International Journal of Engineering, Science and Mathematics
Vol. 6 Issue 8, December 2017 (Special Issue)
ISSN: 2320-0294 Impact Factor: 6.765
Journal Homepage: http://www.ijesm.co.in, Email: ijesmj@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

A Review Of Big Data And Its Current Research Directions

Sangram Keshari Swain*
Srinivas Prasad**
Manas Ranjan Senapati***

Abstract

In recent years, advances in Web technology and the proliferation of sensors and mobile devices connected to the Internet have resulted in the generation of immense data sets available on the Web that need to be processed and stored. At present scenario, companies are starting to realize the importance of using more data in order to support the decision for their strategies. Big Data has become a business priority for companies in the globally integrated economy. Big data is a collection of massive and complex data sets that include the huge quantities of data, social media analytics, data management capabilities, real-time data. Big Data analytics is the process of examining large amounts of data. Technology trends for Big Data embrace open source software, commodity servers, and massively parallel distributed processing platforms. There are enormous opportunities for research in the Big Data field.

Keywords:

Big Data
Parameters
Evolution
Hadoop
HDFS

Author correspondence:
First Author,
Associate Professor, Department of Computer Science & Engineering,
School of Engineering & Technology, Bhubaneswar Campus
Centurion University of Technology and Management, Odisha, India
Second Author,
Professor, Department of Computer Science & Engineering,
K L University, Andhra Pradesh, India.
Third Author,
Associate Professor, Department of Information Technology,
Veer Surendra Sai University of Technology, Burla, Odisha, India

1. Introduction:

We are in the midst of a "Big Data" revolution. Innovations in technology and greater affordability of digital devices have presided over today's Age of Big Data, an umbrella term for the explosion in the quantity and diversity of high-frequency digital data. Big Data is the new experience curve in the new economy driven by data with high volume, velocity, variety, and veracity. They come from various sources that include the Internet, mobile devices, social media, geospatial devices, sensors, and other machine-generated data. Unlocking the value of Big Data allows businesses to better sense and respond to the environment, and is becoming a key to creating competitive advantages in a complex and rapidly changing market. Traditional data processing and analysis of structured data using RDBMS and data warehousing, no longer satisfy the challenges of Big Data. Technology trends for Big Data embrace open source software, commodity servers, and massively parallel-distributed processing platforms [1].

Nowadays companies are starting to realize the importance of using more data in order to support the decision for their strategies. Companies started to realize that they can choose to invest more in processing larger sets of data rather than investing in expensive algorithms [2]. Many believe that “big data” will transform business, government, and other aspects of the economy.

Big Data not only provides solutions to longstanding business challenges, but also helps in finding new ways to transform processes, organizations, entire industries and even society itself. Big Data analysis drives nearly every aspect of the modern society, including mobile services, retail, manufacturing, financial services, life sciences, and physical sciences. This article emphasizes on the meaning, techniques & technologies, trends, and challenges of Big Data [3].

Big data is comprised of datasets too large to be handled by traditional database systems. Big data include structured data, semi-structured and unstructured are those data formatted for use in a database management system. Semi-structured and unstructured data include all types of unformatted data including multimedia and social media content [4].

At present, enterprises have urgent needs to conduct an effective and stable statistical analysis on big data. Big Data is Latest technology, which is very complex to identify the large data sets due to its big size and complexity. It is very difficult to manage in current technology and the era. A collection of data sets are very large and complex that becomes difficult to be processed using on-hand database management tools or traditional data processing applications are called Big Data.

2. Big Data Concept:

Big Data is the word used to describe massive volumes of structured and unstructured data that are so large that it is very difficult to process this data using traditional databases and software technologies. This chapter begins by defining what is truly new and different about big data.

2.1 The sources of big data

The sources and formats of data continue to grow in variety and complexity. A partial list of sources includes the public web; social media; mobile applications; federal, state and local records and databases; commercial databases that aggregate individual data from a spectrum of commercial transactions and public records; geospatial data; surveys; and traditional offline documents scanned by optical character recognition into electronic form. The advent of the more Internet-enabled devices and sensors expands the capacity to collect data from physical entities, including sensors and radio frequency identification (RFID) chips. Personal location data can come from GPS chips, cell-tower triangulation of mobile devices, mapping of wireless networks, and in-person payments.

