iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 **Issue:** X **Month of publication:** October 2016

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

A Research on Big Data Mining with Hadoop F/W and It's Echo Systems

Chiranjeevi. B[1], G. Sravan Kumar[2]
[1]M. Tech Student, [2]Associate Professor
Dept. of CSE, Sreyas Institute of Engineering & Technology, Hyderabad, T.S, India

Abstract: ***Nowadays, most of the information saved in companies are unstructured model. In fact, Big data is term refer to huge data sets, have high Velocity, high Volume and high Variety and complex. Due to characteristic of big data it becomes very difficult to Management, Analysis, Storage, Transport and processing the data using the existing traditional techniques. This paper introduces Big Data Analysis and storage, First we presents the Big data technology alongside it's the significance of big data in the modern world and venture existing which are successful and essential in changing the idea of science into huge science and society as well. Following that, we present How fast Data is increasing and the important of Big Data. Analyzing Big Data is a challenging task as it involves large distributed file systems which should be fault tolerant, flexible and scalable. Map Reduce is widely been used for the efficient analysis of Big Data. With the fast growth of networks now-a-days organizations has filled with the collection of millions of data with large number of combinations. This big data challenges over business problems. It requires more analysis for the high-performance process, The new methods of hadoop and Its' Echo systems are discussed from the data mining perspective.***

Key words: Big Data, Data Mining , Hadoop, HDFS, MapReduce, HIVE, PIG, Flume, HBase, Mahout, Sqoop, Oozie, Spark.

I. INTRODUCTION

Big data is the term used to describe huge datasets having the "3 V" definition: volume, variety, velocity and value(e.g. medical images, electronic medical record(EMR), biometrics data, etc.). Such datasets present problems with storage, analysis and visualization. To deal with these challenges, new software programming to multithread computing tasks have been developed.

A. Characteristics Of Big Data

As the data is too big and comes from various sources in different form, it is characterized by the following 3 v's.

1) *Variety:* Data being produced is not of single category as it not only includes the traditional data but also the semi structured data from various resources like web Pages, Web Log Files, social media sites, e-mail, documents, sensor devices data both from active passive devices. All this data is totally different consisting of raw, structured, semi structured and even unstructured data which is difficult to be handled by the existing traditional analytic systems.
2) *Volume:* The Big word in Big data itself defines the volume. At present the data existing is in petabytes and is supposed to increase to zettabytes in nearby future. The social networking sites existing are themselves producing data in order of terabytes everyday and this amount of data is definitely difficult to be handled using the existing traditional systems.
3) *Velocity:* Velocity in Big data is a concept which deals with the speed of the data coming from various sources. This characteristic is not being limited to the speed of incoming data but also speed at which the data flows. For example, the data from the sensor devices would be constantly moving to the database store and this amount won"t be small enough. Thus our traditional systems are not capable enough on performing the analytics on the data which is constantly in motion.

II. BIG DATA MINING

The term 'Big Data' appeared for first time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the title of "Big Data and the Next Wave of InfraStress" . Big Data mining was very relevant from the beginning, as the first book mentioning 'Big Data' is a data mining book that appeared also in 1998 by Weiss and Indrukya . However, the first academic paper with the words 'Big Data' in the title appeared a bit later in 2000 in a paper by Diebold . The origin of the term 'Big Data' is due to the fact that we are creating a huge amount of data every day. Usama Fayyad in his invited talk at the KDD BigMine'12 Workshop presented amazing data numbers about internet usage, among them the following: each day Google has more than 1 billion queries per day, Twitter has more than 250 milion

tweets per day, Facebook has more than 800 million updates per day, and YouTube has more than 4 billion views per day. The data produced nowadays is estimated in the order of zettabytes, and it is growing around 40% every year. A new large source of data is going to be generated from mobile devices, and big companies as Google, Apple, Facebook, Yahoo, Twitter are starting to look carefully to this data to second useful patterns to improve user experience. Alex 'Sandy' Pentland in his 'Human Dynamics Laboratory' at MIT, is doing research in _nding patterns in mobile data about what users do, and not in what people says they do. We need new algorithms, and new tools to deal with all of this data.

