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 and Hadoop: A Review

Nisha Bhardwaj[1], Dr Balkishan[2], Dr Anubhav Kumar[3]

M. Tech Student, Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, Haryana, India[1]

Assistant Professor, Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, Haryana, India[2]

Head CSE Department, Lingaya's GVKS IMT, Faridabad, Haryana, India[3]

ABSTRACT: We all know that Data stores are growing by 50% each year and that rate of increase is accelerating. For such a huge data we need new modern systems which can handle such data. Now-a-days, Bigdata becomes a dominant thing which has its arms in various fields such as research, science and industry. In this paper, the first part explain the brief introduction to Big data, rest of the part of paper explains the difference between the traditional data and bigdata analytics, bigdata architecture, big data Opportunities, its Issues and Challenges, tools and techniques and brief explanation of Hadoop.

KEYWORDS: Bigdata, Hadoop, Hive, PIG, HDFS, MapReduce, privacy.

I. INTRODUCTION

As we all know that humans create tremendous amount of data everyday and this data comes from every-where. So, for getting the better performances for computation industries have to focus towards the increasing processor speed and memory. Traditional systems were not capable to manage the vast amount of data i.e. bigdata. So, we need some new kind of systems that can handle such bigdata [12].

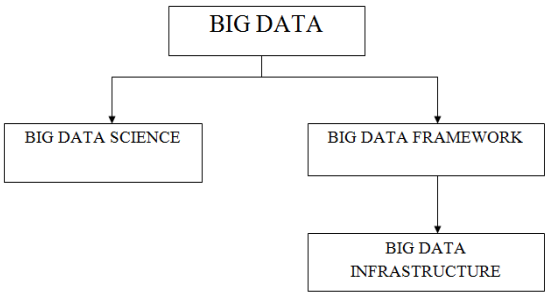

Figure 1: Notations in Big Data environment

Big Data Science is the study of techniques that are the combination of information technology and mathematical approaches which deals with the acquisition, conditioning and evaluation of big data. Big data frameworks are the collection of software libraries with their associated algorithms enabling distributed processing and analysis of big data problems. On the other hand, Big Data Infrastructure refers to the instantiation of one or more big data frameworks including physical and virtual servers, storage structures, networking structures, interfaces and back-up systems.
Bigdata is a combination of transactional data and the interactive data.
When any type of organization wants to deal with huge amount of data i.e. bigdata then they face many problems those are as follows-

- Data acquisition

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

- Information storage and organization
- Information analytics and search
- Data privacy and security

Bigdata technologies is a combinatorial term that includes the tools, techniques and method that can be used to perform manipulate, gather, investigate, process and visualize the large datasets in a reasonable time and memory by increasing the processing speed as compared to the traditional systems.

Bigdata differs in three dimensions from traditional systems which are: volume, velocity and variety [11].

Volume- It represents the large volume and size of data. In the comings days the size of data increased to zettabytes from petabytes.

Velocity- It represents the speed at which data is coming as well as the speed at which the data is outgoing. Traditional systems are not capable to handle such data that is always in motion.

Variety- It represents the inconsistency in the data flow. Bigdata contains the structured and unstructured data. Data Variety means the different varieties of data which includes-

- Structured data – It is organized in a structure. E.g.-SQL data.
- Semi- Structured data – It is the form of structured and unstructured data. E.g.-XML.
- Unstructured data – It includes the data with no identifiable structure. E.g.-Image.

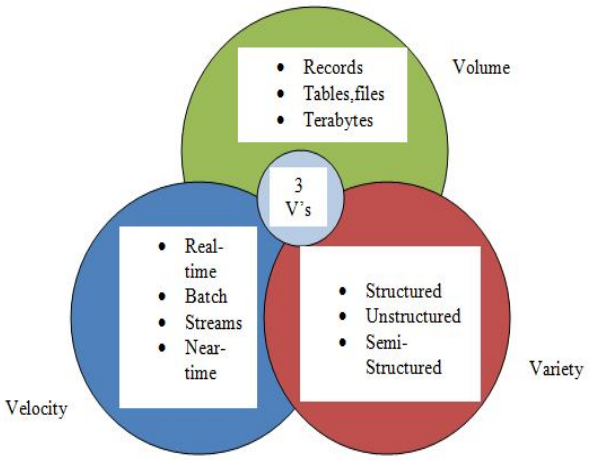

Figure 2: Three V's of Bigdata.