2.2 The “7 Vs” & Complexity:

The seven terms and complexity that signify Big Data have the following properties:

Volume: Refers to the quantity of data gathered by a company. This data must be used further to obtain important knowledge. Every day 2.3 trillion gigabytes of data are produced. Facebook users upload 100 terabytes of data daily. Wal-Mart alone processes over a million transactions per hour.

Variety: Refers to the type of data that Big Data can comprise. This data can be structured as well as unstructured from natural language, geospatial, sensor events: extraction of meaning from such diversity requires increasing algorithmic and computational power.

Velocity: Refers to the time in which Big Data can be processed. Some activities are very important and need immediate responses that are why fast processing maximizes efficiency. 90% of today's data has been created in last two years. 200 million e-mails, 300,000 thousand tweets, and 100 hours of YouTube videos are passing by every minute of the day. Real-time processing will reduce the huge storage requirements.

Veracity: Data streams from various sources have varying signal –to-noise ratios and by the time they reach analysis stage they may have so much of accumulated errors that are difficult to sort out requiring data clean up.

Variability: refers to data whose meaning is constantly changing. (Ex. Natural language processing)-New meanings are created and old are discarded. Interpreting connotations are essential to gauging and responding to social media. This presents a unique decoding challenge.

Visualization: Visual representation of the data and findings that contain dozens of parameters is certainly a challenge.

Value: Big data offers value (in terms of cost reductions to organizations and customers, more effective methods of selling, etc.) to those who can deal with scale and extract the knowledge inherent in the data.

Complexity: Complexity of the data also needs to be considered when the data is coming from multiple sources. The data must be linked, matched, cleansed and transformed into required formats before actual processing.

2.3 Hadoop

Hadoop, which is a free, Java-based programming framework, supports the processing of large sets of data in a distributed computing environment. It is a part of the Apache project sponsored by the Apache Software Foundation. Hadoop cluster uses a Master/Slave structure [5]. Using Hadoop, large data sets can be processed across a cluster of servers and applications can be run on systems with thousands of nodes involving thousands of terabytes. Distributed file system in Hadoop helps in rapid data transfer rates and allows the system to continue its normal operation even in the case of some node failures. This approach lowers the risk of an entire system failure, even in the case of a significant number of node failures.

Hadoop enables a computing solution that is scalable, cost effective, flexible and fault tolerant. Hadoop Framework is used by popular companies like Google, Yahoo, Amazon and IBM etc. to support their applications involving huge amounts of data. Hadoop has two main subprojects – Map Reduce and Hadoop Distributed File System (HDFS).

2.3.1 Map Reduce

Hadoop Map Reduce is a framework used to write applications that process large amounts of data in parallel on clusters of commodity hardware resources in a reliable, fault-tolerant manner [6]. A Map-Reduce job first divides the data into individual chunks which are processed by Map jobs in parallel. The outputs of the maps sorted by the framework are then inputted to the reduce tasks. Generally, the input and the output of the job are both stored in a file-system. Scheduling, Monitoring, and re-executing failed tasks are taken care by the framework.

2.3.2 Hadoop Distributed File System (HDFS)

HDFS is a file system that spans all the nodes in a Hadoop cluster for data storage. It links together file systems on local nodes to make it into one large file system. HDFS improves reliability by replicating data across multiple sources to overcome node failures.

3. Need for Big Data Analysis:

In this era of Globalization, the vast amount of information is available that was neither available nor accessible in the past. Earlier surveys, focus groups, etc. were some of the sources of information; however, now they can be examined directly. Organization scan discovers more through larger samples and crude information. This potential can be realized only if you bring together and analyze all that data. At present, we have gigantic information for analysts to manage, which increases the chances of missing opportunities or risks. Organizations that augment their human experts with Big Data technologies could have competitive advantages by heading off problems sooner, identifying opportunities earlier, and performing a mass customization at a larger scale.

4. Challenges of Big Data:

Like all major technological innovations, Big Data has its own bricks and motors. Following are the challenges faced by the business leaders, business organizations and the IT organizations in adopting Big Data.

4.1. Big Data A Challenge For Business Leaders

• Businesses, consumers, and suppliers are creating and consuming gigantic amounts of information in the global marketplace.

• Gartner predicts that enterprise data in all forms will grow 650 percent over the next five years.
• According to IDC, the world's volume of data doubles every 18 months. This flood of data often referred to as "information overload," "data deluge" and "Big Data," clearly creates a challenge for business leaders.

