A Comprehensive Study on Big Data Techniques, Tools and Challenges

Narender Singh, *Former Assistant Professor, Dept. of Computer Application,*
Rajiv Gandhi Mahavidhyalya, Uchana, Jind, Haryana, India, 126115

Abstract- Big Data is made up of large variety of data that can't be determined by adopting the traditional way of data processing. The use of computing devices such as laptops, smart phones and wearable computing devices has increased widely; the data processing through the internet has increased more than the modern computers can bear. We have put forward a step in the field of big data – therefore the data sets that are divided on the basis of the high volume, velocity, variety, resolution and rationality and flexibility. This provides us the most precious data and also assists us in taking most relevant decisions. Just because of this growing rate, the concept of Big Data is brought into consideration. Thus, this will result in giving rise to the numbers of challenges, such as difference in the collected data and other issues like inefficiency, lack of reliability, time taking process, and security issues. Hence, the big data is something which grants lots of benefits to the big firms, corporates etc in the upcoming era. It can be the most adaptive technology so that the businesses can frame the strategies according to it. It will also help the firms in taking corrective actions. Bring this technology and adapting it, won't be easy. So, it is needed to highlight all the issues and challenges that are associated with it. This study focuses upon throwing the lights on the big data analytics and also finds out the other aspects related with the technology like tactics, issues, challenges, usage, applications, tools and techniques, etc.

Index Terms—Big Data, Legal, Data Analytics, Cloud Computing, Hadoop, Challenges.

I. INTRODUCTION

It has been noticed that from the past few years the Big Data is much adapted in the IT industry. There are number of people who are connected with the internet all over the world leads to the generation of very large amount of data at the increasing rate. Thus, it faces large amount of challenges. There are many advantages of Big Data's to the firms. But on the other hand we should focus on highlighting the challenges and issues which comes across. The Big Data service and tactics market reflects the rapidly growing opportunities all around the world as proved through prediction done by International Data Corporation (IDC). Moreover, the latest IDC forecast shows that it's technology and services market will rise at a 26.4% compound annual growth rate to $41.5 billion by 2019, or about six times the growth rate of the overall information technology market. Hence, by 2020 IDC has a strong faith that believes that the diversified business buyers will help drive analytics beyond its historical sweet spot of relational (performance management) to the double-digit growth rates of real- time intelligence and exploration/discovery of the unstructured worlds. [1].
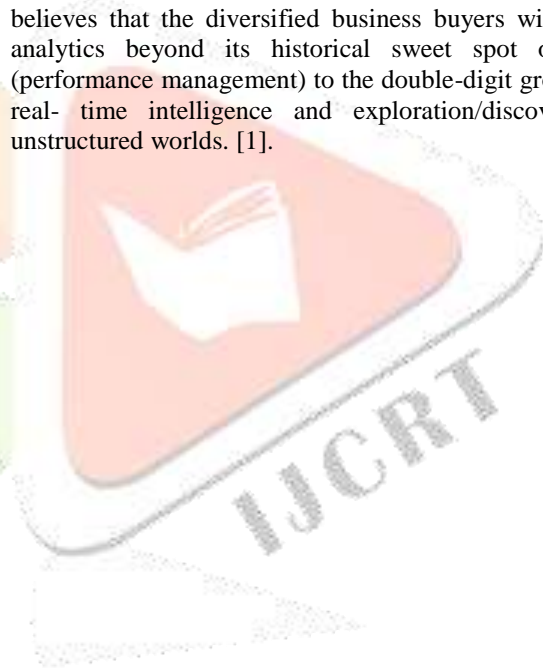
The data is being utilized by the Industries for taking the corrective actions and by making the right decisions. It also assists for business intelligence and predicting the future actions. It is important for getting the useful information. The Big data is also used by the government for making the corrective prediction so to enhance the decision making power and creates a transparency. It is needed for improving the public and national security. Healthcare presents other important aspects. As Big data provides very good opportunities to such areas [2]. It's samples are available in are astronomy, atmospheric science, genomics, biogeochemical, biological science and research, life sciences, medical records, scientific research, natural disaster and resource management, private sector, military surveillance, financial services, retail, social networks, web logs, text, document, photography, audio, video, click streams, search indexing, call detail records, POS information, RFID, mobile phones, sensor networks and telecommunications.

