Distributed Storage Architecture and Hadoop Framework Tools for Crunching Big Data in Clustered Processing System

Ramalingam. M, Prabhusundhar. P, Azhaharasan. V, Prabahari. R
Asst. Professors,
PG & Research Department of Computer Science,
Gobi Arts & Science College (Autonomous),
Gobi, Tamil Nadu, India.

Article Info
Volume 82
Page Number: 4659 - 4664
Publication Issue:
January-February 2020

Article History
Article Received: *18 May 2019*
Revised: *14 July 2019*
Accepted: *22 December 2019*
Publication: *23 January 2020*

Abstract:
Huge amount of terabytes data is being created in today's information technology world such as cloud computing, social media, Internet of Things (IoT) and Internet. Need an important tools to analysis and extract such huge data. Big Data tools allows to extract the information from unstructured format and keep them in the form of events, objects, entities, relations, table format and many other types. Information Extraction is the system where it can extract information from both structured data and non structured data. The basic purpose of this paper is to provide a good understand of Big data tools for extraction and analysis of terabyte data. Big Data has now begun to intervene in a variety of sectors such as astronomy, economics, chemistry, transport and research. Every department has now begun to store very sensitive information for its growth and functioning. Big Data is all about "big data", such as how to extract only the most important information (Data Mining), how to extract the extracted information into a data pipeline and how to submit it to the user. Big data not only stores the information in the orderly format, but also storing information in the unstructured form. Different kinds of logics were chosen to analysis the features suitability of the Big Data and various tools associated with it: the Kibana, Hadoop, HDFS, Pig, Hive and Spark.

Keywords: *Big Data, Hadoop, Pig, ELASTIC search, Structured and unstructured Data.*

I. INTRODUCTION

In this information technology world, 2.5 quintillion bytes of data people have created multiple platforms in a variety of ways. It should be noted that 90% of the data we have now is made in the last two years. The reason why so much of this kind of data is gathered the world and where these data is coming from. Importantly, the public posts their records on public social networks, such as pictures and videos, through the exchange of information. More information is coming from Weather information, Climate change, Information from satellites, GPS signal and purchase transaction. However it is not limited to these issues. High health care related research issues can be found in Husing Kuo et al. Paper [9]. Information that occurs during transactions in banks, to keep this banking detail safe also. This great information, this technique of Big Data was born many years ago but we need it now to handle huge data. Effective integration of technologies and analysis will result in predicting the future drift of events. We need some knowledge and technology to deal with this mega amount of Big Data. Need a lot of mathematical and statistical knowledge and specialized tools and analytic languages to handle this lot of information. Examples of tools that have been released in the modern day include special languages such as the R, Python, hadoop,and so on.

The research issues pertaining to big data analysis are classified into three broad categories namely quantum computing, bio inspired computing, internet of things (IoT) and cloud computing.

Developing tools using different technologies to understand. Eventually they put those tools together and publish them as a package: a tool for big data that the company provides Hadoop, Spark, Druid and ELK are open source software tools for big data that are currently popular in the market. These are provided by companies like Apache, Cloudera, Amazon. Store all kinds of information, such as your mood, likes, dislikes, activities and reviews along with the basic details of the user, such as the phone on Facebook. All of these can be in any form. One can express his mood as words and images. So it cannot predict what shape these will be. These are stored as unstructured data.

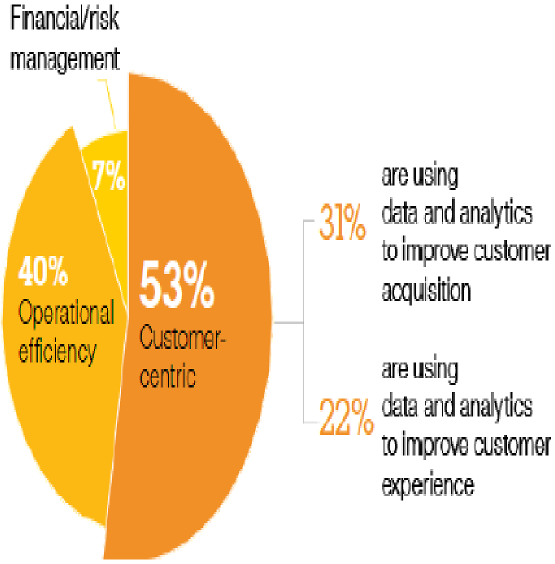


Fig.1 IBM-Analytic Study [2014]