4.2. Big Data A Challenge For Business Organizations

• Most organizations are either getting on a technique related to Big Data or intend to do so in the near future, yet almost no organizations have an expressed strategy for it.
• Big Data initiatives widen an organization's IT setup in new ways and can strain dealings with business units.
• Big Data initiatives originate within business units, putting added pressure on the organization's IT department to get adequately prepared to support them.

4.3. Big Data A Challenge For IT Organizations

• IT organizations should understand that Big Data is not only big volumes. Big Data projects might not succeed if attention is not paid to the variety, velocity, complexity, as well as volume.
• IT departments should be well prepared for things like budget changes and infrastructure re-engineering, as priorities will shift to Big Data processing and analysis.
• IT organizations should be on view that apart from some successful Big Data implementations in well-known organizations, most Big Data projects in recent past are best-described as exploratory.
• IT leaders must make sure that appropriate budget and training should be provided as Big Data requires experienced and skilled IT personnel.

5. Big data advantages:

In Big data, the software packages provide a rich set of tools and options where an individual could map the entire data landscape across the company, thus allowing the individual to analyze the threats he/she faces internally. This is considered as one of the main advantages as big data keeps the data safe. With this, an individual can be able to detect the potentially sensitive information that is not protected in an appropriate manner and makes sure it is stored according to the regulatory requirements. There are some common characteristics of big data, such as

a) Big data, integrate both structured and unstructured data.
b) Addresses speed and scalability, mobility and security, flexibility and stability.
c) In big data, the realization time to information is critical to extracting value from various data sources, including mobile devices, radio frequency identification, the web and a growing list of automated sensory technologies.

All the organizations and business would benefit from the speed, capacity, and scalability of cloud storage. Moreover, end users can visualize the data and companies can find new business opportunities. Another notable advantage with big data is, data analytics, which allows the individual to personalize the content or look and feel of the website in real time so that it suits the each customer entering the website. If big data are combined with predictive analytics, it produces a challenge for many industries. The combination results in the exploration of these four areas:

a) Calculate the risks on large portfolios
b) Detect, prevent, and re-audit financial fraud
c) Improve delinquent collections
d) Execute high-value marketing campaigns

6. Need of security in big data:

For marketing and research, many of the businesses uses big data, but may not have the fundamental assets particularly from a security perspective. If a security breach occurs to big data, it would result in even more serious legal repercussions and reputational damage than at present.

In this new era, many companies are using the technology to store and analyze petabytes of data about their company, business, and their customers. As a result, information classification becomes even more critical. For making big data secure, techniques such as encryption, logging, honeypot detection must be necessary. In many organizations, the deployment of big data for fraud detection is very attractive and useful. The challenge of detecting and preventing advanced threats and malicious intruders must be solved using big data-style analysis. These techniques help in detecting the threats in the early stages using more sophisticated pattern analysis and analyzing multiple data sources.

Not only security but also data privacy challenges existing industries and federal organizations. With the increase in the use of big data in business, many companies are wrestling with privacy issues. Data privacy is a liability, thus companies must be on privacy defensive. But unlike security, privacy should be considered as an asset; therefore it becomes a selling point for both customers and other stakeholders. There should be a balance between data privacy and national security.

7. Big Data Processing Technology:

Theoretically, there are no limits to the improvement of the processing function of big data. At present, in the practical application of the processing of big data, the most often used and mainstream realization technology is the MapReduce technology, parallel database technology, and the hybrid structure technology based on MapReduce technology and parallel database technology.

7.1 MapReduce Technology

In 2004, MapReduce was proposed by Google, it is an object-oriented programming model to deal with the large data, primarily used for processing internet data, such as document capture, inverted index construction. But because MapReduce has a simple and powerful data processing interface, and it hides many details on massively parallel execution, fault tolerance, and load balance implementation, so the technology has been widely applied in the field of machine learning, data mining, data analysis, text tokenization, indexing research, the creation of other kinds of data structures (e.g., graphs).

In a slide presentation, Google offers the following applications of MapReduce: distributed grep, distributed sort, web link-graph reversal, term-vector per host, web access log stats, inverted index construction, document clustering, machine learning, and statistical machine translation. The MapReduce technology realizes the abstract processing of complicated business logics involved in the parallel programming. It realizes complicated the computing process, provides a simple and easily used interactive interface and conceals the specific realization process for the parallel computing, processing, fault tolerance, data analysis and load balancing used in complicated businesses.