II. PAPER ORGANIZATION

The main aim of this study is to focus upon the challenges/issues as well as to identify the various tools and techniques used for the processing of big data and its existing solutions. And it also covers; Open issues in big data and state them properly. Section 1 gives brief introduction to big data. Section 3 analyse the importance of big data analytics, finds out the problems related to it, also pin-point the tools and techniques. Section 4 identifies the other types of techniques used for analysis, processing of the big data. Section 5 provides an insight into big data tools and technologies. Section 5 keeps attention upon different issues and challenges in terms with the different stages of big data and section 6 tells about the future extension of big data with conclusion.

III. BIG DATA

Large data sets are very important for getting the valid information so to up front the designs, trends, patterns and associations are thus, known as big data. It also defines about the high volume of data that are both the defined and undefined form. The quality of the data matters more than its quantity. The motive is to get the valuable output from the gathered, processed and analysed data. It guides us for making the strategies, accordingly and also helps in making right decisions.

Now-a-days, the word "Big Data" has become so popular. Everyone is making the use of it. Even it is being used in fields like academicians and industry experts. Many researchers have described it in different ways. But the

concept of big data dates back to the year 2001, where the challenges of increasing data were addressed with a 3Vs model by Laney (2001). It is also known as the important element of the big data. It reflects the rising Volume, Variety, and Velocity of data. The concept was used to elaborate the meaning of big data. Later on the same concept was brought in use by various enterprises including Microsoft and IBM to define the same [3].

According to Apache Hadoop (2010) defined big data as, a set of data that cannot be stored used, managed through normal computer devices. Following the same path in 2011, McKinsey Global Institute gave the definition of the big data as: datasets which behind the capabilities of typical database software tools to gather, collect, store, manage, and analyse". On the other hand, International Data Corporation (IDC) states big data technologies as a technologies of a new generation of and architectures, which are designed to withdrawal the value economically from much larger volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis" (Gantz and Reinsel, 2011) [4].

Scholars have defined big data as large size of undefined data composed through very high level of performance of different groups of applications that spread from social network to scientific computing applications. The datasets is a variety of a few hundred gigabytes to zeta bytes which means it is beyond the capacity of existing data management tools to capture, store, manage and analyse [5].

The big data has been explained in many ways but it does not have any specific definition. Some says it as what it does while some defined it as what it is. The definition of the big data on the basis of 3Vs is relative. It specifies that what the big data is defined today, may not be the tomorrow. We can understand it with an example: With the changes in the future technologies, the data which is seemed as a very important data today, will be kept safely. Later, for defining big data it was added that there is a need for understanding that how the data can be used to gain valuable information for making right decisions.

A. Dimensions of Big Data

Following are some dimensions on the basis of which the big data is classified, and it is also referred as 3V model:

a. **Volume:** It plays a very significant role. It can be collected and gathered through different sources. It is rapidly increasing from terabytes to petabytes. Thus, it helps in storing large amount of data which was earlier not possible. But now we can store big files in it due to its large storage capacity. The distribution of big data on the basis of volume is done. In other words we can say that it is the type of data which is normally known as Variety, is defines "big" data. The two different types of data that can be considered for an example are text and video of similar volume, can consider distinct type of data management technologies (Gandomi and Haider, 2015) [6].

b. **Velocity:** It refers to the estimated rate of generation of data. Previously the data analytics depends upon the bon periodic updates-daily, weekly or monthly. As the data rate is rising-up, the big data must be handled and

considered in real- or near real-time to make informed decisions. Therefore, the time plays a very critical role in such case [7]. Few domains covers up the Telecommunications, Retail, and Finance produce data with high-frequency. Thereafter, the data which is collected using mobile application covers the factors based on: location, demographics and transaction history that is useful in real-time scenario to offer personalized services to the customers. This will help in retaining the customers as well as raise the service stage.

c. **Variety:** It refers to various kinds of data that is collected and gathered. They are being explored amongst the structured data and come under the division of semi-structured and unstructured data. The data that can be arranged with the help of a pre-defined data model are called as structured data. The Structured data includes: the tabular data in relational databases and Excel. Thus, they create only 5% of all existing data (Cukier, 2010) [8]. The data that is unstructured cannot be formed using the already defined model. It includes image, graphics, , and audio. Semi-structured data falls under the classification of framed and unframed data. Extensible Mark-up Language (XML) is categorized under this category.