A. *Types Of Big Data And Sources*

There are 3 types of big data:

1) *Structured Data:* Data which has proper structure and which can be easily stored in tabular form in any relational databases like Mysql, Oracle etc. is known as structured data. Example- Employee data .
2) *Semi-Structured Data:* Data which has some structure but cannot be saved in a tabular form in relational databases is known as semi structured data.
   Example-XML data, email messages etc.
3) *Unstructured Data:* Data which is not having any structure and cannot be saved in tabular form of relational databases is known as unstructured data.
   Example- Video files, Audio files, Text file etc.

In general, big data shall mean the datasets that could not be perceived, acquired, managed and processed by traditional IT and software/hardware tools within a tolerable time. Big Data describes any massive volume of structured, semi structured and unstructured data that are difficult to process using traditional database system such as RDBMS. An example of big data may be Exabyte's (1024 terabytes) of data consisting of trillions of records of millions of people from different sources such as websites, social media, mobile data, web servers, online transactions and so on. In the past, type of information available was limited. There was a well-defined set of technology approaches for managing information. But in today's world, the amount of data has been exploding. It has grown to terabytes and petabytes. Because in every minute, there are 280,000 tweets, more than 100 millions emails are sent. 2 million searching queries in Google, and more than 350 GB of data is processed in face book in every minute. Some of the applications of big data are in areas such as social media, healthcare, traffic management, banking, retail, education and so on.

III. TOOLS AND TECHNOLOGIES

For the purpose of processing the large amount of data, the big data requires exceptional technologies. The various techniques and technologies have been introduced for manipulating, analyzing and visualizing the big data . There are many solutions to handle the Big Data, but the Hadoop is one of the most widely used technologies .

A. *Why is Hadoop important?*

1) Ability to store and process huge amounts of any kind of data, quickly. With data volumes and varieties constantly increasing, especially from social media and the Internet of Things (IoT), that's a key consideration.
2) Computing power. Hadoop's distributed computing model processes big data fast. The more computing nodes you use, the more processing power you have.
3) Fault tolerance. Data and application processing are protected against hardware failure. If a node goes down, jobs are automatically redirected to other nodes to make sure the distributed computing does not fail. Multiple copies of all data are stored automatically.
4) Flexibility. Unlike traditional relational databases, you don't have to preprocess data before storing it. You can store as much data as you want and decide how to use it later. That includes unstructured data like text, images and videos.
5) Low cost. The open-source framework is free and uses commodity hardware to store large quantities of data.
6) Scalability. You can easily grow your system to handle more data simply by adding nodes. Little administration is required.

B. *Use Cases Of Hadoop*

Linkedin manages over 1 billion personalized recommendations every week. All thanks to Hadoop and its MapReduce and HDFS features!

Hadoop is at its best when it comes to analyzing Big Data. This is why companies like Rackspace uses Hadoop.

Hadoop plays an equally competent role in analyzing huge volumes of data generated by scientifically driven companies like Spadac.com.

Hadoop is a great framework for advertising companies as well. It keeps a good track of the millions of clicks on the ads and how the users are responding to the ads posted by the big Ad agencies!

Hadoop managed by the Apache Foundation is a powerful open-source platform written in java that is capable of processing large amounts of heterogeneous data-sets at scale in a distributive fashion on cluster of computers using simple programming models. It is designed to scale up from single server to thousands of machines, each offering local computation and storage and has become an in-demand technical skill. Hadoop is an Apache top-level project being built and used by a global community of contributors and users.

IV. HADOOP ECOSYSTEM

Hadoop is not like another technology where you can learn or offer services, rather than this Hadoop is a platform where you can build splendid applications. From the rise of the Hadoop there are dozens of software communities have been developing modules which can address variety of problem specs and meeting different needs.