The MapReduce technology includes two basic operation conceptions: Map (Mapping) and Reduce (Simplification). The Map technology mainly processes a group of input data record and distributes data to several servers and operating systems. Its means of processing data is a strategy based on the key/value. The Reduce technology mainly occupies itself in summarizing and processing the result after processing the above key/value. Issues and tasks in the real world may be modeled and described by means of this simple means of processing. Programs realized by this means will be distributed cluster.

The data processing algorithm based on issues will be distributed to the distributed system formed by ordinary computers and then executed. The system will then solve the problem on the details in relation to the input of big data. Then the algorithm programs crossing computer cluster by means of the center server will be dispatched and managed: the management not only relates to the condition of each processing machine but to interactive communication request between the computers. Using such computing mode can help realize the mode of processing and computing of big data based on the distributed system structure of large-scale enterprises, without the need to grasp the process and details of the parallel processing. It will easily cause the realization of unified dispatch, management, storage, analysis and processing of the scattered resource information of the integrated enterprise. It will realize the high-efficiency analysis and utilization of big data of enterprises by using the business data information of each branch of the enterprise.

MapReduce is designed for mass composed of a low-end computer cluster; its excellent scalability has been fully verified in the industry. MapReduce has a low requirement for hardware, MapReduce allows building cluster using inexpensive hardware. As a free, open source system, MapReduce can store data in any format; can achieve a variety of complex data processing function. Analysis based on the MapReduce platform, without the need of complex data preprocessing and writing in the database process, can be directly analyzed based on the flat file, and the calculation model which its use is mobile computing instead of moving data, therefore the analysis delay can be minimized. But the utility software based on MapReduce is relatively little, many data analysis functions requires users to develop their own, which will lead to increased cost. Because the MapReduce does not want to become a database system, so it does not provide SQL interface.

7.2 Parallel Database Technology

During the present phase, the popularizing and applying of relational databases feature the widest scope and they are at a mainstream position in the whole database system field. The original design object of the relational database is to realize the application of the large-scale machines based on the “Host Computer – Terminal Computer” mode; however, its application scope is very limited. With the popularity and application of the “Client - Service”, the relational database system brings about an application era of “Client - Service” and is widely developed and applied. However, with the popularizing of the Internet technologies, the Internet information resources begin increasingly complicated, and the relational database begins to become unable to apply to complicated Internet application and cannot be used to express and administer each type of complicated document type and multi-media resource information. Therefore, the relational database system is improved and adjusted in this regard, such as adding the support function to the database system that is object oriented, at the same time, adding the function on handling each complicated information data.

Database processing technology based on parallel computing is a technology blended with the parallel computing mode and database processing technology. It originated in the seventies of the 20th century, mainly studies the parallelism of the relational algebra operations and the hardware design for implementation of relation operation, hopes to realize some function of the relational database operation by hardware. Unfortunately, the study failed. In the late 80s of the 20th century, the research direction of parallel database technology turned gradually to the general parallel machine, and the research was focused on the physical organization of the parallel database, operative algorithms, optimization, and scheduling policy. From the 90s until now, with the development of the basic techniques for processor technology, storage technology, network technology, the parallel database technology rise to a new level, the focus of the research is also transferred to the time and spatial.

8. Big data challenge:

With the advancement of data, the challenges are also increased to gain the knowledge from the data. Challenges describe how to deal with the data. Security and privacy are the main challenges deals with the data storage for the authentication of the data. The main challenges are related to how to store the data what will be the physical storage medium to store the data from where will be easily accessible and analysis becomes easy and proper understanding. Data have a number of challenges that are related to its complexity which are:

1. How to Understand the Unstructured Data.
2. Capturing of important data.
3. How to store, analyze and understand the data.
4. The most important challenge is Privacy and security.

9. Issues with Big Data:

Security has always been an issue when data privacy is considered. Data integrity is one of the primary components when preservation of data is considered. Access and sharing of Data which is not meant for the public have to be protected. For this type of security, many researchers have been done. Security has always been an issue when data are considered. In the paper, A Metadata-Based Storage Model for Securing Data in Cloud Environment defined the metadata-based approach to secure the large data. They provide the architecture to store the data. Uses cloud computing to make the data unavailable to the intruder. Data integrity is one of the primary components when preservation/security is considered. Hash functions were primarily used for preserving the integrity of the data. The drawback of using hash function is that a single hash can only identify the integrity of the single data string. And because of this drawback, it becomes impossible to locate the exact position within the string where the change has been occurring.

