ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 6, June 2015**

# Investigative Research on Big Data: An Analysis

Shyam J. Dhoble[1], Prof. Nitin Shelke[2]

ME Scholar, Department of CSE, Raisoni college of Engineering and Management Amravati, Maharashtra, India.[1]

Assistant Professor, Department of CSE, Raisoni College of Engineering and Management, Amravati, Maharashtra, India.[2]

**ABSTRACT:** This paper is the analysis paper which gives the details of various analyses done by various companies on Big Data. It has become big news and there are no signs that interest is decreasing. Over the last few years, companies around the world have awakened to a new direction: Big Data brings with it Big Responsibility. Big data is a broad term for data sets so large or complex that traditional data processing applications are skimpy. Challenges include analysis, capture, data creation, search, sharing, storage, transfer, perception, and information privacy. Big data has extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions. Big Data is thus very important to increase productivity growth in the entire world since it is affecting not only software-intensive industries but also public domains like education, health field, education and administrative sectors. Big data "size" is a constantly moving on the target of enlargement of the data as of 2012 ranging from a few dozen terabytes to many petabyte of data. It is defined as the amount of data just beyond technology's capability to store, manage and process efficiently. The ideal of process of huge datasets has been changed from centralized architecture to distributed architecture. In this paper, we provide an extensive analysis of Big data analytics research, while highlighting the specific concerns in Big data world.

**KEYWORDS**: Big data, data perception, data analytics, hadoop.

## I. INTRODUCTION

This paper is about large collections of data. In that we have graph database, big data band.

Fig 1. Perception created by IBM AT multiple terabyte in size, the text and image of Wikipedia are an example of big data.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 6, June 2015**

"Big Data" is a the term use of techniques to capture, process, analyses and visualize potentially large datasets in a reasonable timeframe not accessible to standard IT technologies. By extension, the platform, tools and software used for this purpose are collectively called "Big Data technologies".[2] **Big data** is a broad term for data sets so large or complex that traditional data processing applications are skimpy. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, perception, and information privacy. The term is often refers simply to be use as predictive analytics or the other certain advanced type methods to extract value from data, and refers to a particular size of data set. The Accuracy of big data may lead to the more confident decision making to the system. And off course the better decisions are mean for the greater operational efficiency, cost reductions and reduced risk. Analysis of data sets can find new correlations, to "the spot business trends, and combat crime and so on." Scientists, practitioners of media and advertising and governments are regularly meeting the difficulties with large number data sets in the area including social Internet search, finance and business information. Scientific encounter limitation in E-Science work, including metrology, genomics, connection, complex physics simulations and environmental research.[3]

Data sets grow in size in part because they are increasingly being gathered by cheap and numerous information-sensing mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers, and wireless sensor networks. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012, every day 2.5 Exabyte's ($2.5\times10^{18}$) of data were created; The challenge for large enterprises is determining who should own big data initiatives that straddle the entire organization. Work with big data is necessarily uncommon; most analysis is of "PC size" data, on a desktop PC or notebook that can handle the available data set. Relational database management systems and desktop statistics and visualization packages often have difficulty handling big data. The work instead requires "massively parallel software running on tens, hundreds, or even thousands of servers".[12] What is considered "big data" varies depending on the capabilities of the users and their tools, and expanding capabilities make Big Data a moving target.