II. TRADITIONAL VS. BIGDATA

i. TRADITIONAL DATA ANALYTICS VERSUS BIGDATA ANALYTICS

We all heard about the term 'Big data analytics' it is used in various decisions of many organizations. Basically, Big data Analytics is nothing but it's a analytic technique which is operated on Big or we can say that large data [9].

'Big Data' is the general term which is used to store any kind of data i.e. structured or unstructured and this type of data is not possible to store in a traditional form or Relational form in the organizations databases. The general characteristics of big data are the following which are given below:

- In the big data, data is stored in the range of peta or exa bytes but the current storage of the various Organizations limits only to Terabytes.
- Generally, big data is considered as unstructured data.
- In this data are generated using unconventional methods.

There is another concept which is very close to the concept of big data analytics. That new concept is the data mining. This concept of data mining has been adopted by various industries to the mountains of data [11].

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

Table 1: Difference between traditional and bigdata analytics

Traditional data analytics	Bigdata analytics
1. Suitable for structured and planned data only.	Suitable for any kind of data-structured, semi-structured and unstructured data.
2. Its technology used in Relational database.	Its technology used in Hadoop framework.
3. Regulated.	3. Unregulated.
4. High cost.	4. Generally little or low cost.
5. Centralized, point in time.	5. Distributed, real-time.
6. No open source.	6. Open source.

III. RISE IN USE OF INTERNET: IT RISE THE BIGDATA

Internet is the global system of interconnected computer networks. Now-a-days we can find an internet connection in various institutions, industries, organizations and among various private parties. Internet has grown day by day due to its vast usage by the people. Traditionally only the government agencies make use of internet to their various internal work but in the current time everyone uses the internet even its usage increases to the millions of users for their different purposes according to their needs. In the coming years, internet change the totally means to communicate and way of Business. Internet has given an international or we can say a universal mean to the world. It becomes the global source of information which is used by the millions of people at various places. Everyday internet keeps on changing. There are mainly two things which are evolved in the internet. The first thing is social media and the second is mobile networking. Social media includes the facebook and linkedIn. There are so many users which are continuously using social Medias by which they create large amount of data. Another way of using internet is the mobile networking. This is one of the easiest ways to use internet because of its high reliability, usability and availability. Facebook continues to be the most popular social media site. Internet is used for many things e-mail, online chatting, file transfer and many other documents [9].

Mostly used service on internet is www.Webpages on the internet can be seen and read by anyone. The speed of the broadband connection is very high so the people adopted themselves to that much high speed. People spend so much time online on internet, watched many videos, performed many activities on social networking sites and become the constant content creators.Finally, due to the rise in the internet by social media, humans automatically raise the bigdata.

IV. BIGDATA ARCHITECTURE

Big data can be stored, acquired, processed and analyzed in many ways. Basically, Big Data architecture is composed of a skill set that is used for creating a reliable, scalable and completely automated data pipelines. Every layer in the stack of the Skill set begins with the cluster design. The below diagram shows the working of the complete skill set and also the working of big data architecture. The most important thing about the data pipelines is that they took the useless data and convert it into the useful data. Big data architecture represents the structured and pattern based approach to simplify the task of defining the big data architecture.

i. INGRESS

When we have spun up our cluster, we have to decide how to load data. For doing this, we have basically two main approaches: batch and event-driven. We can say it as batch ingest. In this we are ingesting the data from structured data sources. E.g.RDBMS. so we can say that batch ingest is only used for the structured data. While the Event- driven

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

ingest allows us to define redundant agents. It means event ingest approach is useful for most real-time events such as log or transactional data.

ii. STAGING

Firstly data is arrived. When this process of arriving is finished then the complete process is handover to the staging only. Basically it not only defines the storage of data but it also gives the explanation to store it in a right way and at the right place.