II. CHALLENGES IN BIG DATA ANALYTICS

Over the past few years this Big Data technology has occupied many sectors. These occupied sectors are the public sector such as social network. The fields it occupies are mentioned as follows : occupies fields such as Health and Safety, Public Sector Organizations, Biological Departments, Chemicals, research, Satellite and digital information, Social Websites. We need unique knowledge, technology and tools to deal with the huge amounts of data that can grow in many of these fields. Social computing includes social network analysis, online communities, recommended systems, reputation systems, and prediction markets where as internet search indexing includes ISI, IEEE Xplorer, Scopus, Thomson Reuters etc.Need to store data more securely when we handle growing data, especially in sectors such as the Bank and Stock Stock Market. When dealing with a lot of this information, we deal with it in many different ways such as storing and analyzing information, finding the knowledge, Third, the difficulties that arise when analyzing information, fourth is to find the information that is growing and store it properly, to find emerging data and store it properly, the difficulties that arise when analyzing information. Next we store the information securely and finally get the output (decision making data) from the big data.

III.CLUSTER MANAGEMENT-YARN

It is a software application that enables cluster management of various tasks. DFS (distributed file system) and MongoDB come in many different ways.

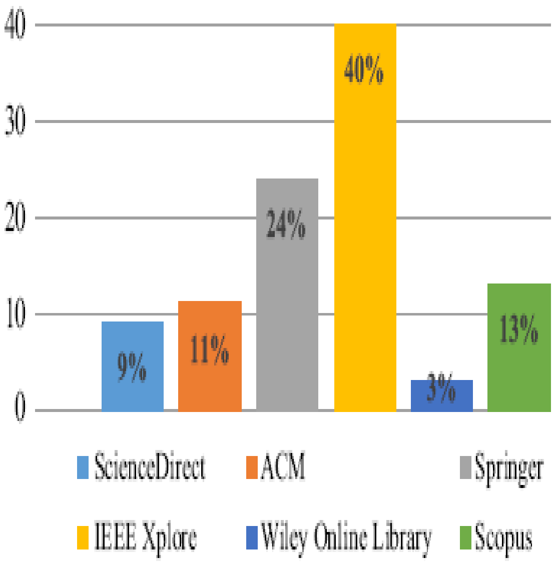
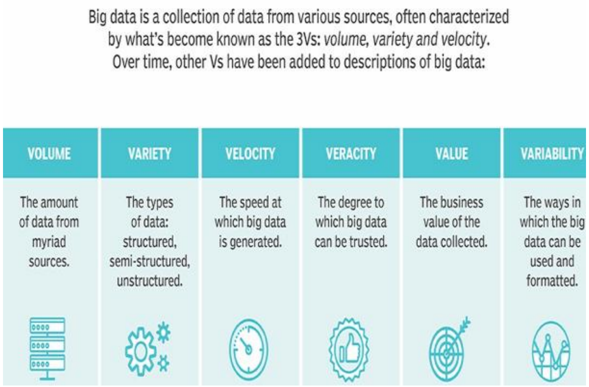


Fig.2 illustrates the selected studies distribution over data sources [KiranAdnan and Rehan Akbar, 2019]

These are storage warehouses that help store information. Mapper takes daughter data and compresses it into key (value) pairs.



