Status
Problem 1	14.31:02	14.31.59	28.74	2	1	Successful
Problem 2	15:02:21	15:02:34	13.01	2	1	Successful
Problem 3	15.43:26	15:43:47	21.30	2	1	Successful

Pig analysis

Pig performs well with huge size of data. Pig executes a well ordered approach as characterized by the developer. If the query given by the developer is not a complex one (query which is included with joins and sorts) then Pig will not work properly. Pig solves every step one by one, which can expend more time in this case. Whenever information need composite job and more joining operation then Pig can deal with it productively by processing every level and persistently processing the next levels. Pig uses Grunt Shell to execute its task. [Table 3] shows pig analysis which is done after executing following script. Solution using Pig, for –

```
Problem 1: grunt> cri1 = LOAD '/user/training/cri ' using PigStorage(',') AS
(mon:int,year:int,report_by:chararray,loc:chararray,losa_code:chararray,losa_name:chararray,type_crime:
chararray);
grunt> cri2 = FOREACH cri1 GENERATE report_by;
grunt> by_loc = group cri2 by report_by;
grunt> count_crimes = foreach by_loc generate group as loc,COUNT(cri2) as TOT;
grant> cri6 = order count_crimes by TOT desc;
grunt> dump cri6;
```

```
Problem 2: grunt> cri1 = LOAD '/user/training/cri ' using PigStorage(',') AS
(mon:int,year:int,report_by:chararray,loc:chararray,losa_code:chararray,losa_name:chararray,type_crime:
chararray);
grunt> cri2 = FOREACH cri1 GENERATE type_crime;
grunt> by_type_crime = group cri2 by type_crime;
grunt> describe by_type_crime ;
grunt> count_crimes = foreach by_type_crime generate group as type_crime,COUNT(cri2) as TOT;
grant> cri6 = order count_crimes by TOT desc;
grunt> dump cri6;
```

```
Problem 3: grunt> cri2 = FOREACH cri1 GENERATE year;
grunt> cri3 = group  cri2 BY year ;
grunt> cri5 = foreach cri3 generate group as year,COUNT(cri2.year) as TOT;
grunt> cri6 = order cri5 by TOT desc;
grant>dump cri6;
```

Table 3: Pig Result Analysis

Tool Pig:	Starting Time	Finishing TIme	Finished In (Sec)	# of Mappers	# of Reducers	Status
Problem 1	14.47:19	14.47.59	33	2	1	Successful
Problem 2	15:10:51	15:11:34	15.56	2	1	Successful
Problem 3	15.52:06	15:52:47	24	2	1	Successful

Map reduce analysis

Here CrimeDriver class contain Mapper and Reducer methods. After performing a following command, it will start execution where mapping, shuffling and reduce process will happen. [Table 4] shows result for three given problem statements using MapReduce Program. And [Fig. 6] gives MapReduce paradigm. In MapReduce scenario, when the database is compiled for the given problem statements 1,2,3 it requires 36,19 and 30 seconds respectively where number of mappers are 2 and reducer is 1.

```
[training@localhost workspace]$ hadoop jar crime.jar CrimeDriver cri MROUT1]
```

Table 4: MapReduce Result Analysis

Tool MapReduce:	Starting Time	Finishing Time	Finished In (Sec)	# of Mappers	# of Reducers	Status
Problem 1	14:22:14	14.22.50	36	2	1	Successful
Problem 2	15.20.24	15.20.43	19	2	1	Successful
Problem 3	15:30:42	15:31:12	30	2	1	Successful

The given problem statements are used to analyze the Crime Data and results of those are shown on above charts and graph using R tool. In first problem statement, we are finding total number of crimes reported by, result is shown in figure [Fig.7(1)]. Cambridge shire Constabulary 454435 (51.3%), City of London Police 41745 (4.3%), Durham Constabulary 390220(44%). In second problem statement, we are finding total number of Each Crime in last 6 Years (2011-2016), result is shown in figure [Fig. 7(2)], where Anti-social behavior 357187 counts more number of crimes from year 2001 to 2016. In third problem statement, we are finding total number of crimes per Year. In this case, total number of crimes are counted for year 2011 to 2016 in which crimes happened in year 2015 is close to 20000 which is again more. Result is shown in figure [7(3)]. Following part will show [Fig. 7(1)], [Fig. 7(2)], and [Fig. 7(3)] and also gives us an analysis using each of the Hadoop eco-system.