In this paper we are going to get brief introduction about the world of Hadoop and its Ecosystem. With Hadoop Ecosystem there are tons of commercial as well as open source products which are widely used to make Hadoop laymen-accessible and more usable. Here we are going to discuss about some specific components which gives core functionality and speed to Hadoop.

A. HDFS – Hadoop Distributed File System

When file size grows from GBs to TBs, we need some concrete system who can manage this entire file. As we all know, that database can be distributed over multiple servers, as the same way when multiple file size increased, a single machine cant process that files, we need distributed file system (who manage storage across network). HDFS is specially designed to store huge amount of file size where operations are write once and read multiple times are done. For Low Latency data access, HDFS is not a good option, as there are lots of small file included in operation for the update purpose.

B. Map Reduce

Hadoop MapReduce is a software framework for easily writing applications which process big amounts of data in-parallel on large clusters of commodity hardware in a reliable, fault-tolerant manner. In terms of programming, there are two functions which are most common in MapReduce. In terms of programming, there are two functions which are most common in Map Reduce.

1) Map
2) Reduce

In map step, Master computer or node takes input (Large data set to process) and divide it into smaller parts and distribute it on other worker nodes (other machine/Hardware). All worker nodes solve their own small problem and give answer to the master node.

In Reduce step, Master node combines all answers coming from worker node and forms it in some form of output which is answer of our big distributed problem.

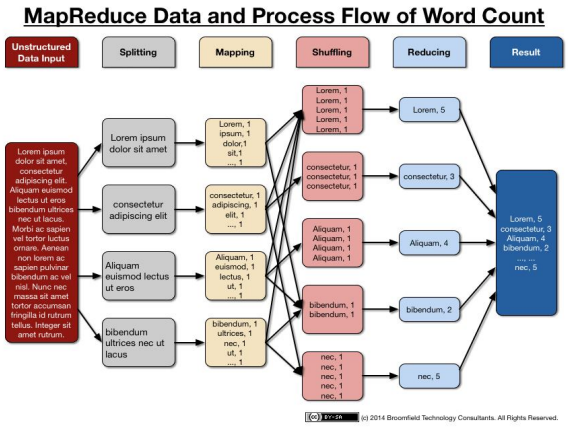


Figure 1.MapReduce Process for WordCount Program

C. Hive

Hive is part of the Hadoop ecosystem and provides an SQL like interface to Hadoop. It is a data warehouse system for Hadoop that facilitates easy data summarization, ad-hoc queries, and the analysis of large datasets stored in Hadoop compatible file systems.
It provides a mechanism to project structure onto this data and query the data using a SQL-like language called HiveQL. Hive also allows traditional map/reduce programmers to plug in their custom map-pers and reducers when it is inconvenient or inefficient to express this logic in HiveQL.
The main building blocks of Hive are –

1) Metastore: To store metadata about columns, partition and system catalogue.
2) Driver: To manage the lifecycle of a HiveQL statement
3) Query Compiler: To compiles HiveQL into a directed acyclic graph.
4) Execution Engine: To execute the tasks in proper order which are produced by the compiler.
5) HiveServer: To provide a Thrift interface and a JDBC / ODBC server.

When all these components are merged, it makes the Hadoop very user friendly.

D. HBase

HBase is a distributed, column oriented database and uses HDFS for the underlying storage. As said earlier, HDFS works on write once and read many times pattern, but this isn't a case always. We may require real time read/write random access for huge dataset; this is where HBase comes into the picture. HBase is built on top of HDFS and distributed on column-oriented database.
Here are the main components of HBase:
HBase Master: It is responsible for negotiating load balancing across all RegionServers and maintains the state of the cluster. It is not part of the actual data storage or retrieval path.
RegionServer: It is deployed on each machine and hosts data and processes I/O requests.

