A Literature Survey on Big Data

Monal Chaudhary
Computer Science Engineering
AMC Engineering College
Bangalore

***Abstract* —There are many changes occurring in Cloud Computing, Big Data and Internet of things since past few years. Big Data is becoming main transformation for the enterprises and scientific society. Large and voluminous quantities of data are difficult to be handled by the traditional data analytics. Big data is the data that is very large in volume and also varied in mixture and is moving with great velocity. The major challenge is analyzing Big Data because it includes vast distributed file systems that should be error lenient, supple and accessible. Hadoop, Map Reduce, Apache Hive, No SQL and HPCC are the technologies used by big data application to handle the massive data. This paper provides a broad review of recent developments within the field of big data and its applications. Today, organizations are putting Big Data into practice in such diverse fields such as healthcare, smart cities, energy and finance.**

I. INTRODUCTION

What is big data?
'Big Data' is a term used to illustrate huge set of data which is bulk in volume and with respect to time is increasing exponentially. Using traditional data management tools such massive data are complicated to process.

Most organizations face difficulty to create, operate, and administer vast amount of data while dealing with large number of datasets. Big data is mainly trouble in business analytics since, standard tools and procedures are not designed to search and analyze substantial datasets.
Due to internet and social media penetration vast amount of data produced significantly in the past five years. Everyday's data generation exceeds 2.5 quadrillion bytes of data.

This vast amount of data accumulates from worldwide info, from the sensors that are used to collect climate information to digital pictures and videos. It also includes posts to social media sites, purchase transaction records, and mobile phone GPS signals. Data is growing with an immense speed than ever before and for every human being on the planet about 1.7 megabytes of new data is being created every second by the year 2020. New data is created every single second.

For example, On google alone every second, if a user performs +40,000 search queries, that will make it 3.5 searches per day and approximately 1.2 trillion number of searches per year.

It is estimated that by the year 2019, big Data will drive $48.6 billion in annual spending and by the year 2020, data production rate will be 44 times larger than it was in the year 2009. More than 70% of the digital universe is created by the individuals. But 80% of big data is stored and managed by enterprises.
It is estimated that hourly collection only by Walmart from its customer transactions is greater than 2.5 terabytes of data. A terabyte is equal to one quadrillion bytes. It is estimated that 1/3rd of all data will be stored, or will have passed via cloud by the year 2020, and a total 45 zettabytes value of data will be created.

When we speak about Big Data, as we have done above, we often identify it as a jargon, which means the enormous volume of data – both structured and unstructured that contains so many huge datasets and the traditional database management techniques and associated software techniques cannot process this large amount of data.
Big Data is a concept and a concept can have various interpretations.

Examples of big data:

- About one terabyte of latest business data/day is generated by the New York Stock Exchange.
- According to the statistics, every day more than 500+ terabytes of newly generated data gets absorbed into the databases of the social media sites like Facebook. This huge data is created in conditions of uploading images and videos, exchanging messages, adding of comments etc.
- In 30minutes of a total flight time a single jet plane engine can create 10+terabytes of data. Thus, creation of data exceeds up to many number of Petabytes.

CHARACTERISTICS OF BIG DATA:

In the year 2001, Gartner's Doug Laney first stated the "three Vs of big data" that described few of the characteristics that make big data unique from new data processing:

- VOLUME: Volume refers to the enormous quantities of data generated every second. Traditional database technology faces difficulty to store and analyze mainly the data sets that is too huge in volume. Most crucial role in determining worth out of data is the role of size of data. A specific data can actually be considered as a big data or not depends on the volume of data. Therefore, while dealing with 'Big Data' we consider **'Volume'** as one of the characteristic.