The solution to overcoming the above problem is to split the data string into the block and then protect each block by the hash function. This also created a drawback that in the case of large data set storing such large number of hashes imposes significant space overhead. In paper Hashing Scheme for Space-efficient Detection and Localization of Changes in Large Data Sets, a method to overcome this problem was described. Certain properties like logarithmic were added instead of linear increase [7]. Whereas the work explained by paper Big Data Privacy Issues in Public Social Media, the very idea of privacy to the people who are using social media was explained. The three techniques to get the location information to stay away from such harmful flood of information were explained [8].

10. Tools to manipulate Big Data:

For the purpose of processing a large amount of data, the big data require exceptional technologies. The various techniques and technologies have been introduced for manipulating, analyzing, and visualizing the big data [9]. There are many solutions to handle the Big Data, but the Hadoop is one of the most widely used technologies.

A. Hadoop

Hadoop is a Programming framework used to support the processing of large data sets in a distributed computing environment. Hadoop was developed by Google's MapReduce that is a software framework where an application breaks down into various parts. The Current Apache Hadoop ecosystem consists of the Hadoop Kernel, MapReduce, HDFS and numbers of various components like Apache Hive, Base and Zookeeper [10]. MapReduce is a programming framework for distributed computing, which is created by the Google in which divide and conquer method is used to break the large complex data into smaller units and process them. MapReduce has two stages which are [11]:

· Map (): - The master node takes the input, divide into smaller subparts and distribute into worker nodes. A worker node further does this again that leads to the multi-level tree structure. The worker node processes the m=smaller problem and passes the answer back to the Master Node.

· Reduce (): - The, Master node collects the answers from all the sub-problems and combines them together to form the output

B. HDFS

HDFS is a block-structured distributed file system that holds a large amount of Big Data. In the HDFS the data is stored in blocks that are known as chunks. HDFS is client-server architecture comprises of NameNode and many DataNodes. The name node stores the metadata for the NameNode. NameNodes keeps track of the state of the DataNodes. NameNode is also responsible for the file system operations, etc [12]. When Name Node fails the Hadoop doesn't support automatic recovery, but the configuration of secondary node is possible. HDFS is based on the principle of "Moving Computation is Cheaper than Moving Data". Other Components of Hadoop [13]:

□ **HBase**: it is open source, Non-relational, distributed database system is written in Java. It runs in the top of HDFS. It can serve as the input and output for the MapReduce.

□ **Pig**: Pig is a high-level platform where the MapReduce programs are created which is used by Hadoop. It is a high-level data processing system where the data sets are analyzed that occurs in high-level language.

□ **Hive**: it is Data warehousing application that provides the SQL interface and a relational model. Hive infrastructure is built on the top of Hadoop that help in providing summarization, query, and analysis.

□ **Sqoop**: Sqoop is a command-line interface platform that is used for transferring data between relational databases and Hadoop.

□ **Avro**: it is a data serialization system and data exchange service. It is basically used in Apache Hadoop. These services can be used together as well as independently.

□ **Oozie**: Oozie is a Java based web-application that runs in a Java servlet. Oozie uses the database to store the definition of the workflow that is a collection of actions. It manages the Hadoop jobs.

□ **Chukwa**: Chukwa is a data collection and analysis framework which is used to process and analyze the large amount logs. It is built on the top of the HDFS and MapReduce framework.

□ **Flume**: it is a high level architecture which focused on streaming of data from multiple sources.

□ **Zookeeper**: it is a centralized service that provides distributed synchronization and providing group services and maintains the configuration information etc.

C. HPCC

HPCC is an open source computing platform and provide the services for the management of big data workflow. The HPCC' data model is defined by the user. HPCC system is designed to manage the most complex and data-intensive analytical problems. HPCC system is a single platform, a single architecture and a single programming language used for the data processing. HPCC system is based on Enterprise control language that is declarative, on-procedural programming language HPCC system was built to analyze the large volume data for the purpose of solving a complex problem. The main components of HPCC are:

- HPCC data refinery: massively parallel ETL engine.
- HPCC data delivery: Massively structured query engine
- Enterprise Control Language distributes the workload between the nodes

Databases	NoSQL, MongoDB, 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

11. Some important current research areas in Big Data field:

✓ **Improving Data analytic techniques** - Gather all data, filter them out on certain constraints and use them to take confident decisions.

✓ **Natural Language processing methods** - Use NL-processing techniques on Big Data to find out the current sentimental trend and it can be used in business, politics, finance etc.

✓ **Big Data tools and deployment platforms** - Conventional tools are inefficient to handle Big Data, Lots of research is needed in these fields.