iii. ACCESS CONTROL

Access control is a kind of restriction to something. Here, the mean to access using, entering or we can say consuming something. Another term is authorization which means to permission of accessing something.
Finally we have to include the Information architecture. It operates on the data in such a way so that various teams can work in sharing mode. It also referred to as multi-tenancy. Because as we all know that it is not possible to maintain data in one directory. If someone wants to read data then also it's not possible to read it from one directory and then make use of that in another directory. So for the clustering we have to make a concise schema that bounded the data for outsider's access. These rules of clustering are defined in the Information architecture.

iv. HADOOP FRAMEWORK

It includes the HDFS and MapReduce. HDFS is responsible for storing the data in a distributed manner across multiple Hadoop cluster nodes. The MapReduce framework provides rich computational API's for developers to code, which eventually run as map and reduce tasks on the Hadoop cluster.

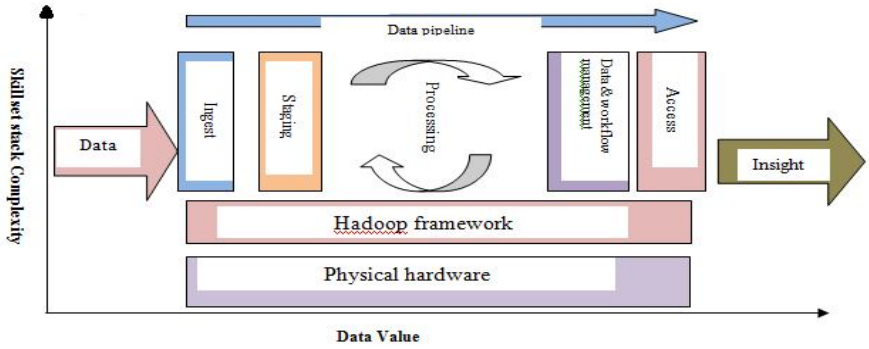

Figure 3: Big Data Architecture

V. OPPORTUNITY FROM BIGDATA

Across the various industries, organizations are using different means for better business solutions by utilizing the useful and essential information. There are many success stories which are entitled with big data that enables innovation and creating new services, also help in creating new solutions for data mining and machine learning techniques [5]. Various organizations experiences in the big data field which have been characterized by the following categories:

- **Innovation –** The innovation is something new or we can say that innovation thing is that which is creative. So it is characterized as we have distributed processing among large datasets, scalability and reliability with the help of cloud computing, virtualization and economic hardware requirement. Various Organizations keep on establishing big data management strategies so that they can scale up the utilization of these innovations, in a cost-effective manner.
- **Acceleration –** Different Organizations have different ways to accelerate their business processes using big data techniques. Various oil and gas companies did processing on the sensor data to operate on maintenance operations and real time operations. Various health related data such as Tele-health, home-health monitoring, electronic health records are operated in a connected health world. Online purchasing and group purchasing help retailers to understand their customer's requirement well so that they engage and act upon their request very frequently. The

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

various imperatives for the enterprise is to understand Enterprise Information Management capabilities and adopt then integrate big data initiatives.

- **Collaboration –** In the big data scenario, Collaboration is adopted as the new trend. In this data assets are shared and referred as a part of the product of data services. In the coming years, there could be new ideas which help in collaborating new opportunities for the big data. There are various data driven companies which offers many services data monetization which means to generate revenue from available data sources or real time streamed data and data democratization which means the acquisition and spread of knowledge amongst the common people.