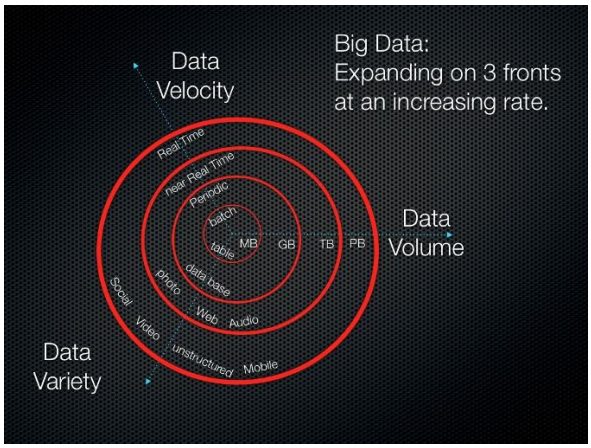

Fig a: Characteristics of Big Data

- VELOCITY: The rate in which the data is formed, accumulated, analyzed and visualized is velocity. Earlier, it was sane to receive an update from the database, every night/even weekly whilst batch processing was common practice. To process the data and modernize the databases large amounts of time was required by computers and servers. In big data period, data is generated in genuine-time or near real-time. With the help of Internet connected devices the wireless or wired, computers and devices can pass on their data the moment they were generated.
  The speed at which data flows in from sources like application logs, networks, business processes and Mobile devices, social media sites, sensors, etc., all deals with the velocity of Big Data. The surge of data is massive and coherent.

- VARIETY: Variety appeals to the various type of data we can use now. Earlier days the focus was only on structured data which properly fitted into tabular columns or relational databases, such as financial data. It is estimated that almost 80% of the data in the world is unstructured (i.e., images, voice, text, video, etc.,). Now we can analyze and combine together various data types like social media conversations, messages, sensor data, video, photos, or voice recordings with the help of big data technology.
  Earlier, the entire data correctly fitted in rows and columns since the data that was created was structured data, but those days are no more. Currently, the data that is created in an organization is 90% of unstructured data. Today, data comes in various variable formats like semi-structured, structured, unstructured data and also complex structured data. These vast range of data requires a variable or unique approach and also diverse methods to accumulate all raw data.

II. TECHNOLOGIES AND METHODS

A. HADOOP:

In earlier days, an enterprise will accumulate and process big data in a computer. Within this approach, the application manages the role of data storage and analysis and the user interacts with this application. This method works well for the applications that are capable of processing less voluminous data. The best application of this method when it processes less huge data that can be accommodated by standard database servers, or up to the limit of the processor that is processing the data. But while dealing with huge amounts of the data it becomes a hectic task to process such amount of data through a single database bottleneck. This problem was solved by Google by introducing an algorithm called Map Reduce. In Map Reduce algorithm the given data is alienated into smaller parts, these parts are then assigned to different computers, and the result is collected from them and are later integrated, to compute the resultant dataset. Doug Cutting with his team developed an open source project named HADOOP which is based on the solution provided by Google.

Hadoop applications also fucntions using this Map Reduce algorithm, in which the data is prepared concurrently with others. In short, Hadoop is useful for developing applications that can perform absolute statistical analysis on vast quantities of data.

Using simple programming models Hadoop allow distributed processing of vast number of datasets athwart clusters of computers. It is written in java and is an Apache open source framework. The environment in which Hadoop applications operates provides computation across clusters of computers and distributed storage. To evolve from a single server to thousands of machines, Hadoop technology is planned and each of this machine provides local calculation and storage.

Hadoop Architecture

Hadoop has 2 major layers at its core, namely:

- Processing or Computation layer (Map Reduce)
- Storage layer (Hadoop Distributed File System-HDFS).

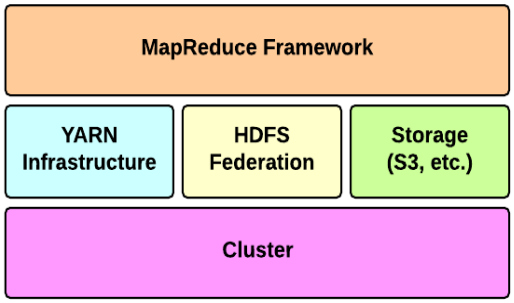

Fig b: Hadoop Architecture

Map Reduce

For writing distributed applications, Map Reduce is used as a parallel programming model which is devised at Google for well-organized dealing out of vast amount of data (multiterabyte data-sets), on bulk collection (thousands of nodes) of product hardware in a dependable, flawless manner. The Map Reduce program functions on Hadoop also that is an apache open source framework.