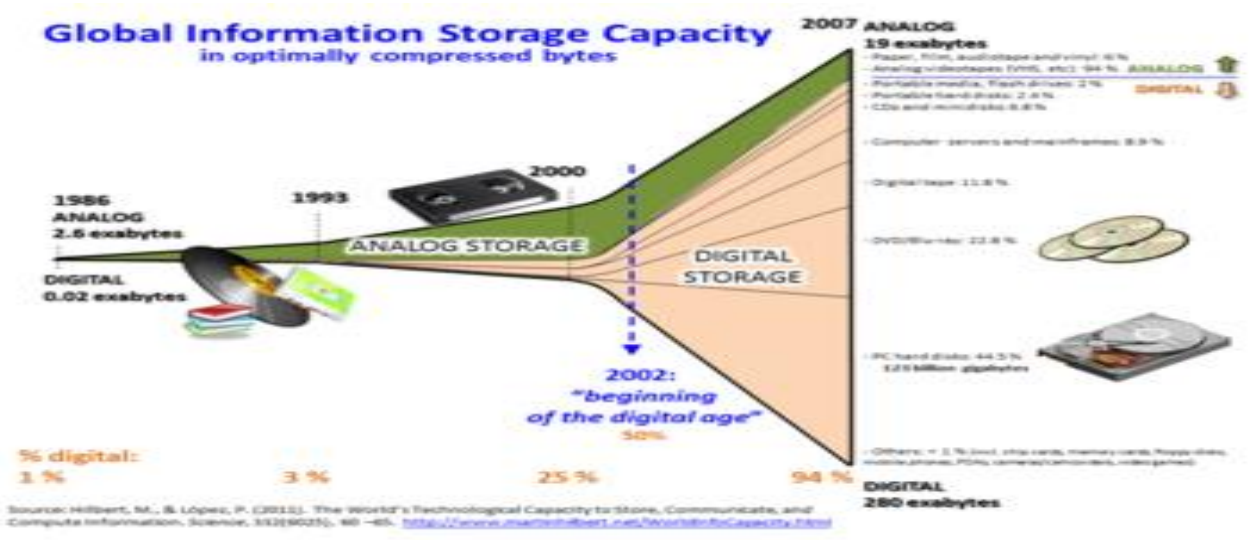


Fig 2. Growth of the Digitization of Global Information Storage Capacity

Relational database management systems and desktop statistics and perception packages often have difficulty handling big data. The work instead requires "massively parallel software running on tens, hundreds, or even thousands of servers". What is considered "big data" varies depending on the capabilities of the users and their tools, and expanding capabilities make Big Data a moving target. Thus, what is considered to be "Big" in one year will become ordinary in later years. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration."[6]

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 6, June 2015**

## II. RELATED WORK

The Applications required more resources that are available on an inexpensive machine. Many companies are finding problem themselves with the business processes applications that no longer fit to a single cost effective computer. A simple but an expensive solution to buy the specialty machines application that has a lot of memory storage and a bunch of CPUs in one system. This solution means supported by the fastest machines available in the market , and usually the only limiting application in the budget. An alternative solution is to build a high-availability cluster. This can be called as a single machine, and typically requires very specialized installation and administrative servicess. Many high-availability clusters are proprietary and well expensive. The more economical solution for acquiring the necessary computational resources which support the cloud computing technology. A same type of data have bulk data that needs to be transformed, where the processing of each data item is essentially independent on the respective data items; that is, using a single-task and multiple-data algorithm. Hadoop provides an open source framework for cloud computing technology, as well as a distributed file system. The Hadoop supports the Map Reduce model, which was introduced by Google as a method of solving a class of petascale problems with large clusters of inexpensive machines. The model is based on two distinct steps for an application.

• **Map**: An initial ingestion and transformation step, in which individual input records can be processed in parallel.
• **Reduce**: an aggregation or summarization step, in which all associated records, must be processed together by a single entity.
The core concept of MapReduceing in Hadoop is that input may be split into logical field, and each field may be initially processed independently, by a map task process. The results of this individual processing field can be physically partitioned into distinct sets, which are then sorted. Each sorted field is passed to a reduce task. Figure.3. illustrates how the Map Reduce model works.

A map task may run on any compute node in the cluster, and multiple map tasks may be running in parallel across the