E. Mahout

A success of big data is depends on how fast and efficiently you can convert a vast amount of information to some actionable information. Whether it is analyzing all log (visited web pages) of Cyberoam for particular company, like Azilen Technologies, to know which site is most frequently opened in Azilen or if you want to proceed thousands of personal email messages and to generate analytics and then to organize and extract information, Mahot is being used here.
If you have seen the English season "Person of Interest" you may found that in this serial Person builds a machine to collect the data of people of New York City to identify terrorists and machine separates these people in two sections relevant and irrelevant. Relevant people are people who can be threat to national security and others are irrelevant. As the Same way Mahout Recommender Engines tries to identify unknown items from their area of interest and distributes in more categorized way.

F. Sqoop

To understand Sqoop, let's say we have released one version of software and now it's in production and you are implementing some new features and then releasing a new version which requires migration from old data to new data. In the same way, even if you love the new concepts of big data, your existing data is still in SQL over distributed server. Now, how to convert/migrate these data in such form that it can be useful to Big Data concepts. This is the scenario where Sqoop comes in the picture.
To load bulk data from production system and to access it in mapreduce application can be a challenging task. Transferring data using SQL or other technology scripts is inefficient and time-consuming. Sqoop is similarly to a SQL Server Integration Service which allows easy import and export of data from structured data such as relational databases, enterprise data warehouses, and NoSQL systems. Sqoop also sliced up data into different partitions and a map-only job is launched with individual mappers responsible for transferring a slice of this dataset.

G. Pig

To analyze and querying a huge data set is not easy with any ordinary language. When we are talking about analyzing Pera bytes of data, it requires extensive level of parallel mechanism to analyze data.
Pig is used to analyze and to fire query on large data set that consists high level language which is for expressing data analysis programs, coupled with infrastructure for evaluating these. Pig Latin has main three key properties:

1) Extensibility
2) Optimization opportunities
3) Ease of programming

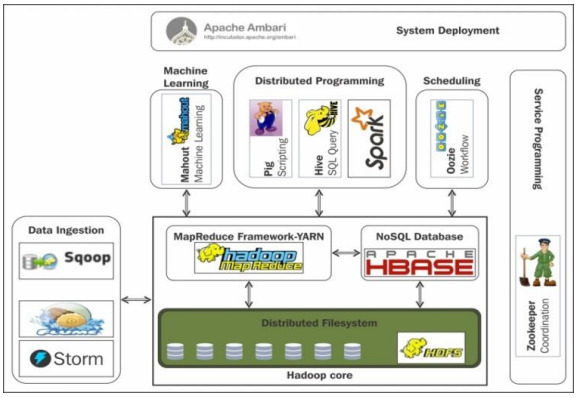
H. Flume

In real world there are many common problems which can't be solved by the existing components like if I have well defined structured database and ask help to some of software engineer then each one can implement their logic and do analytics for me. But, data is not always in structured manner. Let's take one example, if you assume Google is taking logs. I assume that Google has very strategic and good way to do this but this is only an assumption that on each day every keyword typed in Google search box are written in file. Now, do analytics for Google. I guess you are getting my points. All logs, whether they are event based or specific action based, are most difficult to analyze. these data that is where I may present Flume.

Flume is a utility for harvesting, aggregating and moving huge amounts of log data or text files in and out of Hadoop. Input of flume can be anything like Avro (another Hadoop component), files, system logs, HDFS, HBase. Flume itself has a query processing engine, so it's easy to transform each new batch of data before it is shuttled to the intended sink.

I. Apache Spark

Apache Spark is a general compute engine that offers fast data analysis on a large scale. Spark is built on HDFS but bypasses MapReduce and instead uses its own data processing framework. Common uses cases for Apache Spark include real-time queries, event stream processing, iterative algorithms, complex operations and machine learning.



Pigure 2. Understanding Hadoop Ecosystem

From All above discussed components just few are well known components which are developed by famous software companies and each one is designed for very specific purpose so, for Hadoop We can say that it's not single person or company's idea to build it. With Hadoop and its ecosystem most interesting thing I found is names of each component. You can also see that each component has unique name and every name has its own story. Hadoop is powerful because it is extensible and it is easy to integrate with any component.