After sometime few other dimensions were added, which are as follows:

d. **Veracity:** It is the process of analysing whether the unreliability is concern with the data sources (Gandomi and Haider, 2015). For instance, sentiment analysis using social media data (Twitter, Facebook, etc.) is subject to uncertainty. This is very important to do the differentiation between the reliable data from uncertainty and imprecise data. And also, to manage the uncertainty related with the data.

e. **Variability:** The SAS considered the Variability and Complexity as the most vital aspects. Infect, discrepancy creates a difference in flow rate of data in the big data velocity, which commonly known as variability (Gandomi and Haider, 2015). Data is gathered through variety of sources. Thus, there are higher chances of increasing complexity in managing data ranging from transactional data to big data. The data collected from many geographical locations have several semantics (Zikopoulos et al., 2012; Forsyth, 2012) [9].

f. **Low-Value density:** the collected data is useless in its original form. Therefore, the data is examined properly in order to extract the meaning out of it (Sun and Heller, 2012). Jotting the Illustration to specify the statement, logs from the website cannot be used in its initial form to obtain business value. One should properly examine it before making the prediction about the behaviour of the user.

B. Sources of Big Data

After knowing about the big data and its dimensions it is needed to understand the different ways of it. It is described briefly. The new way of extracting the data is done through digitization of content (Villars et al., 2011) Increment in technology results in generating the high rate of data

generation. For example, one of the biggest surveys in Astronomy, Sloan Digital Sky Survey (SDSS) has recorded a total of 25TB data during their first (2000-2005) and second surveys (2005-2008) combined. With the advancements in the resolution of the telescope, the amount of data collected at the end of their third survey (2008-14) is 100 TB. Use of "smart" instrumentation is another source of big data. Smart meters in the energy sector record the electricity utilization measurement every 15 minutes as compared to monthly readings before.

Now-a-days social media has also become one of the convenient ways of getting the data. The data related to the hospital, industry, hotel, agriculture, medical care, etc industries of the smart cities developed based on IoT is gathered.
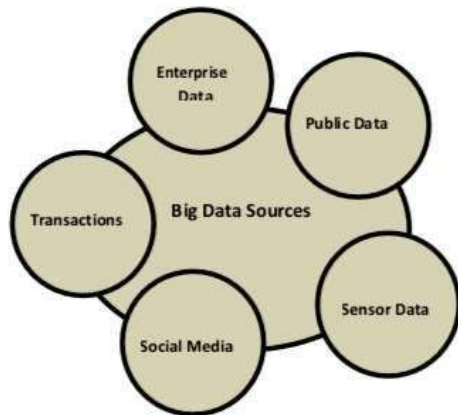

Fig.1. Some Sources of Big Data.

C. Need of Big Data Analytics

The big issue which comes in front is storing the large amount of structured as well as unstructured data at a described time period. Few of those loopholes led to handle and process vast amount of data with the traditional storage techniques which pin-point to the emergence of the term Big Data. The big data has gained lots of attention due to emergence of Internet. But there is a big difference between the both. There are no similarities between them. It has become quite easier to get the information and share it across the world using the internet. Even we can share the raw data through it. It shows that how we can store, use, manage, process the data so that it can be useful in future whenever needed. It is useful for making the prediction about future actions.

Marketers keeps an eye upon the target audience, like insurance agents focus on doing the customized insurances to their customers, and healthcare providers concentrates on providing high quality and low-cost treatment to patients. Though, there are lots of improvement have been done regarding data storage, examination, collection of data, future prediction etc. Still, it is required to understand various other factors that play a vital part in making strong models. It handles very large data and provides accuracy (Boyd and Crawford, 2011) [11].

The main aim of current big data analytics is to analyze traditional ways. It covers pattern mining, decision trees, rule-based systems and other data mining techniques to establish rules of the business effectively and efficiently at large

concern. It can be attained either by establishing algorithms that makes the use of assigned data storage, in-memory computation or by using cluster computing for parallel computation. This process was followed by making the use grid computing in earlier days. But now it has been left behind by using the cloud computing in recent days.

D. Phases in Big Data Processing

The big data should obtain the data and record it from different sources before processing it. Once the data is recorded it should be filtered properly. So, that it can be clearer and preside. Only the appropriate data should be recorded after doing the filtration of data. The aim is to discard the irrelevant information. For filtering the data several specialized tools are used such as ETL. ETL tools that shows the ways in which the data gets loaded into the warehouse. The figure 3 shows the stages in the process.
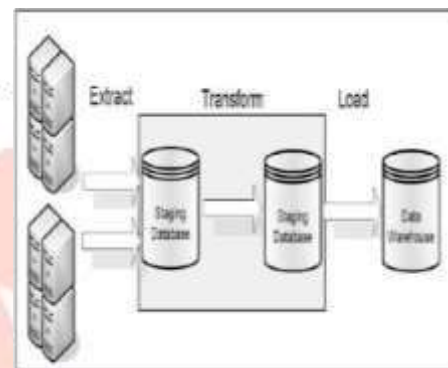

Fig. 2. ETL process.

Sr. No.	Phase	Description of phase
1.	Extraction	In this stage only relevant information is derived out of available data. The data extracted is very relevant and useful as it is extracted using the most effective tool i.e. ETL process.
2.	Transformation	1. Data is transferred from variety of phases. The phases are Data analysis; 2. Definition of transformation workflow and mapping rules; 3. Documentation; 4. Resolution; and 5. Backflow of cleaned data.
3.	Loading	At the last, after the data is in the required format, it is then loaded into the data warehouse.

E. Big Data Value Chain

In the year 1980, Porter highlights the concept of Value Chain [12]; he defined it as a set of activities which are carried out by a firm to increase the value at each step of delivering a product/service to its customers. We can also define it as the framework that deals with a set of activities to create value from available data. It can be divided into seven phases: data generation, data collection, data transmission, data pre-processing, data storage, data analysis and decision making.

1. **Data Generation:** The very first step in value chain is the generation of data. The data can be generated from a ton of sources that include data from Call Detail Records (CDR), blogs, Tweets and Facebook Page.

2. **Data Collection:** In the second stage, the data is abstracted from every possible data sources (Miller and Mork, 2013; Chen et al., 2014). For instance, to predict the behaviour of the consumers towards Telecommunication. Thus, the information can be gathered using the CDRs and through the reviews (either positive or negative) of the users on Social Media Sites such as Instagram, Facebook, Twitter, etc. The tactics which are generally used are: sensors, log files, web crawlers and network monitoring software (Chen et al., 2014) [13].

3. **Data Transmission:** As soon as the information is obtained, than the data is transferred and stored in a storage device. Then, the data is processed for further examination and analysis. It is done through two ways: Inter-Dynamic Circuit Network (DCN) transmission and Intra-DCN transmissions. Inter-DCN transmission manages the transfer of data from the data source to the data centre. Later it assists in sharing the data to the data centre. Instead of only storing the data it also helps in gathering, organizing and arranging the data.

4. **Data Pre-processing:** The data collected from different kind of sources might be quite excessive, noisy and irregular. Hence, the data is already processed to facilitate and improve the quality of the data that is necessary for examination. It even helps in enhancing the exactness and efficiency of the analysis and even cuts off the cost of storing the data.