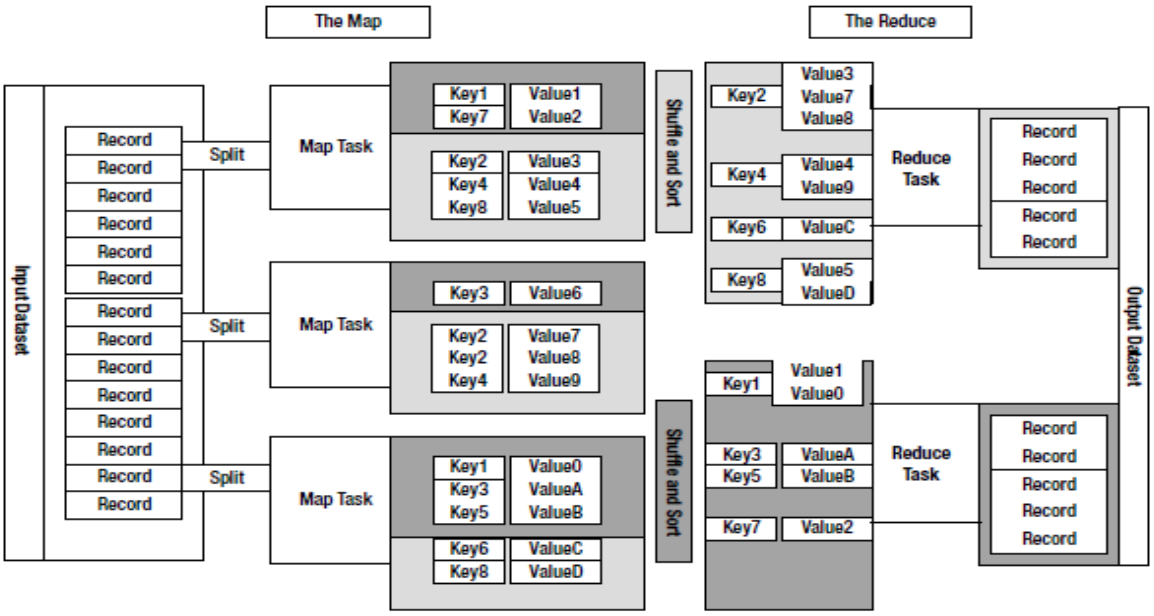


Fig. 3 the map reducing techniques.

cluster. The map task is responsible for transforming the input records into key/value pairs. The output of all of the maps will be partitioned in different fields, and each Fig partition will be sorted. There will be one partition for each

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 6, June 2015**

reduce task. Each partition's sorted keys and the values associated with the keys are then processed by the reduce task. There may be multiple reduce tasks running in parallel on the cluster. The application developer needs to provide **only four items** to the Hadoop framework: the class that will read the input records and transform them into one key/value pair per record, a map method, a reduce method, and a class that will transform the key/value pairs that the reduce method outputs into output records. The Hadoop Map Reduce framework requires a shared file system. This shared file system does not need to be a system-level file system, as long as there is a distributed file system plug-in available to the framework. When HDFS is used as the shared file system, Hadoop is able to take advantage of knowledge about which node hosts a physical copy of input data, and will attempt to schedule the task that is to read that data, to run on that machine. The Hadoop Distributed File System (HDFS) Map Reduce environment provides the user with a sophisticated framework to manage the execution of map and reduce tasks across a cluster of machines.

The user is required to tell the framework the following:-

• Location(s) in the distributed file system of the job input.
• Location(s) in the distributed file system for the job output.
• Input format.
• Output format.
• Class containing the map function.
• Optionally:- the class containing the reduce function.
• JAR file(s) containing the map and reduces functions and any support classes.

If a job does not need a reduce function, the user does not need to specify a reducer class, and a reduce phase of the job will not be run. The framework will partition the input, and schedule and execute map tasks across the cluster. If requested, it will sort the results of the map task and execute the reduce task(s) with the map output. The final output will be moved to the output directory, and the job status will be reported to the user.

The Map Reduce is oriented around key/value pairs. The framework will convert each record of input into a key/value pair, and each pair will be input to the map function once. The map output is a set of key/value pairs are nominally one pair that is the transformed input pair, but it is perfectly acceptable to the output of multiple pairs. The map output pairs are been grouped into one and then sorted by unique key. The reduce function is called one time for each key, in sort sequence, with the key and the set of values that share that key. The reduce method may output an arbitrary number of key/value pairs, which are written to the output files in the job output directory. If the reduce output keys are unchanged from the reduce input keys, the final output will be sorted.