J. Power of Hadoop

Well, Hadoop is powerful yet widely used by the big brands in the real world. In this section I am going to share some case studies which tell their stories itself.

Case Study – 1: Last.fm

Last.fm is personalized radio which is having 40M unique users and 500M different page view each month. Every page view leads to at least one log line and different actions leads to multiple logs. Site is doing analytics of this vast data set and they find site states, reporting charts via Hadoop.

Case Study – 2: Our favorite Facebook

To manage large size of Facebook users, to determine application quality, to Generate statistics about site usage, Facebook is using multiple Hadoop Clusters. With use of Hadoop Cluster, Facebook is loading 250 GB every day and have hundreds of jobs running each day on these data. An amazingly large fraction of Facebook engineers have run Hadoop jobs at some point.

Case Study – 3: Google itself

Google is always looking for better & faster sorting algorithm. Google was able to sort 1TB size of files stored on Google file system in 68 Seconds with use of Hadoop MapReduce Programs. It's a Google so they are not stopped with 1TB files they have experiment to sort 1PB data in (1000 TB) 6 hours and 2 minutes using 4000 machines.

Case Study – 4: Yahoo!Inc

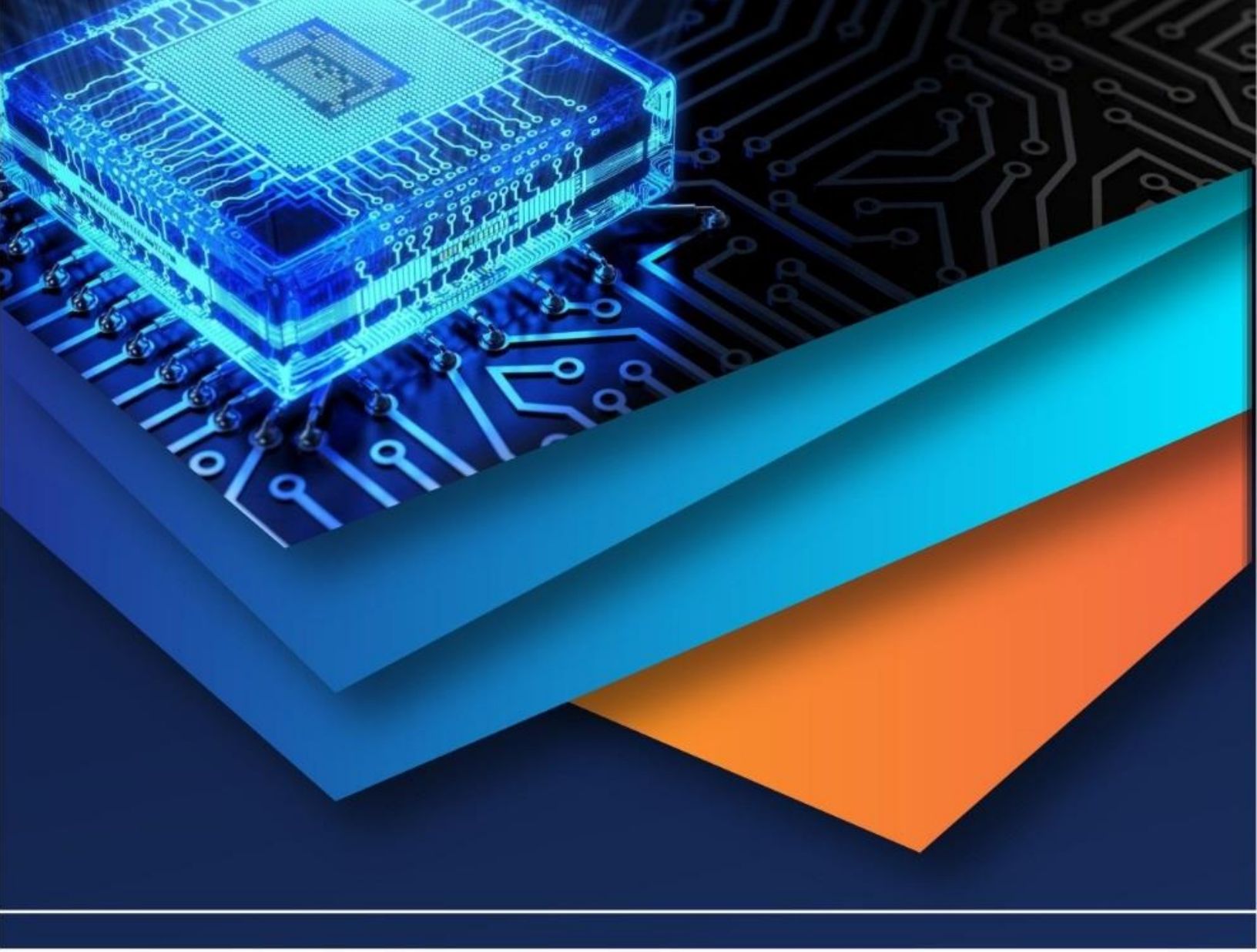
It was Yahoo!Inc. that developed the World's biggest application of Hadoop on February 19, 2008. In fact, if you've heard of 'The Yahoo! Search Webmap', it is a Hadoop application that runs on over 10,000 core Linux cluster and generates data that is now extensively used in each query of Yahoo! Web search.