Big data is a collection of data from various sources, often characterized by what's become known as the 3Vs: *volume, variety and velocity.* Over time, other Vs have been added to descriptions of big data:

VOLUME	VARIETY	VELOCITY	VERACITY	VALUE	VARIABILITY
The amount of data from myriad sources.	The types of data: structured, semi-structured, unstructured.	The speed at which big data is generated.	The degree to which big data can be trusted.	The business value of the data collected.	The ways in which the big data can be used and formatted.

Fig.3 Volume, Variety and Velocity: Big Data

Reducer is a Mapper: compressor that can re-compress and store in a meaningful way. Mapper and Reducer are both called 'MapReducer'. It is a library written using Java. Pig & Hive are languages that use the Map Reducer library. All of these are created using a structure called Hadoop. Next up is ELK, a framework that combines Elastic Search, Logstash & Kibana 3 to help authorities report as they wish. Elasticsearch is an engine that helps to store first level information. Logstash is a tool that lets to enter information in a file format or website into the Engine. Kibana is a tool that searches and discloses information from the Engine for reporting purposes. Like Spark and Druid, each has its own structure. The company, along with the tools they developed, came up with the name "ELK Stack" as a tool for big data. Information from the Engineer is required for reporting T is a tool that can be used to reveal the functions of each of these tools.

ELK Stack: An Introduction ELK Stack is a combination of 3 separate open source software tools called Logstash, Elastic Search, and Kibana. These were developed by individual individuals in 2009, 2010 and 2011 respectively and emerged as separate open source tools. Elastic Search ElasticSearch is a storage area & search engine that enables To quickly display information that the hear when ask for it. It is designed to help in search. GitHub, Google, Stackoverflow, Wikipedia and so on. In addition to storage, it also accelerates users' ability to search for information at their discretion. For example, when it search for something on Google, it will first find links that are ompletely relevant to the words entered, followed by links that have somewhat related words. Likewise, when hit words to search for, it would suggest us words that are relevant to the search in the name of 'suggessions'. These are the features of Elastic Search.

IV. HADOOP FRAMEWORK

Already seen that "Hadoop is not a unique tool; it is a combination of various miniature tools!" The most important members are HDFS and MapReducer.

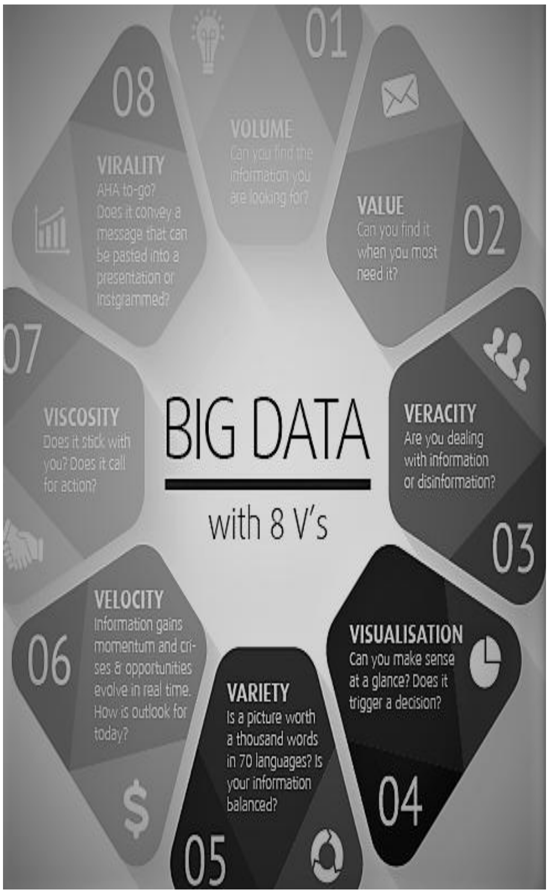
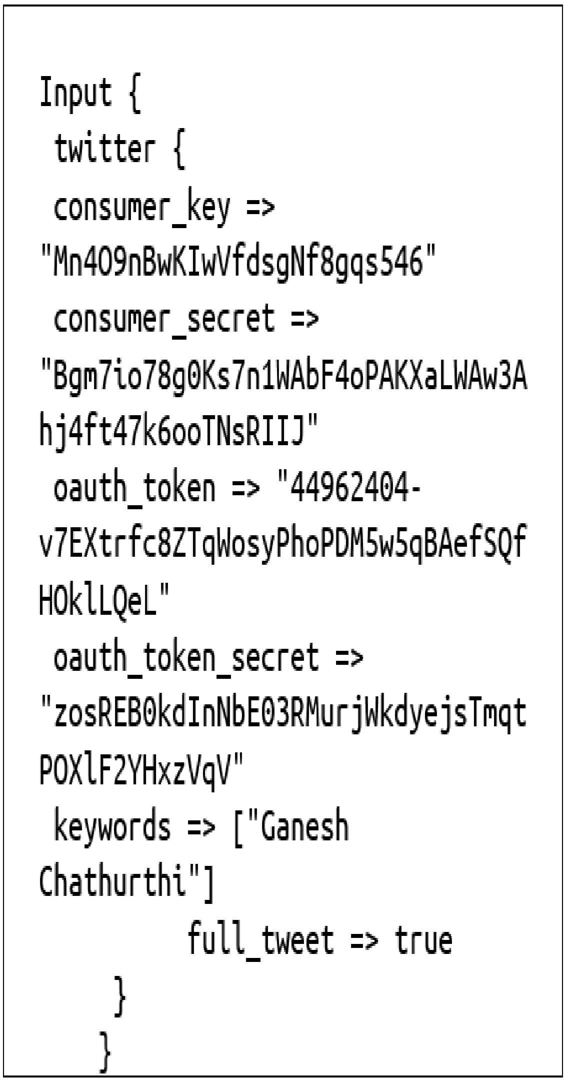