The framework provides two processes that handle the management of Map Reduce jobs :-
• Task Tracker manages the execution of individual map and reduces tasks on a compute Node in the cluster.
• Job Tracker accepts job submissions, provides job monitoring and control, and manages the distribution of tasks to the Task Tracker nodes.

Generally, there is one Job Tracker process per cluster and one or more Task Tracker processes per node in the cluster. The Job Tracker is a single point of failure, and the Job Tracker will work around the failure of individual Task Tracker processes.

## III. CHARACTERISTICS OF BIG DATA

Big data can be described by the following characteristics:

- **Volume** – The amount of the data that is generated which is very important in this context. It is the memory allocation size of the data which determines the value and potential of the data under utilization and whether it

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 6, June 2015**

can actually be considered Big Data or not. The name 'Big Data' itself contains a term which is related to size allocation.

- **Variety** - The next aspect of Big Data is its types of variety. This means that the type of category to which Big Data belongs to it, also a very essential factor that needs to be understood by the data analysts. This helps the people, who are closely working on that data and also associated with it, to effective use the data to their advantages and the upholding of importance to its Big Data.
- **Velocity** - The 'velocity' in the context refers to the speed of generation of the data or how fast the data is been generated and to be processed to meet the demands, queries and the challenges which occurs in the way of growth and development process.
- **Variability** - This is a factor is the most important factor which can became the problem for those who analyses the data. This refers to the inconsistency and incompatibility to which can be shown by the data at times, thus execution process of being to be able to handle and manage the data effective way.
- **Veracity** - The quality of the data should be created and captured greatly and high definition format.. Accuracy of analysis depends on the veracity of the source data collection.
- **Complexity** - Data management process can become a very chard to implement process, especially when large amount of the data come from multiple sources databases. These data need to be formed in the format of linked to which it can be connected and correlated in order to be able to extract the information that is supposed to be distributed by these data. This factor, is therefore, termed as the 'complexity' of Big Data.

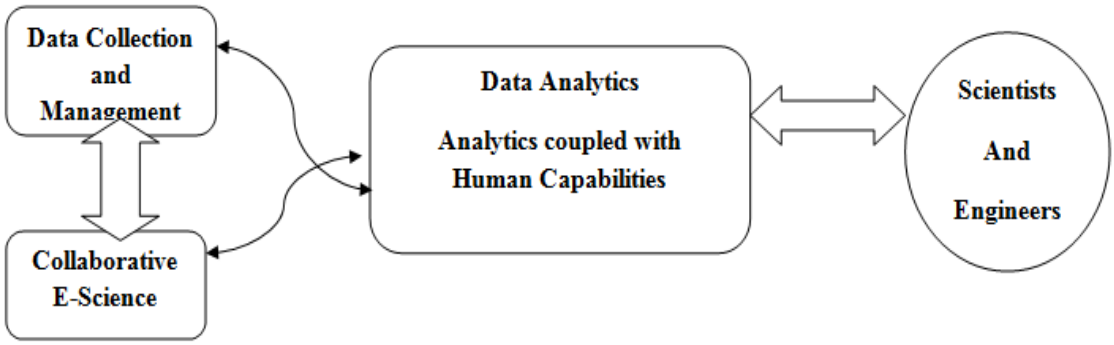


Fig. 3: For regulatory science, one of the most important aspects of big data research is Data Analytics, which is a key for moving from data to human insights.

**Factory work and Cyber-physical systems may have a 6C system:**

1. Connection (sensor and networks),
2. Cloud (computing and data on demand),
3. Cyber (model and memory),
4. content/context (meaning and correlation),
5. community (sharing and collaboration), and
6. Customization (personalization and value).

In this concept and in order to provide useful insight to the structured management and gain correct data content, and that data has to be processed with advanced tools (analytics and algorithms) to generate correct information. By considering the presence of visible and invisible issues in an industrial area, for the information generation of the algorithm has to be capable of finding out the addressing invisible issues such as machine degradation, component wear, etc. in the structure of factory floor.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 6, June 2015**

## IV. SYSTEM ARCHITECTURE

As we know that companies are uses the value of business intelligence. Business data is check for many purposes: the log analytics and social media analytics for risk evaluation by the company, customer possession, and brand management, and so on. Typically, such variation tasks are handled by different systems, even if each system includes same steps of information extract, the data wiping, relational-like processing, statistical and predictive modeling, and to explore and visualization tools as shown below.