Table 2: Bigdata value across industries

	Volume of Data	Velocity of Data	Variety of Data	Under - Utilized Data ('Dark Data')	Big Data Value Potential
Banking and Securities	High	High	Low	Medium	High
Communications & Media Services	High	High	High	Medium	High
Education	Very Low	Very Low	Very Low	High	Medium
Government	High	Medium	High	High	High
Healthcare Providers	Medium	High	Medium	Medium	High
Insurance	Medium	Medium	Medium	Medium	Medium
Manufacturing	High	High	High	High	High
Chemicals & Natural Resources	High	High	High	High	Medium
Retail	High	High	High	Low	High
Transportation	Medium	Medium	Medium	High	Medium
Utilities	Medium	Medium	Medium	Medium	Medium

VI. ISSUES AND CHALLENGES

There are various Issues and Challenges which are [7]-

- Heterogeneity (or Variety)
- Scale (or Volume)
- Speed (or Velocity)
- Accuracy and Trust
- Privacy Crisis
- Interactiveness
- Garbage mining

Heterogeneity (or Variety) - The stored data is not all of the same type or category. The different types are given below-

- Structured data - data that is organized in a structure so that it is identifiable e.g. SQL data.
- Semi-structured data - a form of structured data that has a self-describing structure yet does not conform with the formal structure of a relational database e.g. XML.
- Unstructured data - data with no identifiable structure e.g. image.

Scale (or Volume) - The "Big" in Big data and represents the large
Volume or size of the data. At present the data existing is in petabytes and is supposed to increase to zettabytes in the near future
For example big social networking sites are producing data in order of terabytes everyday and this amount of data is difficult to handle using traditional systems.
Speed (or Velocity) - It represents not only the speed at which the data is incoming, but also the speed at which the data is outgoing.
Traditional systems are not capable of performing analytics on data that is constantly in motion.

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

Accuracy, Trust and Provenance - As the data sources are belonging to different regions, not all of them are well known and not all of them are verifiable. Therefore, the accuracy and trust of the source data become an issue.
The vast volume of big data attributes additional characteristics- high dynamics and evolution. So an adequate system is required which allow dynamic changing and evolution of the data items for big data management and analysis. This makes data provenance an integral feature in any system that deals with big data.
Interactiveness - Great interactiveness boosts the acceptance of a complicated mining system and its mining results by potential users. Sufficient interactiveness is especially important for big data mining.
Garbage mining - Garbage has no value. No one wants garbage. Everyone wants to get rid of garbage. In the real world, garbage collection is a business with profits.
Garbage does not speak : " I am garbage, recycle me!"
We propose applying data mining approaches to mine garbage and recycle it. "One man's trash is another's treasure". Garbage definition remains one of the greatest challenges.

VII. TOOLS AND TECHNIQUES

The tools which are typically used in Bigdata are as follows [10]:
NoSQL
DatabasesMongoDB, CouchDB, Cassandra, Redis, BigTable, Hbase, Hypertable, Voldemort, Riak, ZooKeeper
MapReduce
Hadoop, Hive, Pig, Cascading, Cascalog, mrjob, Caffeine, S4, MapR, Acunu, Flume, Kafka, Azkaban, Oozie, Greenplum
Storage
S3, Hadoop Distributed File System
Servers
EC2, Google App Engine, Elastic, Beanstalk, Heroku

Processing
R, Yahoo! Pipes, Mechanical Turk, Solr/Lucene, ElasticSearch, Datameer, BigSheets, Tinkerpop
To handle large amount of data various techniques and technologies have been introduced to manipulate, process, analyze and visualize the bigdata which are discussed below [13]:

NOSQL databases
For storing the large or huge data traditional systems were not effective. So, new breed comes into existence that is called NOSQL databases, which are mainly used to work with bigdata. These are non-relational databases.

MapReduce approach
A programming task which is divided into multiple identical subtasks and which is distributed among multiple machines for processing is called Map task. The results out of these Map tasks are combined together into one or many reduce tasks.
Overall approach of computing tasks is called a MapReduce approach.

Hadoop
It is openly available software based on java programming which processes big datasets in a distributed environment. It is designed in such a way if there any hardware failure occurs then it automatically handled by the software by the framework.
It provides the following parts-
Hadoop Common
Hadoop Distributed File System (HDFS)
Hadoop YARN
Hadoop MapReduce

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

Hive
Hive runs on the top of Apache hadoop and provides data warehouse capabilities. The Apache Hadoop framework is difficult to understand, and it requires a different approach from traditional programming to write MapReduce-based programs. With Hive, developers do not write MapReduce at all. Hive provides a SQL like query language called HiveQL to application developers.