HDFS Architecture

The underlying file system of a Hadoop cluster is Hadoop Distributed File System (HDFS). This system provides scalable, fault-tolerant storage intended to be deployed on export hardware. HDFS is set apart from other distributed file systems because of several attributes in it.. Among them, some of the key differentiators are that HDFS is:

- designed with hardware failure in mind
- built for large datasets, with a default block size of 128 MB
- optimized for sequential operations
- rack-aware
- cross-platform and supports heterogeneous clusters

In Hadoop cluster, data is broken down into minor units (called blocks) and distributed all through the cluster. Each block is duplicated twice (for a total of three copies) and the two replica is stored in a rack somewhere else in the cluster on two nodes. Since the data has a default replication factor of three, it is highly available and fault-tolerant. HDFS can automatically re-replicate a copy elsewhere in the cluster, if it is lost (for example, because of machine failure). Thereby, ensuring that the threefold replication factor is maintained.

YARN

YARN (Yet Another Resource Negotiator) is the shell that is responsible for resources for application execution.
YARN consists of three core components:

- ResourceManager (one per cluster)
- ApplicationMaster (one per application)
- NodeManagers (one per node)

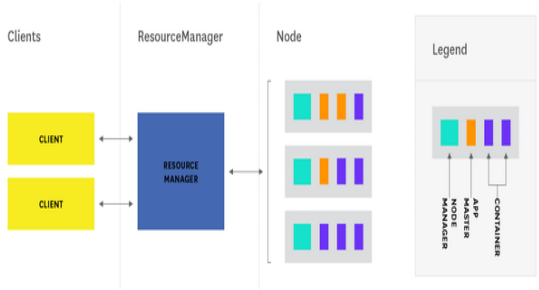

Fig c: Components of YARN

Resource Manager:

It is responsible for taking inventory of available resources and runs several critical services, the most important of which is the Scheduler.

Node Managers:

The Node Manager is a per-node agent tasked with overseeing containers throughout their lifecycles, monitoring container resource usage, and periodically communicating with the Resource Manager.

WORKING OPERATION OF HADOOP?

We can bind collectively numerous commodity computers with a single CPU as an alternative to build bigger servers with heavy configurations that handle large scale processing. Practically, to offer much advanced throughput these clustered machines can interpret the dataset at the same time. Furthermore, it is cheaper than one high end server. Thus hadoop runs across clustered and low-cost machines hence, this is the first important reason behind using Hadoop.
Hadoop runs its code transversely in a group of computers. The above process includes the subsequent core tasks that is performed by Hadoop:

- Initially data is divided into directories and files. These files are further alienated into equivalent size of blocks of 128M and 64M.
- For further processing these files are then distributed across a range of cluster nodes.
- Being on peak of the local file system, HDFS oversees the processing.
- For handling hardware failure, blocks are duplicated.
- Verifying that the code was executed effectively.
- The sort that takes place between the map and reduce stages is performed.
- The sorted data is sent to a assured computer.
- The debugging log for each job is written.

B. MAP REDUCE FRAMEWORK:

A map reduce job generally divides the input data set into autonomous chunks in a totally simultaneous way which are processed by the map tasks. The **frame _work** sorts the outputs of the maps and then goes to the reduce tasks. Typically both the input and the output of the job are accumulated in a file-system. The basic unit of information used in Map Reduce is a Key and value pair. Before feeding the data to Map Reduce model entire type of structured and unstructured data needs to be translated to this basic unit. As the name suggests, Map Reduce model consist of two separate routines, namely Map-function and Reduce-function. When coupled with HDFS, map Reduce can be used to handle big data.

The computation on a set of pairs in Map Reduce model occurs in three stages:

a. Map stage
b. Shuffle stage
c. Reduce stage.

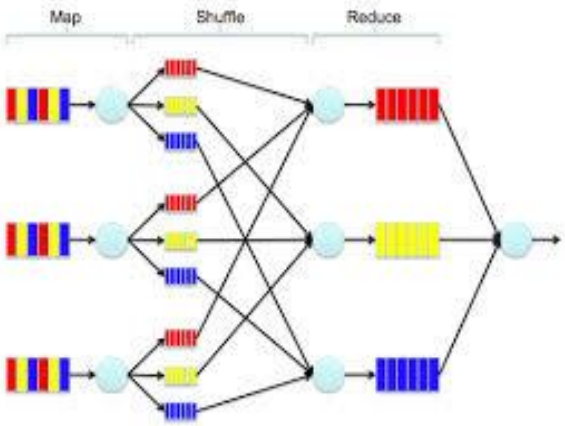

Fig d: Stages of Map reduce

a. Map Stage:

Map Reduce logic, is not restricted to just structured datasets. It has an extensive capability to handle unstructured data as well. Map stage is the crucial step which makes this possible. Mapper brings a structure to unstructured data. For example, incase I want to calculate the number of photographs on my system based on the location (city), of the photograph I will have to analyze the same. The mapper makes key and value pairs from this data set. In this case, key will be the location and value will be the photograph. We have a structure to the entire data-set, after mapper is done with its task. In the map stage, the mapper accepts a single key and value pair as input which generates multiple number of key and value pairs as output. A map function is designed by the user that corresponds an input key and value pair to many number of output pairs for the map phase. Mostly all the time, to identify the preferred position of the input value by altering its key the map phase is simply used.

b. Shuffle Stage:

The Map Reduce framework automatically handles the shuffle stage. The fundamental system executing map reduce directs all of the values that are associated with an entity key to the same reducer.

c. Reduce Stage:

In the reduce stage, all of the values linked with an single key k and outputs multiple number of key and value pairs are accepted by the reducer. This bring out one of the main aspects of the map reduce computation- before reduce stage begins all of the maps must finish. The reducer can perform sequential computations on the values since, all access to the values with the same key is inbuilt in reducer. In the reduce step, by observing that reducers operating on different keys can be executed at the same time the parallelism is exploited.

C. HIVE:

Hive is a data warehousing infrastructure built above hadoop. Providing query, data summarization, and analysis is the primary liability. Analysis of large datasets stored in Hadoop's HDFS is supported by it. An SQL-like interface is provided by Hive to query data that is stored in various databases and file systems that integrate with Hadoop.

Analysis of large datasets stored in Hadoop's HDFS and compatible file systems such as Amazon S3 filesystem is supported by Apache Hive. It provides called Hive query langauge, an SQL-like language with schema on read and clearly transforms queries to map or reduce. Embedded metadata is stored in Apache Derby database, and other client-server databases like MySQL can alternatively be used.

Alternative features of Hive include:

- To give acceleration indexing types like compaction and bitmap index as of 0.10 and extra index types are designed.
- Various types of storages such as plain text, HBase , ORC, RC File and others.
- The time to perform semantic checks during query execution are considerably reduced using metadata storage in an RDBMS.
- Condensed data is stored in the Hadoop ecosystem can be operated using algorithms such as DEFLATE, SNAPPY, BWT, etc.
- Management of strings, dates, and other data-mining tools are done using built-in user defined functions. To handle use cases not supported by built in functions hive supports extending the user defined function set.
- Implicitly, SQL-like queries such as Hive query language are transformed into map reduce or Spark or Tez jobs.

Hive query language do not severely chase all the SQL-92 standard although hive is based on SQL. Extensions of hive query language offers not in SQL, including *multi table* inserts and create table as select, but only provides basic support for indexes. Hive query language only has inadequate subquery support and hence lacks support for transactions and materialized views.

III. OBSTACLES IN BIG DATA IMPLEMENTATIONS

Heterogeneity and timeliness, security, incompleteness and scalability of the data are the biggest obstacles in analyzing big data.

Skilled people are required for the shift to Big Data. It requires people in the area of system analysis, domain knowledge, data analytics, database management and software developers. Large number of open source technologies available in the market for Big Data.

At every step of extraction and analysis the big data creates many challenges. In order to achieve success, we need to overcome many obstacles in the Big Data & Analytics

process. First time the obstacles are encountered they take an wide quantity of time to be solved.

Many are trying to recognize the benefits of big data, while more than 1/5th of the respondents are still trying to gain knowledge more about big data. There are significant reliant of data management professionals trying to understand the basics, although the industry has written innumerable blogs and articles, white papers about big data.

Cost implications is one of the vital practical challenges faced by big data. Even though it's been a decade since the implementation of Big Data analytics has started, the cost allegation of storing vast amount of data is still a subject of concern.

a. Data Collection:

The companies not only have to find and examine the appropriate data they need they must also find it quickly due to today's hypercompetitive business environment. The most important problem in the data collection step is heterogeneity of the data sources. Collecting the data from multiple sources is the first step of general big data schema. Challenges arise when the data sources are complex and sophisticated. The main source of data for Big Data stream is rapidly shifting from manual data entries to the data collected from sensors, social networks.

b. Integration:

The data that has been transferred must be stored in some form. Every day we create so much data that it costs companies fortune to store it in order for them to improve their business. The demand for storing the big data has increased so immensely and in such a fast pace.