01
VOLUME
Can you find the information you are looking for?
02
VALUE
Can you find it when you most need it?
03
VERACITY
Are you dealing with information or disinformation?
04
VISUALISATION
Can you make sense at a glance? Does it trigger a decision?
05
VARIETY
Is a picture worth a thousand words in 70 languages? Is your information balanced?
06
VELOCITY
Information gains momentum and crises & opportunities evolve in real time. How is outlook for today?
07
VISCOSITY
Does it stick with you? Does it call for action?
08
VIRALITY
AHA to-go? Does it convey a message that can be pasted into a presentation or Instgrammed?
BIG DATA
with 8 V's

Fig.4 Big data with eight V's

They perform important tasks of storing and analyzing data, respectively. These include "Spark, Hive, Pig, HBase, Phoenix, , Zookeeper, Sqoop, oozie" and

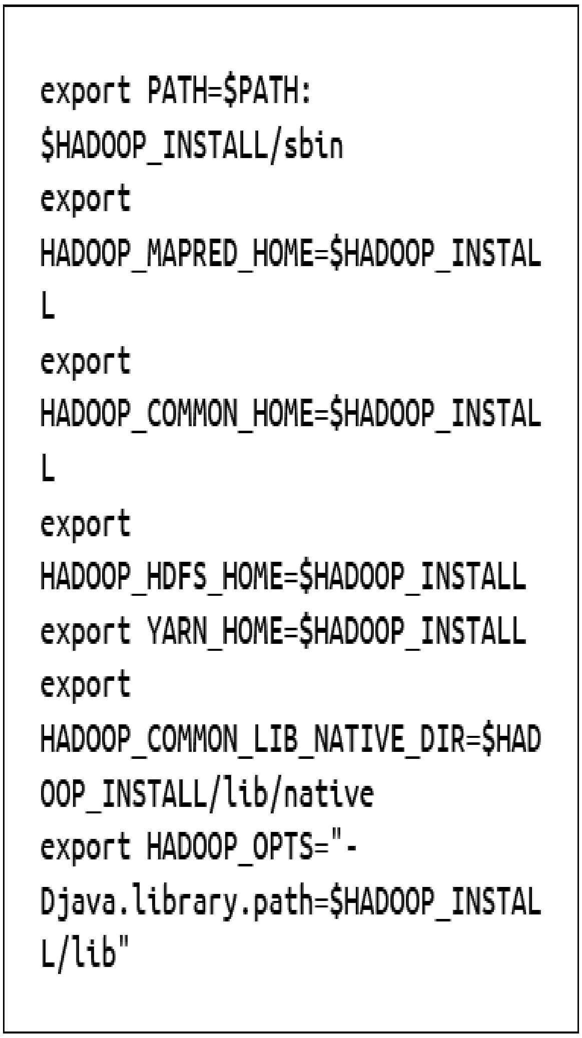
queues and shell command outputs help to get programs from a variety of sources. What it refer to as the show is just information! This is the first stage. Next, they convert the information receive to the data processing it need.



```
Input {
 twitter {
 consumer_key =>
"Mn409nBwKIwVfdsgNf8gqs546"
 consumer_secret =>
"Bgm7io78g0Ks7n1WAbF4oPAKXaLWAw3A
hj4ft47k6ooTNsRIIJ"
 oauth_token => "44962404-
v7EXtrfc8ZTqWosyPhoPDM5w5qBAefSQf
HOklLQeL"
 oauth_token_secret =>
"zosREB0kdInNbE03RMurjWkdyejsTmqt
POXlF2YHxzVqV"
 keywords => ["Ganesh
Chathurthi"]
          full_tweet => true
     }
    }
```

Fig.5 Hadoop-env.sh

Logstash receives events: Its name may be called LogStash. For this it is not a log only tool. Retrieves all status data in a variety of formats such as Files, sockets, script outputs.



```
export PATH=$PATH:
$HADOOP_INSTALL/sbin
export
HADOOP_MAPRED_HOME=$HADOOP_INSTAL
L
export
HADOOP_COMMON_HOME=$HADOOP_INSTAL
L
export
HADOOP_HDFS_HOME=$HADOOP_INSTALL
export YARN_HOME=$HADOOP_INSTALL
export
HADOOP_COMMON_LIB_NATIVE_DIR=$HAD
OOP_INSTALL/lib/native
export HADOOP_OPTS="-
Djava.library.path=$HADOOP_INSTAL
L/lib"
```

Fig.6 Config File-"# diwali"