Big Data use of separate systems in this fashion which becomes expensive for the given large size of the data sets. Similar declarative Specifications are required at higher levels to meet the programmability and composition needs of these analysis pipelines. Besides the basic technical need, there is a strong business imperative as well. Businesses typically will outsource Big Data processing, or many aspects of it. Since the point of the out-sourcing is to specify precisely what task will be performed without going into details of how to do it. Declarative specification is needed not just for the pipeline composition, but also for the individual operations themselves. Each operation (cleaning, extraction, modeling etc.) potentially runs on a very large data set. Furthermore, each operation itself is sufficiently complex that there are many choices and optimizations possible in how it is implemented. In databases, there is considerable work on optimizing individual operations, such as joins. It is well-known that there can be multiple orders of magnitude difference in the cost of two different ways to execute the same query. Fortunately, the user does not have to make this choice – the database system makes it for her. In the case of Big Data, these optimizations may be more complex because not all operations will be I/O intensive as in databases. Some operations may be, but others may be CPU intensive, or a mix. So standard database optimization techniques cannot directly be used. However, it should be possible to develop new techniques for Big Data operations inspired by database techniques.

The very fact that Big Data analysis typically involves multiple phases highlights a challenge that arises routinely in practice: production systems must run complex analytic pipelines, or workflows, at routine intervals, e.g., hourly or daily. New data must be incrementally accounted for, taking into account the results of prior analysis and pre-existing data. And of course, provenance must be preserved, and must include the phases in the analytic pipeline. Current systems offer little to no support for such Big Data pipelines, and this is in itself a challenging objective.

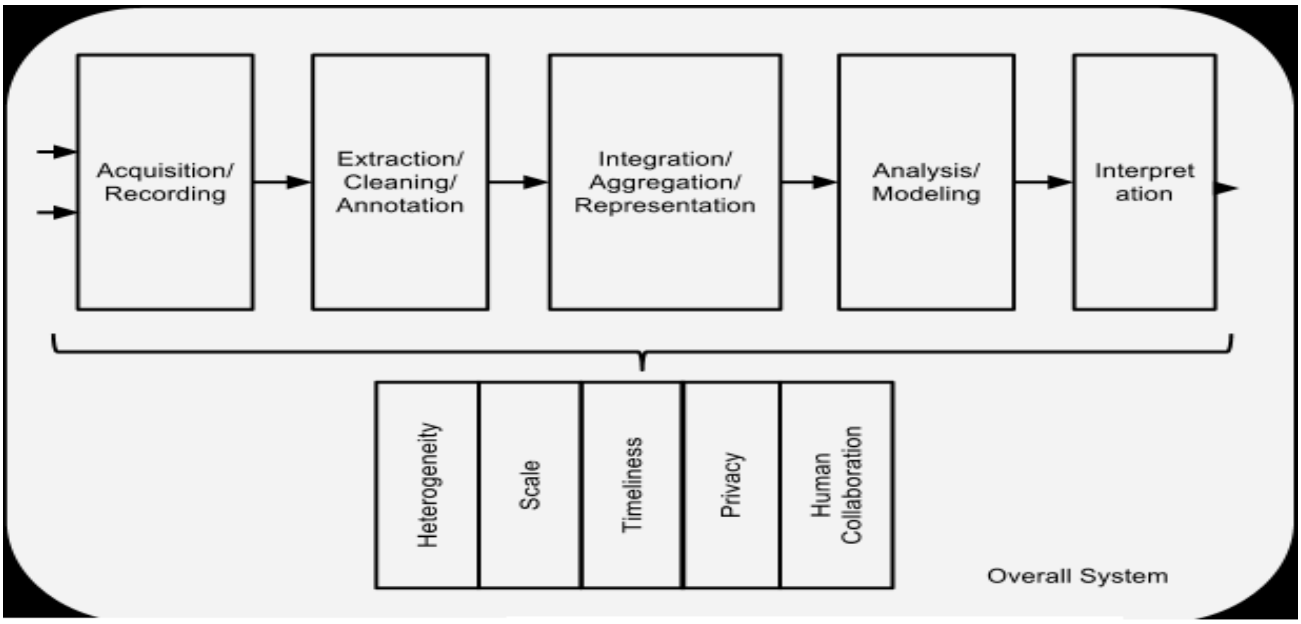


Fig. 4: The big data analysis pipeline.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 6, June 2015**