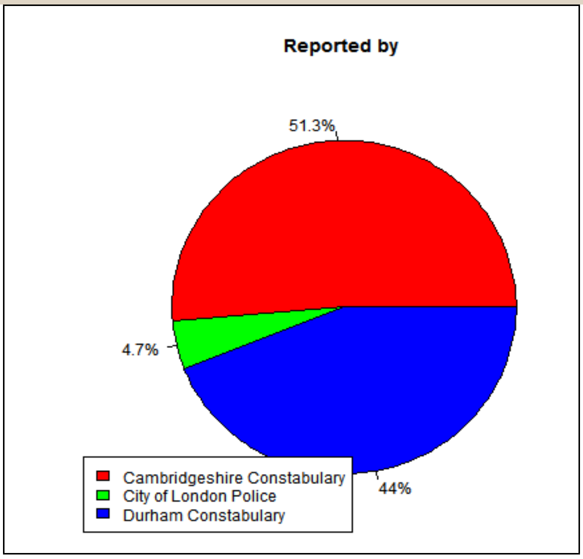

Fig. 7(1): Result for Problem #1

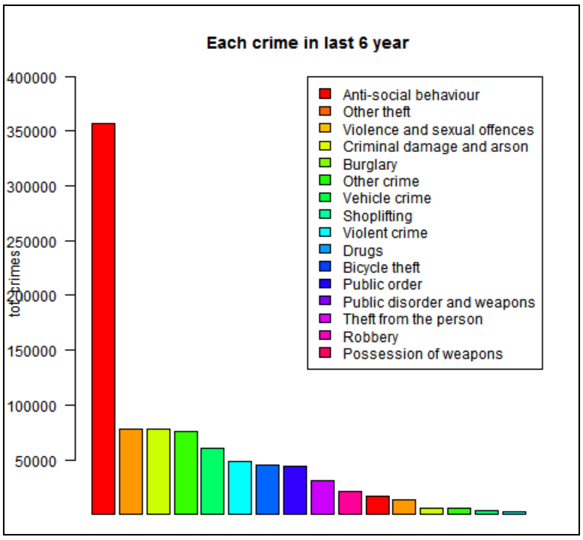

Fig. 7(2): Result for Problem #2

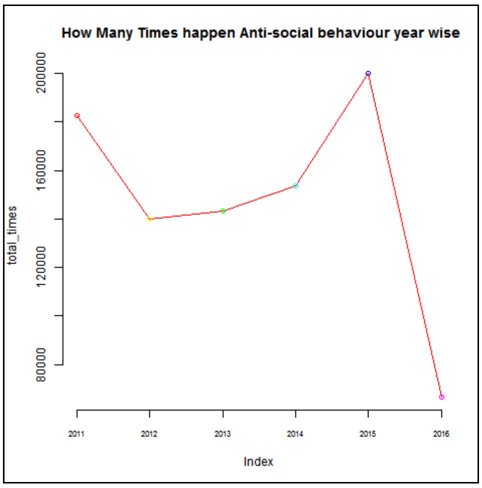

Fig. 7(3): Result for problem #3.

[Fig. 8] shows a graph where, we have analyzed the performance for each of the Hadoop Eco-system tool – Hive, Pig, MapReduce. Results of the graph shows that Hive works efficiently as compared to Pig and

MapReduce for the given use case scenario: Crime Data Analysis. Similarly, we can analyze different problems related to crime data and results of this analysis which can be helpful for restricting crimes in future. If we consider average execution time for Hive, Pig and MapReduce, we can conclude that Hive is 2.23 times faster than MapReduce and 1.90 times than Pig for the given scenario. Also, average line of code used by Hive and Pig are very lesser than MapReduce Program. [Table 5] describes the comparison of each tool with respect to average time complexity and number of code lines.

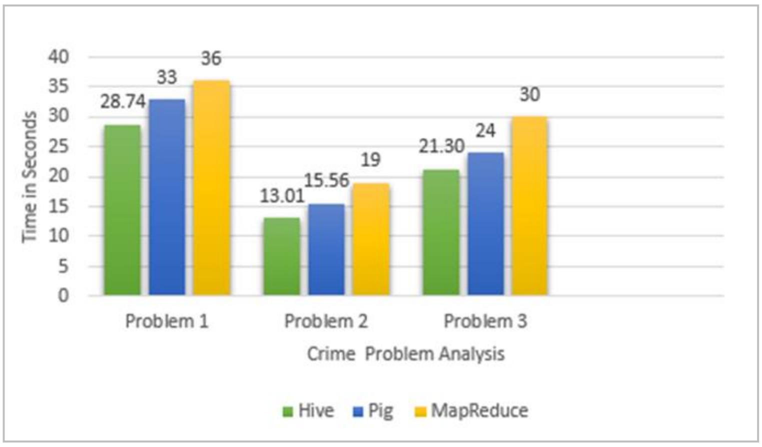

Fig. 7(4): Enhanced hadoop performance analysis.