5. **Data Storage:** The big data storage systems offer a big space for storing large data. The different types of data storage devices for big data must undertake the factors like consistency (C), availability (A) and partition tolerance (P). As per the CAP theory introduced by Brewer in the year 2000 has stated that; the 2 different kinds of requirements cannot be fulfilled by different kinds of storage devices, which means, either consistency and availability or availability and partition tolerance or consistency and partition tolerance but not all requirements simultaneously (Gilbert and Lynch, 2002) [14]. The research on Considerable is still on in the area of big data storage mechanism. A little bit timely changes are to be done in relation with Google File System (GFS), Dynamo, Big Table, Cassandra, Couch DB, and Dryad.

6. **Data Analysis:** After the collection, transformation and storage of the data, the stage comes for data analysis. It can be done using the steps as follows:
   a) Define Metrics
   b) Select architecture based on analysis type
   c) Selection of appropriate algorithms and tools
   d) Data Visualization

7. **Decision Making:** After doing all the analysis and visualizing the result, it becomes quite difficult for the decision maker to decide what is right and what is not. It becomes difficult to choose whether and how to reward a good behaviour and change the bad one. One can go through in details about a particular problem so, to understand it better and provide the better solutions to it. It also helps in taking right actions and finding out the best ways (Miller and Mork, 2013) [15].

IV. TOOLS AND TECHNIQUES FOR BIG DATA HANDLING

A. Big Data Processing Tools [16]

Hadoop divided the file system (HDFS) and lot of similar elements such as Apache Hive, HBase, Oozie, Pig and Zookeeper and thus, they are divided into following ways:

HDFS: It is a highly faults tolerant distributed file system which is required for storing data in the clusters.

Map Reduce: It is a type of technique used for processing the high amount of distributed data.

HBase: A column-oriented distributed NoSQL database for random read/write access.

Pig: It is a high-level data programming language used for identifying data of Hadoop computation.

Hive: It is a data warehousing application that provides relational model and a SQL-like access and.

Sqoop: It is a type of project used for sharing the data between relational databases and Hadoop.

Oozie It is a kind of workflow management that is totally dependent Hadoop jobs.

B. Big Data Analysis Tools [17]

Hadoop and Map Reduce: For processing huge datasets and works on the principle of divide and conquer it is regarded as a programming model. Map step and Reduce Step are the two points which are: instrumented in divide and conquer method.

Following are the two types of nodes: Master node and worker node. In map step we perform the Division of input into tinier sub-problems and then disseminate them to worker nodes. Subsequently, the master node the outputs for the entire sub- problems are combined in reduce step. Furthermore, with the help of Hadoop and Map Reduce fault-tolerant storage and high output data processing can be possible.

Apache Mahout: For big data it provides scalable and commercial machine learning techniques and smart data analysis. Some core algorithms of the mahout are Clustering, classification, pattern mining, regression, dimensionality reduction, evolutionary algorithms, and batch based collaborative filtering. For attenuating big challenges, Apache mahout is to provide a tool.

Apache Spark: Speed processing, and sophisticated analytics is given by this framework. In java, scala, or python we are able to write applications by the help of it. Map Reduce operations, SQL queries, streaming data, machine learning, and graphs data processing are supported by it. Three elements that are driver program, cluster manager, and worker nodes are included inside it. On the spark cluster, driver program work as the major application point of the application. To perform the data processing in the form of tasks, cluster manager assigns the resources and worker nodes. This is considered as the main advantage: Deploying spark applications in an existing Hadoop clusters.

Dryad: For executing parallel and distributed programs it is considered as the programming model, and also for managing large databases. It has number of computing nodes. It is a sort of tool that runs the data-parallel

programs. On the other hand, Dryad made the use of different kinds of machines with various processors and performs concurrent programming.

Storm: It is divided and combined tool of processing unstructured streaming data. It is preferable for real-time processing in comparison of batch processing. It is very simple to use and easy to understand. It also helps in getting effective results. It is not so costly and open source distributed real-time computation system.

Apache Drill: It is a distributed system for interactive analysis of big data. It handles various queries related with languages and many kinds of data formatting. It is very useful in exploring the fixed data. It has an extraordinary ability to line up the tons of servers. It also expands the massive data (petabytes to zeta bytes) and records trillions of data within a second. It is very fruitful to make use of the HDFS for storage and map reduce for batch analysis by the drill.