The increasing amount of data and a need to analyze the given data in a timely manner for multiple purposes has created a serious barrier in the big data analysis process.
Storing such a huge sized data requires enormous amount of energy and resources. One of the problems of the Big Data is to find the best located servers to store the data. The server locations must also be energy efficient and scalable. The location is important due to the speed of transfer of the stored data to do the analyses.

c. Analysis:

The problems that data analysts face when dealing with an average size of data set emerge in more severe form for the big data. Most business decisions need to be made in a punctual manner. The companies that cannot modify their behavior to the changes in the market behavior in a timely manner have serious problems and will likely face severe problems in the future.

False and transitory correlation will most likely result in wrong decisions that can damage the company in the long term.

There are many huge analytical challenges that come along with big data.

Large number of advance skills are required to perform every type of analysis on enormous quantity of data. That can be unstructured, semi structured or structured.

d. Privacy and security:

The most important test in big data comprises responsive, conceptual, technical and also legal importance.

The personal records of any individual when combined with exterior huge number of data sets, it leads to the interpretation of latest facts about that individual. There can also be possibility that these kinds of details about the individual are private. This individual may not want the data owner or any other individual to know about that individual.

In order to add importance to the business of the organization, information concerning the individuals is collected and used. This is prepared by creating perceptions in their lives which they are ignorant of.

Big data predictive analysis advantages will be utilized by a well-educated person. On the further side underprivileged will be effortlessly recognized and treated inferior. In the future, this would be another very important arising outcome.

IV. APPLICATIONS

Big Data is gradually turning popular. Each area in marketing, healthcare, automation, manufacturing industries can now implement big data analytics.

1. Improving Security and Law Enforcement:

In improving security and enabling law enforcement, big data is applied greatly. In U.S, The National Security Agency (NSA) makes use of big data analytics to outwit terrorist strategies. Cyberattacks are detected and prevented using big data techniques, even police services use big data tools to catch criminals and still predict criminal activities. Big data techniques are also used by credit card companies to spot fake dealings.

2. Healthcare Sector:

Inspite, of the fact that the healthcare sector have access to vast quantities of data. The health care sector has been overwhelmed by failures in utilizing the data to restrain the amount of rising healthcare and by unproductive systems that throttle faster and better healthcare benefits athwart the board.
It is majorly because of the aspect that the electronic data is out of stock, scarce, or unfeasible. Furthermore, all the databases that seize healthcare-interrelated information has made those information complex to link data that can demonstrate patterns helpful in the medical field.
Few hospitals like Beth Israel makes use of data composed from a phone application, as of millions of patients, to permit doctors to make use of fullproof based medicines as conflicting to administering many medical or lab tests to all patients that go to the hospital. A series of tests can be resourceful, but they

can also be costly and frequently turns out to be unproductive.

The University of Florida uses free public health data and Google maps, to build visual data that will allow for quicker identification and proficient analysis of healthcare information that can be used in tracking the increase of chronic disease.

3. Improving Sports Performance:

Big data analytics have now been embraced by nearly all privileged sports. The IBM Slam Tracker tool is used for tennis tournaments, also we make use of video analytics that helps in tracking the quality performance of each player in games like football or baseball game. Also, to get feedback (via smart phones and clo/ud servers) on our game and how to improve it, sensor technology in sports equipments such as basket balls or golf clubs is used. Smart technology is used to track nutrition and sleep, as well as social media conversations to monitor emotional wellbeing. Numerous privileged sports teams also track athletes outside of the sporting environment using these smart technologies.

The NFL has developed personal platform of applications to aid all the thirty two teams in making the finest decisions, based on everything from the weather, to the situation of the grass on the field, to statistics of each and every player's performance in university. This is entirely in name of tactics and also as plummeting player injuries.

4. Civilizing and Optimizing City and Countries:

Various aspects of different cities and countries can be progressed by using big data. Big data can help cities to optimize traffic flows based on actual time traffic information and also social media and weather data. Many cities at present are monitoring big data analytics with the intention of turning themselves into Smart Cities, in which the transport infrastructure and convenience processes all are tied up.