Table 5: Hive, Pig, MapReduce Performance Comparison

Tool	Average Total Time Taken (seconds)	Time Relative to MapReduce	Line of Code Usage
Hive	38.05	2.23 times	2-4
Pig	72.56	1.17 times	10-12
MapReduce	85	1 .0 times	70-100

CONCLUSION

This paper presents data analysis of huge dataset utilizing three unique things that are a piece of the Hadoop environment – Hive, Pig and MapReduce. The application presented here is a crime Data investigation. The issue is clarified top to bottom and after the simulation, results are shown for the three tools. Complete dataset is accessible from https://data.police.uk/data/ and after successful operations, additionally the R techniques used to display the analysis and plot the outcomes. Results are appeared for each of the three tools with 8lacs set of records. Results show that Hive is more efficient when compared to Pig and MapReduce which shows that it is enhancing Hadoop Performance for huge data set. Hive is 2.23 times faster than MapReduce and 1.90 times than Pig. Also, line of code used by Hive and Pig are very lesser than MapReduce Program. In the future part, focus should be on finding the different parameters which can minimize the Hadoop performance for large size of data using different strategies.

CONFLICT OF INTEREST

There is no conflict of interest.

ACKNOWLEDGEMENTS

I express my sincere gratitude to Prof. Jabanjalin Hilda, Senior Assistant Professor in VIT University - Vellore, Department of Software Systems (SCOPE), for her expert guidance and invaluable support in Big Data – Hadoop Technology. I am extremely thankful to VIT University – Vellore for providing the infrastructure facilities for this project and also thankful to my Departmental Dean, HOD, all the Staff Members and my batch mates for their encouragement throughout the course of present work. Finally, I would like to acknowledge with gratitude, the support, and love of my family – my parents, friends without them it would not have been possible.

FINANCIAL DISCLOSURE

None

REFERENCES

[1] "Introduction to Crime Data Analysis", Developed by Garner Clancey.

[2] Chun-Wei Tsai, Chin-Feng Lai, Han-Chieh Chao and Athanasios V. Vasilakos, "Big data analytics: a survey", Journal of Big Data 2015.

[3] Pual C. Zikopoulos, Chris Eaton, Dirk deRoos, Thomas Deutsch, George Lapis, "Understanding Big Data – Analytics for Enterprise Class Hadoop and Streaming Data", McGraw-Hill Osborne Media ©2011 book.

[4] H. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, et al., "Big data and its technical challenges," Communications of the ACM, vol. 57, pp. 86-94, 2014.

[5] S Vemula. "Hadoop Image Processing Framework."

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

[7] T. White, Hadoop: The definitive guide: "O'Reilly Media, Inc.", 2012.

[8] Herodotou H.[2011] Hadoop performance models". arXiv preprint arXiv:1106.0940

[9] Ammar Fuad, Alva Erwin, Heru Purnomo Ipung, [2014]Processing Performance on Apache Pig, Apache Hive and MySQL Cluster", International Conference on Information, Communication Technology and System,

[10] Swathi Prabhu, Anisha P Rodrigues, Guru Prasad MS & Nagesh HR.[2015] Performance Enhancement of Hadoop MapReduce Framework for Analyzing BigData, Electrical, Computer and Communication Technologies (ICECCT), 2015 IEEE International Conference on March 2015.

[11] Shiju Sathyadevan, Devan M.S, Surya Gangadharan. S, "Crime Analysis and Prediction Using Data Mining", Networks & Soft Computing (ICNSC), 2014 First International Conference on August 2014.

[12] Hamoud Alshammari, Jeongkyu Lee and Hassan Bajwa, "H2Hadoop: Improving Hadoop Performance using the Metadata of Related Jobs", IEEE TRANSACTIONS ON Cloud Computing 2015.

[13] Rachel Boba "Introductory Guide to Crime Analysis and Mapping", November 2001 Report to the Office of Community Oriented Policing Services.

[14] Arushi Jaina, Vishal Bhatnagara, [2015] Crime Data Analysis Using Pig with Hadoop", International Conference on Information Security & Privacy (ICISP2015), 11-12, Nagpur, INDIA

[15] Prachi Pandey, Sanjay Silakari, Uday Chourasia,[2016]A Comparative Study of Hadoop Family Tools", International Journal of Computer Science and Information Technologies, 7(3): 1620-1623

[16] http://sqoop.apache.org/docs/1.4.0-incubating/SqoopUserGuide.html

[17] http://hive.apache.org

[18] http://pig.apache.org

[19] Lu Jiamin, Feng Jun.[2015] A Survey of MapReduce based Parallel Processing Technologies", Big Data, Cloud & Mobile Computing,China Communication, Vol 11

[20] Yunquan Zhang, Ting Cao, Shigang Li, Xinhui Tian, Liang Yuan, Haipeng Jia, and Athanasios V. Vasilakos,[2016] Parallel Processing Systems for Big Data: A Survey, Proceedings of the IEEE 104(11).