PIG
Another abstraction layer on the top of MapReduce is provided by Apache Pig. It provides a language called Pig Latin. Pig Latin is a programming language that creates MapReduce programs using Pig. It is a high-level language for developers to write high-level software for analyzing data. Pig code generates parallel execution tasks, therefore effectively uses the distributed Hadoop cluster.

HDFS
It is responsible for storing the data in a distributed manner across multiple hadoop cluster nodes.
HDFS, being a distributed file system, has the following major objectives to satisfy to be effective:

- Handling large chunks of data.
- High availability and handling hardware failures seamlessly.
- Streaming access to its data.
- Scalability to perform better with addition of hardware.
- Durability with no loss of data in spite of failures.
- Portability across different types of hardware/software.
- Data partitioning across multiple nodes in a cluster.

VIII. HADOOP : IT'S A SOLUTION FOR BIGDATA PROCESSING

Hadoop is a framework which is freely available and based on java programming which enables the processing of big datasets in a distributed manner. Hadoop is created by Google and provided by apache software foundation. Hadoop delivers distributed processing power at a remarkably low cost, making it an effective complement to a traditional enterprise data infrastructure [13].

i. HADOOP ECOSYSTEM

Apache Hadoop ecosystem enables us to effectively apply the concepts of the MapReduce paradigm at different requirements. It also provides end-to-end solutions to various problems that are faced by us every day. Apache Hadoop ecosystem is vast in nature. It has grown drastically over the time due to different organizations contributing to this open source initiative. Due to the huge ecosystem, it meets the needs of different organizations for high performance analytics. The ecosystem of Apache Hadoop consists of the following [2]-
Hbase: it directly runs on the top of HDFS. It allows developers to directly read or write data from HDFS. It is called as NOSQL database as it not supports SQL.
Pig: Another abstraction layer of apache hadoop is provided by the apache pig. Pig code generates parallel code execution. It provides a language called pig latin.
Hive: it runs on the top of Apache hadoop and provides data warehouse capabilities.
Sqoop: it allows developers to import/export easily from various data sources. It establishes the connectivity between non hadoop data sources and HDFS.
Mahout: Mahout is an open source machine learning software library. Mahout is highly effective over large datasets; the algorithms provided by Mahout are highly optimized to run the MapReduce framework over HDFS.
HCatalog: HCatalog helps any third party software to create, edit, and expose (using rest APIs) the generated metadata or table definitions. HCatalog provides metadata management services on top of Apache Hadoop. It means all the software that runs on Hadoop can effectively use HCatalog to store their schemas in HDFS.
Ambari: Ambari provides a set of tools to monitor Apache Hadoop cluster hiding the complexities of the Hadoop framework. It offers features such as installation wizard, system alerts and metrics, provisioning and management of Hadoop cluster, job performances, and so on.

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

Flume: it is a distributed data collection service which gets their data from different sources then aggregates them and puts into HDFS.
Zookeeper: it main task is to maintain the coordination among various nodes. Except coordination it also did many other jobs such as maintaining configuration information, group services and distributed coordination at very high speed.

ii. RUNNING A MAPREDUCE JOB

In MapReduce programming model, the basic unit of information is the key/ value pair. The MapReduce program reads sets of such key /value pairs as input and gives new key/value pairs as output. The overall approach occurs in three different stages which are map, shuffle and reduce [14].

iii. MAP AND REDUCE TASKS

Map () – It is written by the user. It is a programming task which is divided into multiple identical subtasks and which is distributed among multiple machines for processing is called map tasks. All of its stages are stateless; it enables them to run independently in a distributed environment. All of the map tasks should finish before the start of reduce phase.
Reduce () – It is also written by the user. It can act upon multiple pairs. The result which comes out from the map tasks are combined together into one or more reduce tasks. This allows us to handle lists of values that are too large to fit in memory [14].