Different types of plugins work with logstash to extract events from different sources. These are called input plugins. For example, file, irc, jdbc, kafka, github exec, eventlog, http, imap etc files, servers, webhooks. Zeam out message This is the basis on which information is divided, what should be changed and what should be deleted. This is the secondary. The last formatted data is stored in a variety of repositories. There are various plugins to add with Zoom Out. These are called output plugins. For example, Csv, datalog, email, irc, jira, exec, kafka, elasticsearch etc can be added to a variety of store locations such as files, servers, message queues, storage engine, databases. This is the third stage. Following is the configuration file for embedding current data such as Twitter. This can now be understood by themselves. Here, have taken those who tweeted "# diwali" and put them into Elastic Search via logstash. The input plugin for twitter {} has been used and information has been entered. They are inserted into elastic search under the index name 'twitter _ elastic _ example' without any changes. **Kibana Kibana** is a Visual Interface which helps to convert data in Elasticsearch into graphs. Kibana's maps help to make some important decisions by keeping the data in ElasticSearch. It can also be called ReportingTool. It is a little more difficult to make a few important decisions with just a handful of information.

V. CONCLUSION

The amounts of data increases exponential, the current techniques are becoming obsolete. Dealing with Big Data requires comprehensive coding skills, domain knowledge and statistics. Developing tools using different technologies to understand. Eventually they put those tools together and publish them as a package: a tool for big data that the company provides Hadoop, Spark, Druid and ELK are open source software tools for big data that are currently popular in the market. Social computing includes prediction markets,reputation systems,, recommended systems, online communities and social network analysis where as internet search indexing includes Thomson Reuters, IEEE, Scopus, ISI etc. Discussed the ELK Stack and its combination of 3 separate open source software tools called Logstash, Elastic Search and Kibana. This paper explains, how to uses the Pig and to handle files when they are the same. It discussed the programming language called PigLatin for data processing.