V. CONCLUSION

The amounts of data is growing exponentially worldwide due to the explosion of social networking sites, search and retrieval engines, media sharing sites, stock trading sites, news sources and so on. Big Data is becoming the new area for scientific data research and for business applications. Big data analysis is becoming indispensable for automatic discovering of intelligence that is involved in the frequently occurring patterns and hidden rules. Big data analysis helps companies to take better decisions, to predict and identify changes and to identify new opportunities. In this paper we discussed about the issues and challenges related to big data mining and also Big Data analysis tools like Map Reduce over Hadoop and HDFS which helps
organizations to better understand their customers and the marketplace and to take better decisions and also helps researchers and scientists to extract useful knowledge out of Big data. In addition to that we introduce some big data mining tools and how to extract a significant knowledge from the Big Data. That will help the research scholars to choose the best mining tool for their work.

REFERENCES

[1] Richa Gupta, Sunny Gupta, Anuradha Singhal, (2014), "Big Data:Overview", IJCTT, 9 (5)
[2] Wei Fan, Albert Bifet, "Mining Big Data: Current Status and Forecast to the Future", SIGKDD Explorations, 14 (2), pp1-5
[3] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, "Data Mining with Big Data"
[4] Puneet Singh Duggal, Sanchita Paul, (2013), "Big Data Analysis:Challenges and Solutions", Int. Conf. on Cloud, Big Data and Trust, RGPV
[5] OnurSavas, YalinSagduyu, Julia Deng, and Jason Li,Tactical Big Data Analytics: Challenges, Use Cases and Solutions, Big Data Analytics Workshop in conjunction with ACM Sigmetrics 2013,June 21.
[6] Apache Hadoop. Available at http://hadoop.apache.org
[7] Apache HDFS. Available at http://hadoop.apache.org/hdfs
[8] Apache Hive. Available at http://hive.apache.org
[9] Apache HBase. Available at http://hbase.apache.org
[10] A community white paper developed by leading researchers across the United States "Challenges and Opportunities with Big Data"
[11] Proceedings of ICETECT 2011 Application of hadoop map reduce technique to VDS.
[12] Apache Hive. http://hadoop.apache.org/hive.
[13] Tom White, (2009) "Hadoop: The Definitive Guide. O'Reilly", Scbastopol, California.
[14] "Why Big Data is a must in E-Commerce", Guest post by Jerry Jao, CEO of Retention Science. http://www.bigdatalandscape.com/news/why-big-data-is-a-must-in-ecommerce



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY