A Review on the Study of Big Data with Comparison of Various Storage and Computing Tools and their Relative Capabilities

Yogesh Kumar Gupta
Assistant Professor, Banasthali University
Newai, Rajasthan, India

Dr. C.K. Jha
Associate Professor, Banasthali University
Newai, Rajasthan, India

Abstract- The Big Data era is in full force nowadays because the world is changing. The name "Big Data" referred to as the massive amount of gigantic dataset in Zetabyte or bigger-sized dataset generated from various sources in daily life such as medical science or healthcare data, social media sites, satellites data, various sensors etc., with very high velocity. This tremendous data is no more time stable in working area; rather it is updated according to time at rapid speed with different format, so that the conventional database management system or algorithms is unable to handle such kind of biggest dataset, thus the big data need to describe a very novel tools, technique and algorithms to handle such massive amount of huge data to acquire, store, distribute, handle and analyze. The process for analyzing the complicated data normally concerned with the disclosing of hidden patterns. This paper firstly give the introduction of big data in detail and mainly it concerned about two big aspects related to storage and computing tools, to overcome the several critical challenges. Also, we are going to compare some available well known computing and storage tools and techniques on the behalf of most prominent parameters or capabilities that a glance helps us for prior to selecting a tools with its application province to hold Big Data.

Keywords -Big Data, Storage Tools, Computing Tools, Comparative Capabilities, Challenges, opportunities.

I. INTRODUCTION

"Big data" is relatively new concept or most emerging area nowadays in IT companies, industry, academic, medical science and different kinds of business trades. However, several scientist and trainees has been utilizing the term big data for analysis. For instance, referred to big data as a large amount of researcher data for visualization and processing.
The Big Data came into real world when the traditional relational database management system(RDBS) was unable to maintain the Unstructured (text, audios and videos media, pictures, social updates, human behavior, medical science and marketing data) and semi-structured (weblogs files) data that is near about 90% according to IBM survey, generated today by various categories of organization, medical science, social media websites, sensors, devices, industry or from any source which producing data [6]. The data that is so massive in volume has different form in variety or moving with such velocity is known as Big Data [6]. Big Data is more challenging for storing, processing and analyzing. It involves large distributed file systems in commodity hardware for storing which should be more flexible, fault tolerant, scalable and reliable [6]. The tools and techniques used for big data analysis to manage massive amount of huge data are Hadoop, Map Reduce, NoSQL database, HPCC and Apache Hive. These technologies manage large volume of huge data in KB, GB, MB, TB, YB, PB, EB and ZB [6].

The rapid increment of Internet has been led to massive volume of huge data available online [3]. This data required to be stored, processed, analyzed, and extract to get valuable or meaningful Information in less redundant and well managed form. In order to store, access, process and manage massive volume of huge data available online form many resources and the data that is generated in structured, semi-structured and unstructured form, Data computational statistical tool and methods are needed which specify the need of searching, analyzing, data mining, and visualizing the massive amount of data and information [3]. There are various data computational and analyzing technologies (DFS in parallel, map reduce model) are available which perform computation on massive amount of huge data and dedicate most of their processing time to input/output and updating of such data [3]. The process for analyzing the complicated data normally concerned with the disclosing of hidden patterns.

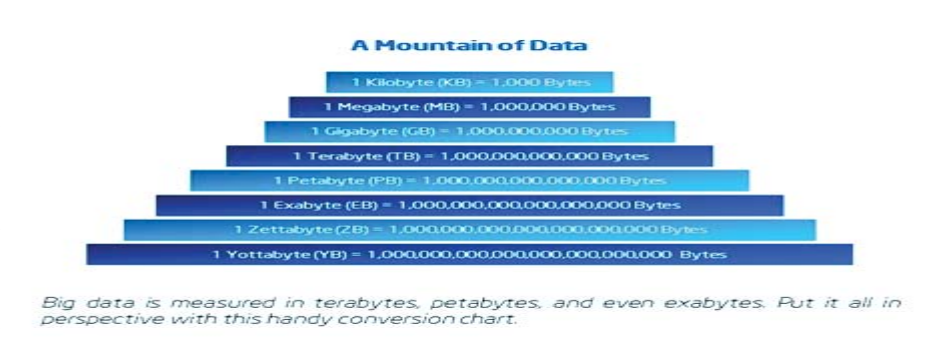

Figure 1 – Illustrates how big data is measured [25]

CHARACTERISTICS OF BIG DATA

As we all know the big data referred to as a large volume of huge dataset which is categorized into five main characteristics or 5V's such as Velocity, Volume, Variety, Veracity and Value. Each aspect puts challenges in processing and handling massive volume of huge dataset to extract some meaningful information. Such kind of challenges could be in data acquisition, recording, searching, sorting, retrieving, analyzing, and visualizing from a variety of already described key features of the Big Data.

Figure 2: Big Data Characteristic [23]

1. **Volume**: It is referred to the size of dataset in different amount such as KB, MB, GB, TB, PB, YB, EB and ZB of information. So the massive amount of huge data results into large data files [8].
2. **Variety**: Data generating sources (even from same categories or from different) are excessively heterogeneous. The files are coming from different sources, in different formats and types, may be unstructured, semi-structured or structured [8] .On twitter 400 million tweets are sent daily and there are 200 million active users on it [14].
3. **Velocity**: The large amount of data comes from various sources at high speed on which it is developed and processed. For example social media posts [8].
4. **Veracity:** - means accuracy or uncertainty of data. Data is uncertain due to the incompleteness and inconsistency [13]. Veracity is the very critical and big challenge in data analysis when doing comparison of two things like velocity and volume [19].
5. **Value:** - Massive amount of huge data has different kinds of values such as statistical data, events, correlations, hypothetical data etc.

BIG DATA FORMATS

The Big Data sources mainly categories into following formats:

- **Structured Data** – the data produced from several research article and generals, business applications such as retail, finance, bioinformatics and other such traditional databases in various sources such as RDBMS, OLAP and data warehousing etc [2].
- **Unstructured Data** – Data created by the users as healthcare data, trading markets data, social media sites, Web forums, feedback, emails, comments, audios, images, videos etc. or it may be created by machine as various sensors data, web logs, online transaction data etc [2].
- **Semi-structured Data** - XML formatted data, HTML, CSV and RDF.

HISTORY OF BIG DATA FROM VARIOUS SOURCES

A recent study projected that every minute the users of E-mail send more than 200 million emails, Twitter users produce 277,000 tweets [22], users of YouTube upload 72 hours of video, Google receives more than 4 million

searching queries and users of Facebook share more than 2 million pieces of substance and more than 350 GB of data is processed. The approximated figure about the tremendous amount of data is that, it was estimated near about 5 exabytes (EB) till 2003, 2.7 Zetabytes (ZB) till 2012 and it is expected to grow near about 4 times greater till 2016 [20]. The Figure 3 shows the data in Terabytes (TB) for the year 2001-2012 [10]. From Year 2005-2012, it would appear from this graph that the amount of data was exponentially grow within this period due to the significant contribution of Big Data Analytics. And figure 4 shows the Forecast of Transition and Size of Big Data Analytics Market from the year 2012 – 2020[18].

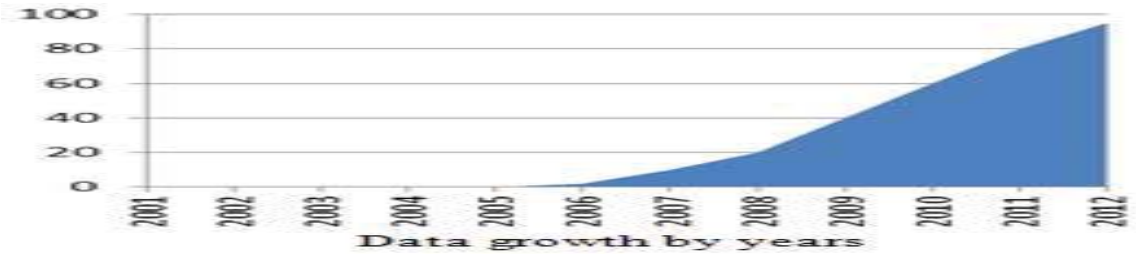

Figure 3: History of Big Data [22]

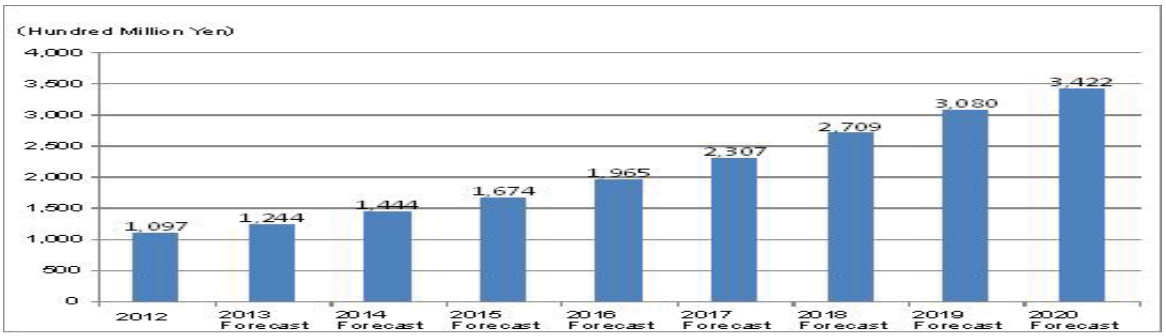

Figure 4: Forecast of Transition and Size of Big Data Analytics Market [24]

II CHALLENGES AND OPPORTUNITIES

As we all know that the tremendous amount of data generated day to day life from various sources such as online transaction, e-mails, research generals and articles, social media sites (Facebook, Twitter, WhatsApp etc.) and web forums, different sensor's data composed from various sources such as healthcare science, environmental organizations, meteorological department, business strategically data, trading market, company data being generated daily life etc. in different format such as structured , semi structured and unstructured etc. With a great Velocity is normally referred to as Big Data. As an outcome, the conventional database computation tools and algorithms as well as data storage and management techniques has not able to deal with these data. So we need to look on big data challenges and design some computing models for efficient analysis of data [14].

CHALLENGES WITH BIG DATA [15]

1) Heterogeneity and Incompleteness - When we want to process the data, it should be structured but when we deal with the Big Data, data may be unstructured, semi-structured or structured as well. Heterogeneity is the tremendous challenge in data Analysis and analysts need to survive with it. For example a patient in a Hospital, We will make each record for each diagnosis test. And we will also make a record for hospital stay. This will be different for all patients. This design is not well structured. So managing with the Heterogeneous and incomplete is required. A good data analysis should be applied to this [13].

2) Scale - The continuously growing large amount of huge data is a big challenging issue from many past years. These challenges was moderated by processors to achieve high speed, prescribed Moore's law, use to allow the various resources required to manage large amount of huge data. Commonly, as usual an aspect running now that the large data is scaling much faster than computing CPU speeds and resources are static[11].

3) Timeliness - The important aspect of size or volume is speed. When processing a large dataset, it will take much more time to analyze them. To design a model that can successfully deals to process a given size of dataset as much as faster. However, this speed is not enough in the perspective of Velocity of Big Data. Relatively, there is an

acquisition rate challenge, For example, if there is any fraud transaction, it should be analyzed before the transaction is completed [15]. So some new system should be designed to meet this challenge in data analysis.

4) Privacy - Privacy of data is another big problem with big data. In some countries there are strict laws regarding the data privacy, for example in USA there are strict laws for health records, but for others it is less forceful. For example in social media we cannot get the private posts of users for sentiment analysis [13].

5) Human Collaborations - In malevolence of the tremendous advanced computational models which have many textures information's that users can simply identify but the algorithms have difficulties for searching. Preferably, Big Data analytics is not capable for computing all task, somewhat, it is designed externally which have a human interface. In present era, A Big Data analytics tools must accept the input from multiple experts with several domains to recognize what is currently going on, and distributed the results of exploration. These Experts may be divided into time and space when it is more costly to accumulate a whole team collectively in one room. Thus the aspect of big data has to support this distributed parallel expert input, and their association [15].

BIG DATA OPPORTUNITIES [7]

Now this is Data Revolution time. Big Data is giving so many opportunities to business organizations to grow their business to higher profit level. Not only in technology but big data is playing an important role in every field like health, economics, banking, and corporate as well as in government.

1) Technology - Almost every top organization like Google, Facebook, IBM, Yahoo have adopted the Big Data and investing on them. A current study projected that every minute the users of YouTube upload 72 hours video, Google receives more than 4 million searching queries, users of e-mail send more than 200 million emails, users of Facebook share more than 2 million pieces of substance and processed the data more than 350 GB on it, every minute more than 570 websites are created and Twitter users produce 277,000 tweets [16]. From these stats we can say that there are a lot of opportunities on internet and social media.

2) Government - The analytics of big data can be used to handle the problems faced by the government. Obama government declared research on big data and development initiative in 2012. Big data analytics played an important role of BJP winning the elections in 2014 and Indian government is applying big data analysis in Indian electorate.

3) Healthcare - According to the survey of IBM Big data for Healthcare, 80% of medical data is unstructured. Organizations of Healthcare are adapting big data technology to get the complete information about a patient. To improve the healthcare and low down the cost big data analysis are required and certain technology should be adapted.

4) Science and Research - Big data is a most emerging area of research. Many researchers are working on big data. There are so many papers being published on big data. The center of NASA for climate simulation stores 32 petabytes (PB) for observations [14].

5) Media - Media is using big data for the promotions and selling of products by targeting the interest of the user on internet. For example social media posts, data analysts get the number of posts and then analyze the interest of user. It can also be done by getting the positive or negative reviews on the social media.

III BIG DATA COMPUTATION MODEL

For retrieving the meaningful information from the massive volume of huge data or big data, there are some suitable tools and techniques needed to achieve data acquisition, data storage and computation for a variety of analytical perspectives. The Big Data computation process model is shown in Figure 5.

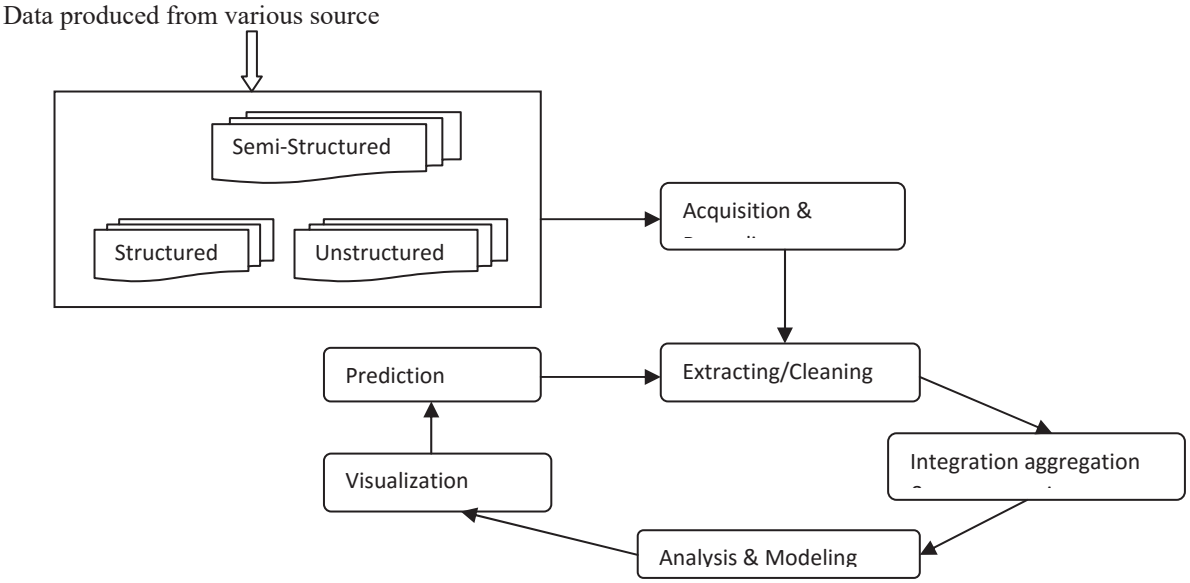

Figure 5:- Process Flow model for Big Data computation [4]

PAPER ORGANIZATION

The whole paper is divided into five sections. The section one introduces the concept of Big Data, heterogeneity in various formats and sources, history around the world and their characteristics. The section two introduces the opportunities and challenges related to big data. The section three introduces the computation model of big data basis on some terminologies. The section four discusses the requirement analysis of tools and technology to handle the processing and storage of Big Data with several application domains. It shows the comparison of some well known processing and storage paradigm on the behalf of their limitations and abilities. At last section five concludes this study with their some positive implication and blessing.

IV DISCUSSION AND RESULTS

The whole organization of big Data includes data acquisition, data storage, processing and analyzing it for a variety of purposes; such as to visualize the infrastructure, to make the hold on Big Data related tasks. Using the deep analysis of different computational and storage tools, we have got various prominent parameters or capabilities that help us to do comparative study about these tools. The several pros/cons of these tools and technologies show the correctness of the tools in different types of application domain.

COMPARITIVE STUDY OF PROCESSING TOOLS ON VARIOUS PARAMETERS [4]

Computing Tools/Parameters	**Hadoop & MapReduce**	**Cloudera Impala RTQ**	**IBM Netezza**	**Apache Giraph**
Scalability	Yes	Yes	Yes	Yes
Distributed Architecture	Yes	Yes	Yes	Yes
Parallel Computation	Yes	Yes	Yes-Asymmetric Massively Parallel Processing	Yes-Bulk Parallel processing
Fault Tolerance	Vary High	Yes	Yes-using redundant SMP hosts	Yes- By Check Point
Single Point Failure	Yes- At master nodes	Yes -If any node fail query execution then whole query process is terminated	At SMP server level	No - Multiple master threads running
Query Speed	Slow	High	High	High
Real Time / Response Time Analytics	No	Yes / in seconds	Yes / in seconds	Yes / Vary Less

Streaming query Support	No	No	Yes	No
ETL Required	No	No	No	No
Data Format for analytics	Structured/ Unstructured	Structured/ Unstructured	Structured (RDBMS)	Graph Database
I/O optimization	No	Yes	Not Required	Not Required
Optimized query Plan	Not Applicable	Yes	Yes	Yes – reference to graph query/ algorithms
Latency Time for query	Not relevant	Low - owing to employ a devoted distributed query engine	Very Low latency- in seconds, owing to database processing and parallelism in memory	Low - In memory computation

COMPARITIVE STUDY OF STORAGE PARADIGMS/TOOLS ON VARIOUS PARAMETERS [4][15]

Storage Tools	**Hbase**	**Apache Hive**	**Neo4j**	**Apache Cassandra**	**MongoDB**	**Redis**	**Drizzle**
Open Source	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Distributed	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Scalable	Yes	Yes-Good	Yes	Yes - Vast	Yes	Yes	Yes
Data Storage Format	Structured- that means table but not exactly full Relation	Structured/ Unstructured	Non-tabular or relational i.e. graph database(schema less)	Structured / Semi-structured / unstructured (schema less)	Documents storage	Key-Value store	Relational DBMS
ETL Required?	Yes	Yes - in minutes a little bit high latency	No	No	No	No	No
Failover Recovery	More time at node level failure where as at Region Server level	Yes - supports node level recovery	Yes - Select the new master	Yes- for Recovery performance it is optimized	Yes	Yes	Yes
Concurrency	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fault Tolerance	Yes	Yes-replication method to have synced with meta-store	Yes -support by ACID Transaction	Yes -Optional	Yes	Yes	Yes
MetaData Store	Yes	Yes	Yes -not mandatory	Yes –owing to support flexible	Yes	Yes	Yes
Access Control	Yes	No- built-in security provisions	Yes	Yes -Provided by the DataStax Enterprise	Yes	Yes	Yes
Acid Transaction Support	Yes-RollBack Support	Yes	Yes	Yes-Provides AID only	Yes	Yes	Yes
Real Time Query/ OLTP	No	No	Yes-In form of graph traversal, Insertion and deletion of nodes	Yes	No	No	No

Replication Method	Selected Replication factor	Selected Replication factor	Selected Replication factor	Selected Replication factor	Master – Slave Replication	Master – Slave Replication	Master – Master Replication Master – Slave Replication
Stream Query Support	NO-Partially	No	No	Yes	No	No	Yes
Durability	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Range Of SQL support Queries	No support of SQL-can support when integrated with HIVE	Limited-through HiveQL that has been extended through writing custom functions	Queries in the form of graph traversal	Yes-Through CQL whereas JOINs and the most SQL search are supported by defining schema	No	No	No
Single Point Failure	Yes-At region server level	Yes-At master node of underlying Hadoop	Yes- At master level responsible for write replicas	No-hance high Availability	Yes	Yes	Yes

V. CONCLUSION

The main objective of this study is to describe the fundamental aspect of Big Data such as heterogeneity in various formats and sources, about history and computation model basis on some terminologies, opportunities and challenges linked with it; and the characteristics along with 5Vs such as volume, veracity, velocity, value and variety of Big Data. As we all know that big data is most emerging area that is needs to store and analyze by using some well known existing tools and technique, so this paper mainly focuses to find and compare some available computing and storage tools and techniques based on some parameters that are going to be used in current scenarios to specify the several challenges of Big Data computation. The comparison is done in the behalf of most prominent parameters or capabilities that a glance helps us for prior to selecting tools with its application province to hold Big Data. The process for analyzing the complicated data normally concerned with the disclosing of hidden patterns.

REFERENCES

[1] Bhardwaj, N. and Balkishan, Kumar, A., " Big Data and Hadoop: A Review", International Journal of Innovative Research in Science, Engineering and Technology, An ISO 3297: 2007 Certified Organization, Vol. 4, Issue 6,ISSN-2319-8753 June 2015.

[2] Gupta, Y. K. and Jha, C. K., "Study of big data with medical imaging communication", International conferences on communication and computing systems(ICCCS-16), CRC Press (Taylor & francis group)-2016 –Paper accepted.

[3] Fadnavis, R. A. and Tabhane, S. (2015) "Big Data Processing Using Hadoop", International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 6 (1), 2015, 443-445.

[4] Prasad, B. R. and Agarwal, S., "Comparative Study of Big Data Computing and Storage Tools: AReview", International Journal of Database Theory and Application Vol.9, No.1 (2016), pp.45-66 (2016).

[5] Shilpa and Kaur, M., "BIG Data and Methodology-A review", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 10, October 2013.

[6] Sabia and Arora, L. (2014), "A survey: Technologies to handle big data", International Conference on Communication, Computing & Systems (ICCCS–2014).

[7] Duggal, P. S. and Paul, S., "Big Data Analysis: Challenges and Solutions", International Conference on Cloud, Big Data andTrust 2013, Nov 13-15, RGPV.

[8] Patel, A. B., Birla, M. and Nair, U. (2012),"Addressing Big Data Problem Using Hadoop and Map Reduce", International Conference on Engineering (NUiCONE), 2012, Nirma University, 1 – 5.

[9] Dhomse, G., Komal, K., Manali, L., Latika, A., "A Review Approach for Big Data and Hadoop Technology", International Journal of Modern Trends in Engineering and Research, www.ijmter.com,e-ISSN No.:2349-9745, Date: 2-4 July, 2015.

[10] Jha, A., Dave, M., Madan, S., "A Review on the Study and Analysis of Big Data using Data Mining", International journal of latest trades in engineering and technology(IJLTET)ISSN- 2278-621X -2016.

[11] BIG DATA: Challenges and opportunities, Infosys Lab Briefings,Vol 11 No 1, 2013.

[12] Sriramoju, S. B., "A Review on Processing Big Data", International Journal of Innovative Research in Computer and Communication Engineering ,ISSN(Online): 2320-9801, (An ISO 3297: 2007 Certified Organization), Vol. 2, Issue 1, January 2014.

[13] Agrawal, D., "Challenges and Opportunities with Big Data", Acommunity white paper developed by leading researchers across the United States.

[14] Beakta, R.,"Big Data And Hadoop: A Review Paper", Volume 2, Spl. Issue 2 (2015)e-ISSN: 1694-2329 | p-ISSN: 1694-2345, " RIEECE.

[15] Bhosale, H. S.and Gadekar, D. P., "A Review Paper on Big Data and Hadoop", International Journal of Scientific and Research Publications, Volume 4, Issue 10, ISSN 2250-3153, October 2014.

[16] Tiwarkhede, A. S. and Kakde, V., "A Review Paper on Big Data Analytics", IJSR, Volume 4, 2015.

[17] Elgendy, N. and Elragal, A., "Big Data Analytics: A Literature Review Paper", P. Perner (Ed.): ICDM 2014, LNAI 8557, pp. 214–227, 2014. © Springer International Publishing Switzerland 2014.
[18] Thomas, M., "A Review paper on BIG Data", International Research Journal of Engineering and Technology (IRJET), e-ISSN: 2395 -0056, Volume: 02 Issue: 09 | Dec-2015.
[19] K.Arun and L.Jabasheela, "Big Data: Review, Classification and Analysis:Survey", International Journal of Innovative Research in Information Security (IJIRIS) ISSN: 2349-7017(O),Volume 1 Issue 3 (September 2014),ISSN: 2349-7009(P).
[20] Garlasu, Sandulescu, D., *et al.,* "A Big Data implementation based on Grid Computing", 17-19 Jan. 2013.
[21] Kaur, G. and Kaur, M., "REVIEW PAPER ON BIG DATA USING HADOOP", International Journal of Computer Engineering & Technology (IJCET), Volume 6, Issue 12, pp. 65-71, Dec 2015 .
[22] http://www.domo.com/blog/2014/04/data-never- sleeps-2-0/
[23] http://www.slideshare.net/anric/140228bigdata-volumevelocityvarietyvaracityvalue140228064400phpapp02
[24] https://www.yanoresearch.com/press/press.php/001191
[25] http://blogs.intel.com/policy/2013/10/09/a-vision-for-big-data/