REFERENCES

[1]. Singh M. PetterSanchitaPaul, "Big Data tool and Challenges and Solutions‖", International Conference Big Data, December- 2012.

[2]. Lisa M, "Five Emerging Challenges and Big Data and warehouse", IJRES ,IISN:234-123-2014

[3]. R. Jeeven, et al., "Research challenges and opportunities of big data analytics", ACM conference, 2013

[4]. E.L Chatter and T. Rewql, "developing tool and challenges of big data computing in health care science", National Conference on Optimization tools-Big Data Research, Page number- 2 to 11, year:2013.

[5]. J. Merlin and W.x. Vermel, "Big data and machine learning methods", International Journal of Science, Page number 50 to 57, Publication year 2014.

[6]. V.Thiyagu and Shan X, "Unsupervised learning and data analysis", Journal of Big Data Research", Page number 59 t0 64. Publication year:2015

[7]. Haider M, "Big data concepts methods and analytic", IJIM journal of computer science, Page number 137 to 144, publication year 2017.

[8] E.Booms and V. Varma, "mathematical and statistical knowledge and specialized tools", Journal of ACM, Page number: 36 to 44, Publication year 2018

[9] Zlhu, Xyuu and Y. Pelyt, "Security analysis of big data in bio-metric security", International Conference on Information Technology and Management Innovation, Page number 1041 to 1044, Conference year: 2015.

[10] E. L Tongjun, and Wuxing, "Research on data science and information security in big data tools", Ijisr journal, Page number 1 to 6, publication year 2014.

[11] MerelliPerez-sanchez H, and Simbath D, "Integrating big data in bioinformatics and security", Journal of Bio-Med , Page number: 1 to 13, publication year: 2015.

[12] Mishra N, Lin and Chang H, ": open problems and future perspectives of Data analysis", Distributed Sensor Networks:International Journal, Page number: 1 to 14, Publishing year:2015.

[13]. V. Ragu, http://thing solve r.com/anoma ly-detec tioni/, Data ElasticSearch in Big data.

[14]. E. QuestrnWang, "Intrusion detection systems and Kibana search systems", Comput Netw, page number 48 to 55, year of publication :2017.

[15]. E.F. Code, "Data Analysis in Western tamilnadu systematical data", WWW.Tn-satc",

[16].E.P.Palanisamy ,Archive,http://elts.wide.ad.jp/mawi/.Elastic Search, Project:year: 2012.

[17]HadoopProject,http://www.hadoop .ad.jp/.

[18]. T.T. Devi,"Spark data tools and its advantages" , Journal of GASCDE, 2017.

[19] R.K.T ashokvi, "Hive tools and its impact on big data", GDCOVAI, Conference on big data, 2016.

[20] M. Ramalingam, P. Prabhusundhar, R.Prabhahari, V.Azhaharasan, K. Yuvaraj, "Classifications of Biometric Security Authentication and Cluster Based Intrusion Detection, Prevention Technique in Mobile Ad-hoc Network", International Journal of Scientific Research in Computer Science Applications and Management Studies, Journal ISSN 2319 1953, 7th volume of 4th issue, Publication of year: 2018-July

[21]Understanding-lstm-and-its-diagrams, medium, WWW.medium.com/mlrev iew/under stand ing-lstm-and-its-diagrams.

[22] C. Noel and Osindero Dogwild S, "distributed data and GIS data analysis", LPT-workshop on Distributed machine learning , Workshop year: 2014.

[23] P.L naki, "Research issues in big data tools and data analytics", Page number 228 to 232, Journal of data analysis in Management, Year of publication: 2015.