Smart water meters is being used by the city of Long Beach, California to identify illegitimate watering in actual time and is being used to aid some homeowners cut their water usage by as much as 80%. The above becomes crucial
whilst the state is going through the most terrible drought in recorded history ever and governor has enacted the first state wide water limitations.
In Los Angeles, to manage traffic lights and to also control the congestion of traffic about the city makes use of data from magnetic road sensors and also traffic cameras. Approximately, forty thousand traffic signals around the city is controlled by computerized system which has reduced traffic jamming by 16%.

5. Financial Trading:

The area where big data finds a lot of use today is HFT (High Frequency Trading). Algorithms of big data are used to formulate trading decisions. Nowadays, the greater part of impartiality trading currently takes place by means of data algorithms that rapidly take in account signals from the social media networks and also news websites to create, buy and sell decisions in few seconds.

The markets can be scanned by computers which are programmed with intricate algorithms for a set of customizable circumstances and explore for trading opportunities. Depending on the needs and desires of the client all these programs can be designed to function with zero human interaction / with human interaction,.

The most sophisticated of these programs rather than being hardcoded are also designed to modify as markets transform.

V. FUTURE WORK

Big data has taken the business world by storm, we can concur that, but what comes after that?
Large data will it continue to grow?
What kinds of technologies will expand around it?
Will the big data turn into an artifact as rapidly as the next trend, cognitive technology? Fast data? — appears on the horizon.

These are various number of the predictions from the foremost experts in this field and also how much likely they are to approach to exceed.

1. It is estimated that sixty percent and above of companies are currently investing in big data and analytics tools to help make their HR departments more data driven.
2. The market of big data is estimated to develop further by 25% a year. This numeral comes from market research specialists International Data Corporation. IDC predicts that by the year 2018, there will be a compound annual growth rate of 26.4 % to $41.5 billion of the big data technology and services market. Which is about 6 times the growth rate of the whole information technology market.
3. More than 30 billion devices will be wirelessly connected by 2020.
4. There will be continuity in growth of data volumes. Absolutely, there are no questions that we will continue creating superior and superior volumes of data. Particularly taking into account, the exponential growth of total number of handheld devices and internet connected devices.
5. Data analyzing ways will improve. Spark is rising as a balancing tool for analysis while SQL is still the standard and according to ovum, it will prolong to grow.
6. In the business analytics software the prescriptive analytics will be built in. According to IDC by the year 2020, ½ of all business analytics software will include the intelligence where it is desired.

7. Together with autonomous vehicles, robots, smart advisers and virtual personal assistants the *Autonomous agents and things* will continue to be a vast trend.
8. According to IDC(International Data Corporation),
   to incorporate architects and experts in data management big data recruitment shortages will enlarge from analysts and scientists.

VI. CONCLUSION

In this literature survey, we have discussed Big Data in detail until its existing state. It elaborates on the concepts of big data followed by the technologies and methodology used in this field. In this paper we have also discussed regarding the obstacles faced by it and also about the applications. Lastly we have discussed topic related to future facilities which can be harnessed in big data. Big data is a progressing field, and alot of research is however to be done. Currently, Hadoop is the software that handles the big data. Neverthless, the invariable growth in volumes of data is making Hadoop inadequate. In future, extensive research needs to be carried out and also revolutionary technologies need to be developed to tackle the potential of big data completely.

REFERENCES

[1] Samiddha Mukherjee, Ravi Shaw,"Big Data – Concepts, Applications, Challenges and Future Scope", International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 2, February 2016.
[2] Ms. VibhavariChavan, Prof. Rajesh. N. Phursule,"Survey Paper On Big Data",VibhavariChavan et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 2014, 7932-7939.
[3] SabithaM.S, Dr.S.Vijayalakshmi, R.M.RathikaaSre,"Big Data – Literature Survey",alue: 13.98 ISSN: 2321-9653 International Journal for Research in Applied Science & Engineering Technology .
[4] Chun‐Wei Tsai, Chin‐Feng Lai, Han‐ChiehChaoand Athanasios V. Vasilakos,"Big data analytics: a survey", Springer.
[5] Sreedhar C, Dr. D. Kavitha, K. Asha Rani," Big Data and Hodoop", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 5, May 2014 2.
[6] Ms. VibhavariChavan, Prof. Rajesh. N. Phursule, JSPM's Imperial College of Engineering and Research, Pune, "(IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 2014, 7932-7939.