iv. THE MAP SIDE

While a map task is running, its map function is emitting key-value records to an in memory buffer class called MapOutputBuffer2. If not explicit specified, all the entities discussed during this section are located inside this class. In the previous section we stated that these records are separated into partitions corresponding to the reducers that they will ultimately be sent to. This is done by a partitioned component, which assigns records output by the map tasks to reducers based on a partitioning function [1].

v. STORING LARGE DATA SET IN HDFS

HDFS is a subproject of Apache foundation. It is designed to maintain the large files or records in a highly distributed manner. It uses master-slave architecture and reliably run on low cost hardware. It provides very high speed data access and also provides APIs to manage its distributed file system. To handle failures of nodes, HDFS effectively uses data replication of file blocks across multiple Hadoop cluster nodes, thereby avoiding any data loss during node failures. HDFS stores its metadata and application data separately [2].

IX. CONCLUSION AND FUTURE WORK

As we know that in the big data era where enormous amounts of heterogeneous, semi-structured and unstructured data are continuously generated at tremendous scale. Big data discloses the limitations of existing data mining techniques, which results in a series of new challenges to the big data mining. In the coming years, this will be the new trend of research. From the big data demand, there will be created more than 4 million jobs. Along with china, in the coming years India could also be one of the biggest suppliers of manpower for the 'Big data' industry.
This paper initiates a research to examining the traditional view of data analytics and big data analytics, introduces new techniques and technologies to handle big data, explain big data issues and challenges. We identified various issues related to big data which includes its storage, management and processing. We also have identified some challenges regarding big data which big data users specifically face. Our future research work emphasis on creating a more complete understanding of the issues and challenges which are associated with big data.

REFERENCES

[1] Jeffrey Dean, Sanjay Ghemawat, "**MapReduce: Simplified Data Processing on Large Clusters"**, pp 1-13.
[2] Hrishikesh Karambelkar, **"Scaling Big Data with Hadoop and Solr"**, Birmingham B3 2PB, UK, 2013.
[3] John Lenhart, **"Big Data: Issues, Challenges, tools and Good practices"**, pp 1 – 15.
[4] Danyel Fisher, Rob DeLine, Mary Czerwinski, Steven Drucker, **"Interactions with Big Data Analytics"**, pp 50-59, May + June 2012.

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

[5] Soumendra Mohanty, Madhu Jagadee, Harsha Srivatsa, **"Big Data Imperatives"**, Apress.
[6] Martin Hahmann, Gunnar Schröder, Phillip Grosse, **"Data Analytics Methods and Techniques"**, Forschungskolleg, pp 1-45
[7] Dunren Che, Mejdl Safran, Zhiyong Peng, **"From Big Data to Big Data Mining: Challenges, Issues and Opportunities"**, pp 1-15.
[8] Neil Raden, Hired Brains, **"Big Data Analytics Architecture"**, Teradata Aster, 2012.
[9] Jose Ramon G. Albert, **"Challenges, Opportunities and Issues on Using Big Data for Meeting Current and Emerging Demands on Measuring Progress and Development"**, United Nations Economic Commission For Europe (ECE) Conference of European statisticians, European Commission Statistical Office of the European Union (EUROSTAT), Organization for Economic Cooperation and Development (OECD) Statistics Directorate, United Nations Economic and Social Commission for Asia and the Pacific (ESCAP) Meeting on the Management of Statistical Information Systems(MSIS 2014) (Dublin, Ireland and Manila, Philippines 2014).
[10] Marko Grobelnik, **"Big Data Tutorial"**, Stavanger 2012.
[11] Vibha Shukla, Pawan Kumar Dubey, **"Big Data: Moving Forward with Emerging Technology and Challenges"**, International Journal of Advance Research in Computer Science and Management Studies, volume 2 Issue, pp 187-193, 2014.
[12] Nidhi Grover, **"Big Data'- Architecture, Issues, Opportunities and Challenges"**, International Journal of Computer and Electronics Research (IJCER) volume 3 Issue, 2014.
[13] Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar, **"A Review paper on Bigdata and Hadoop"** ,International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014.
[14] Shilpa, Manjit kaur, **" BIG Data and Methodology-A review"** , International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 10, October 2013.