Jaspersoft: It is a type of software that is used for creating reports from database column. It is generally; used as an analytical tool that gives the clear vision about the storing data. It includes Mango DB, Cassandra, and much more. There is no such tool that can analyze the thing without extraction, changes and loading (ETL) except one, i.e. one of the Jaspersoft does. It is very interesting that it does not require ETL. It can create HTML reports and dashboards directly from big data store as it is not dependent upon ETL. It provides exact and accurate results and it can be shared anywhere, at any time.

Splunk: It makes the complete use of the big data generated using the machines. It is a real-time and smart platform. It makes the use of the advanced and updated technologies. Big data are very simple and easy. It allows the access of the web to the users. The outputs of the Splunk are shown or represented in the form of pictures, graphs, digits, reports, and alerts. It is used looking after all kinds of data including structured, semi-structured and unstructured. It is good at resolving problems for system and information technology infrastructures.

V. CHALLENGES AND ISSUES

Due to the several characteristics of big data such as: velocity, volume, validity, value, variety etc give rise to various issues. The Figure 3 shows different types of challenges come across big data. The following figure shows the list some of the challenges in big data along with its impact and risks involved. [18]

1. Finding relevant data: It is always needed to explore new things, technologies, tools and tactics in order to bring something new. It can be done through finding the useful data which is very essential for using new tactics, tools etc. For e.g. selecting representative (diversity).
2. Machine Learning: The better integrate HPC is very necessary. High performance computing refers to any computational activity requiring more than a single computer to execute a task. Supercomputers and computer clusters are used to solve advanced

computation programs.") And Data Analytics (Machine learning).

3. Need of Data Management and Data Analytics: Both management and analytics require different software and hardware characteristics, therefore, we will need machines that do support parallel machine learning to do analytics and management together
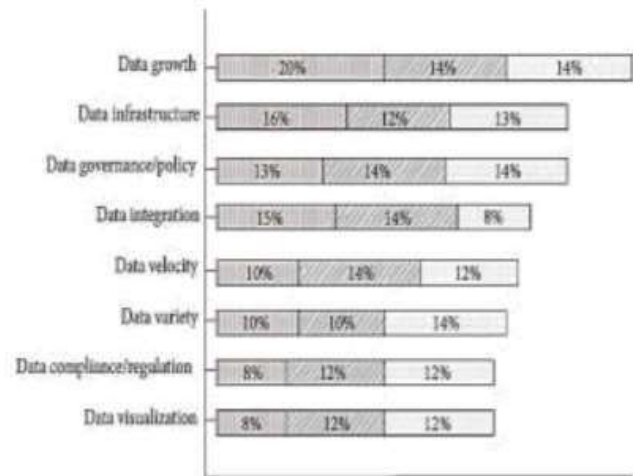

Fig.3. Challenges in Big data.

.

4. Need for Efficient Parallel Algorithms: There is need of developing efficient parallel algorithms for purpose of streaming and multiscale adaptive algorithms which reduce time complexity of O (N2) to O (N).
5. Need of Secure Cloud: It is very important to keep the data safe and secure i.e. it should not be leaked. In order to secure the data it is important to find the better solutions.
6. Need for Better Adaptations to Classical Algorithms: It is important for handling the rising problems without many efforts.
7. Need for Better Transfer of Data from One Location to Another: It helps in easy transfer of data.
8. Security and Privacy of the Data: Using highly advanced encrypted algorithm for the security of the information stored online.
9. Increased Need to Store Data: It is very convenient to store the data online. It provides safety, recovery of data, privacy of data etc.
10. Requires Massive Parallelism: We require massive parallelism in computation as well as storage; it could be achieved by following ways

Data Processing across millions of servers

Data Management across millions of data devices

Managing and establishing new data infrastructure system;

MapReduce helps in achieving above, it is a programming model to support distributed computing but still, we need some really good tools and techniques to acquire parallelism.

11. Needs Big Data Resource Management: Relational databases are outdated for big data problems because of lack of scalability. Non- relational databases (NoSQL) are best for data store relaxing consistency; it provides volume scalability and better capacity to digest a variety of data.