## V. IT SCENE IN BIG DATA

Big data projects increases data growth, as 60 percent of those analyses predict that these increased growth in unstructured information, presenting in IT with further data management has a difficult task for. Managing the flow of unstructured information is a problem for organizations as (31%) of respondents have seen that challenge in response to their organizational approach. Respondents cite emails (52%) and customer databases (49%) as the current most common big data sources. Analytics solutions top the list of technologies for managing and extracting value from business data, cited by more than half (52%) of those analyses. Thirty-six percent of those analysised plans to invest in data perception technology over the next 12 months, while nearly as many plan to invest in predictive analytic products (33%). Meanwhile, a quarter of respondents are investing in big data training services.

Investments in production technologies such as storage (55%), servers (53%) and Networking products (45%) remain strong. However, the companies are investing in Hadoop based big data preservation which face further complications. 56% say that they will have to re-architect the data center network to some extent.

## VI. CONCLUSION AND FUTURE WORK

As we have entered in an era of Big Data, there is the potential for making profits in many scientific disciplines and enterprises through better analysis of the large volumes of data that is becoming available. However, many technical challenges like data Perception and implementation is to be taken into consideration in future. This is just the analysis paper which shows the demand of big data and how IT are taking interest in it. We must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data. As the Hadoop has extends into new markets and sees some new use cases with security and flexible difficult tasks, the advantages of processing sensitive and legally protected information with all Hadoop projects and HBase must be join with protection for the private information that limits performance impact.

## REFERENCES

[1] LaValle et al: "Big Data, Analytics and the Path From Insights to Value", (Dec 2010)
[2] McKinsey Global Institute, Big Data: "The next frontier for innovation, competition and productivity" (June 2011).
[3] Advancing Personalized Education. Computing Community Consortium. Spring 2011,jan1.
[4] "Getting Beneath the Veil of Effective Schools": Evidence from New York City. Will Dobbie, Roland G. Fryer, Jr. NBER Working Paper No. 17632. Issued Dec. 2011.
[5] "Smart Health Wellbeing": Computing Community Consortium. Spring 2011.
[6] "Big data: The next frontier for innovation, competition, and productivity". James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. McKinsey Global Institute. 15 May 2011.
[7] "A Sustainable Future": Computing Community Consortium. Summer 2011.
[8] "Using Data for Systemic Financial Risk Management" :Mark Flood, H V Jagadish, Albert Kyle, Frank Olken, and Louiqa Raschid. Proc. Fifth Biennial Conf. Innovative Data Systems Research, Jan. 2011.
[9] "The Emerging Big returns on big data" :TCS-Big-Data-Global-Trend-Study-2013,mar 21,2013
[10] "Challenges and Opportunities with Big Data" :Divyakant Agrawal, Philip Bernstein, Elisa Bertino ,Susan Davidson, Umeshwas Dayal, 1-1-2011
[11] IDG ENTERPRISE RESEARCH REPORTS, jan 6,2014 ,http://www.idgenterprise.com/report/big-data-2
[12] Grand Challenge: "Applying Regulatory Science and Big Data to Improve Medical Device Innovation" : Arthur G. Erdman∗, Daniel F. Keefe, Senior Member, IEEE, and Randall Schiestl, IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, VOL. 60, NO. 3, MARCH 2013.
[13] "Big Data A new World of Opportunities" :Nessi White Paper , December 2012.
[14] "Big Data Process Analytics: A Analysis Emerging Research in Management &Technology" ISSN: 2278-9359 (Volume-3, Issue-7) Sameera Siddiqui Deepa Gupta.