✓ **Better data mining techniques** - Data mining is the method to grab data from various platforms. Improved distributed crawling techniques and algorithms are needed to scrape data from multiple platforms.

✓ **Algorithms for Data Visualization** - In order to visualize the required information from a pool of random data, powerful algorithms are crucial for an accurate result.

✓ **Hardware improvements** - Amazon's ElastiCache feature helps make everything faster; cheaper SSD technologies for quicker read/write times.

✓ **Improvements in DB languages/platforms** - Amazon's DynamoDB is a recent entry in the NOSQL pool, but comes with elastic processing/storage, built in, with Amazon support - no more dedicated tech team constantly worried about falling back end server support, or server capacity, or spikes in reading/writing, etc.

✓ **Improved role specific DB solutions** - Neo4j specifically addresses graphing relationships of data sets and nodes. What about DB languages optimized for biotechnology? Astrophysics? Health care management?

✓ **Improvements in data visualization** - With the advent of touch sensitive navigation, interactive visualization technologies and themes are being taken to another level. Whether the user is a data analyst or a stay at home mom, the ability to understand and act on data is going to be democratized with these new visualization tools.

✓ **Cross-collaboration between unrelated industries** - Twitter uses Redis to handle tweets blazingly fast - Every second, on average, around 6,000 tweets are tweeting on Twitter, which corresponds to over 350,000 tweets sent per minute, 500 million tweets per day and around 200 billion tweets per year! How can gene sequencing companies and cancer research facilities learn from Twitter [14]?

Big Data is an emerging field of technology and is growing by leaps and bounds. There are numerous fields in which research can be done when it comes to Big Data.
Here is a list:

- ➢ Health Sector
- ➢ Retail Sector
- ➢ Banking
- ➢ FMCG

- Telecom
- Digital Media Solutions
- Machine Learning
- Sentiment Analysis, etc.

12. Conclusions:

Big Data has proved to create value at almost all fronts be it organizations, customers, workforce, or society as a whole. The new paradigm moves towards NoSQL databases, massively parallel and scalable computing platforms, open-source software, and commodity servers Researchers, businesses, and entrepreneurs, equally vehemently point to concrete or anticipated innovations that may be dependent on the default collection of large data sets. As organizations continue to collect more data at this scale, formalizing the process of big data analysis will become paramount. To remain competitive business executives need to adopt the new technologies and techniques emerging due to big data.

References:

[1] An Architecture for Big Data Analytics, Joseph O. Chan, Roosevelt University, USA jchan@roosevelt.edu Communications of the IIMA ©2013 Volume 13 Issue 2.
[2] Perspectives on Big Data and Big Data Analytics, Elena Geanina ULARU, Florina Camelia PUICAN, Anca APOSTU, Manole VELICANU, Institute of Doctoral Studies Bucharest.
[3] https://en.wikipedia.org/wiki/Big_data.
[4] The emergence of "big data" technology and analytics, Bernice Purcell, Holy Family University, Journal of Technology Research.
[5] http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-multi-node-cluster/.
[6] Parallel Processing of cluster by Map Reduce, International Journal of Distributed and Parallel Systems (IJDPS) Vol.3, No.1, January 2012, Madhavi Vaidya, Department of Computer Science, Vivekanand College, Chembur, Mumbai.
[7] Algorithm and approaches to handle large Data- A Survey, IJCSN International Journal of Computer Science and Network, Vol 2, Issue 3, 2013, ISSN (Online): 2277-5420, 1Chanchal Yadav, 2Shuliang Wang, 3Manoj Kumar.
[8] Disaster Prevention and Preparedness, Lelisa Sena, Kifle W/Michael, Jimma University In collaboration with the Ethiopia Public Health Training Initiative, The Carter Center, the Ethiopia Ministry of Health, and the Ethiopia Ministry of Education 2006.
[9] Big Data: Moving Forward with Emerging Technology and Challenges, International Journal of Advance Research in Computer Science and Management Studies, Volume 2, Issue 9, September 2014, Vibha Shukla1, Pawan Kumar Dubey2.
[10] https://hadoopecosystemtable.github.io/.
[11] https://en.wikipedia.org/wiki/MapReduce/.
[12] https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsUserGuide.html/.
[13] https://www.dezyre.com/article/hadoop-components-and-architecture-big-data-and-hadoop-training/114/.
[14] https://www.quora.com/What-are-the-most-important-research-topics-in-the-Big-Data-field/.