12. StorageTechnology:Non-VolatileMemory

Evolution: Managing input data is a hard task and so, merge and management of intermediate results is important, the solution to the storage problem is better and advanced storage strategies for grasping high-performance in-memory key/value databases to speed up data-intensive tasks.

VI. CONCLUSION

Every day there is very large number of data being produced. Due to very large file of data it becomes quite challenging to attain efficient processing, using the existing traditional tools. Big data is the data that exceeds the working capabilities of normal database systems. In this study, the fundamental aspects of the Big Data are brought into consideration. It also covers Big Data, its characteristics, the problems and issues arise out of it and techniques for handling big data.

In the upcoming future, the issue of lack of expertise to perform best practices will be vanished. There will be lots of variety of products and services for different types of big data. Hence, a ton of firms will give the cloud platform. The numbers of competitors, competing with the cloud service providers. A strong digital platform will be formed that will grow more, provides better data consistency and better results. In comparison with today, the media analytics will be increased about 10 folds. Cognitive computing will be used on regular basis by the people. Hadoop will get better in future. Many new technologies and tools with the high capability of storing data will be used to store the data, do examination, measure and incorporate all forms and formats of data present surround us.

REFERENCES

[1] https://www.idc.com/prodserv/4Pillars/bigdata.

[2] T. Murdoch, A. Detsky, "The Inevitable Application of Big Data to Health Care", JAMA, 2013, 309(13):1351-1352.

[3] Meijer, E., 2011. The world according to LINQ. ACM Communications 54 (10), 45-51.

[4] Gantz, J., Reinsel, D., 2011. Extracting value from chaos. Tech. rep., IDC.

[5] Cuzzocrea, A., Song, I.-Y., Davis, K. C., 2011. Analytics over large-scale multidimensional data: the big data revolution! In: Proceedings of the ACM 14th International workshop on Data Warehousing and OLAP (DOLAP '11). ACM, New York, NY, USA, pp. 101-104.

[6] Gandomi, A., Haider, M., 2015. Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management 35 (2), 137-144.

[7] Foster, I., Zhao, Y., Raicu, I., Lu, S., 2008. Cloud computing and grid computing 360-degree compared. In: Grid Computing Environments Workshop, 2008 (GCE '08). pp. 1-10.

[8] Cukier, K., 2010. Data, data everywhere: A special report on managing information. The Economist http://www.economist.com/node/15557443, Accessed on: June 15, 2015.

[9] Mukherjee, A., Pal, A., Misra, P., 2012. Data analytics in ubiquitous sensor-based health information systems. In: 6th International Conference on Next Generation Mobile Applications, Services and Technologies (NGMAST). pp. 193-198.

[10] Minelli, M., Chambers, M., Dhiraj, A., 2013. Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today"s Businesses, 1st Edition. Wiley Publishing, Wiley CIO Series.

[11] Boyd, D., Crawford, K., 2011. Six provocations for big data. In: A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society.

[12] Porter, M., 1980. Competitive Strategy: Techniques for Analyzing Industries and Competitors. The Free Press, New York.

[13] Chen, M., Mao, S., Liu, Y., 2014. Big data: A survey. Mobile Networks and Applications 19 (2), 171-209.

[14] Gilbert, S., Lynch, N., 2002. Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services. SIGACT News 33 (2), 51-59.

[15] Miller, H. G., Mork, P., 2013. From data to decisions: A value chain for big data. IT Professional 15 (1), 57-59.

[16] D. P. Acharjya and Kauser Ahmed P, "A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools" International Journal of Advanced Computer Science and Applications(IJACSA), 7(2), 2016.

[17] Getaneh Bernie Tarekegn PG, Department of Computer Science, College of Computing and Informatics, Assosa University, Assosa, Ethiopia Yirga Yayeh Munaye MSC, Department of Information Technology, College of Computing and Informatics, Assosa University, Assosa, Ethiopia "BIG DATA: SECURITY ISSUES, CHALLENGES AND FUTURE SCOPE", July–Aug 2016, pp. 12–24.

[18] https://www.progress.com.