An overview of decadal progress in Big Data

Harsh Thakkar[1], *Research Scholar* and **Prasenjit Majumder**[2], *Assistant Professor*
DA-IICT, Gandhinagar, Gujarat, India
[1]harsh9t@gmail.com
[2]prasenjit.majumder@gmail.com

***Abstract*—Big data is a term for massive data sets having large, more varied and complex structure with the difficulties of storing, analyzing and visualizing for further processes or results. The process of research into massive amounts of data to reveal hidden patterns and secret correlations named as big data analytics. These useful informations for companies or organizations with the help of gaining richer and deeper insights and getting an advantage over the competition. For this reason, big data implementations need to be analyzed and executed as accurately as possible. This paper presents an overview of big data's content, scope, samples, methods, advantages and challenges and discusses privacy concern on it.**

***Index Terms*—Big Data, hadoop, MapReduce, Big Data challenges, Parallel processing.**

I. Introduction

THIS world is witnessing data growth as never before. Increasing amounts of data are streaming into contemporary organizations as a result of the rapidly growing quantity of data being generated not only by the organizations themselves but also in the organizations business environments by both their stakeholders and other entities operating there. Thus, it is in this context that such expressions as "a data-centric world" have become more and more common [1].

As of 2012, limits on the size of data sets that are feasible to process in a reasonable amount of time were on the order of exabytes of data [2]. Scientists regularly encounter limitations due to large data sets in many areas, including meteorology, genomics, [3] connectomics, complex physics simulations, [4] and biological and environmental research [5]. The limitations also affect Internet search, finance and business informatics. Data sets grow in size in part because they are increasingly being gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies (remote sensing), software logs, cameras, microphones, radio-frequency identification readers, and wireless sensor networks [6], [7]. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; [8] as of 2012, every day 2.5 exabytes (1 exabyte = 10^{18} bytes) of data were created.[15] The challenge for large enterprises is determining who should own big data initiatives that straddle the entire organization [9]. In 2013 the total created data is estimated to 4 zettabytes (1 zettabyte = 10^{21} bytes)

Modern e-Science infrastructures allow targeting new large scale problems whose solution was not possible before, e.g. genome, climate, global warming. e-Science typically produces a huge amount of data that need to be supported by a new type of e-Infrastructure capable to store, distribute, process, preserve, and curate these data [10], [11]: this is referred as the new infrastructures as Scientific Data e-Infrastructure (SDI).

Within the next decade, number of information will increase by 50 times however number of information technology specialists who keep up with all that data will increase by 1.5 times. [30].

The article is worded as follows: Section II presents big data characteristics, architecture, and techniques in brief. In Section III, the important success stories of big data analytics are are reviewed. Section IV presents potential barriers, challenges and obstacles of big data. Section V concludes the work.

II. Big Data

The term "Big data" is used for large data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time. Big data sizes are a constantly moving target currently ranging from a few dozen terabytes to many petabytes of data in a single data set [12]. Difficulties include capture, storage, search, sharing, analytics and visualizing. Typical examples of big data found in current scenario includes web logs, RFID generated data, sensor networks, satellite and geo-spatial data, social data from social networks, Internet text and documents, Internet search indexing, call detail records, astronomy, atmospheric science, genomics, biogeochemical, biological, and other complex and/or interdisciplinary scientific research, military surveillance, medical records, photography archives, video archives, and large-scale eCommerce. Big Data impacts include Walmart handles more than 1 million customer transactions every hour, which is imported into databases estimated to contain more than 2.5 petabytes of data - the equivalent of 167 times the information contained in all the books in the US Library of Congress, Facebook handles 40 billion photos from its user base and so on.

1) Big Data characteristics: Big data can be defined, called the five "V's" of big data [31], with the following properties associated with it:

- **Variety**:
  Data being produced is not of single category as it not only includes the traditional data but also the semi structured data from various resources like web Pages, Web Log Files, social media sites, e-mail, documents, sensor devices data both from active passive devices. All this data is totally different consisting of raw, structured,

semi structured and even unstructured data which is difficult to be handled by the existing traditional analytic systems.

- **Velocity**:
  Velocity in Big data is a concept which deals with the speed of the data coming from various sources. This characteristic is not being limited to the speed of incoming data but also speed at which the data flows. For example the data from the sensor devices would be constantly moving to the database store and this amount wont be small enough. Thus our traditional systems are not capable enough on performing the analytics on the data which is constantly in motion.
- **Volume**:
  The Big word in Big data itself defines the volume. At present the data existing is in petabytes and is supposed to increase to zettabytes in nearby future. The social networking sites existing are themselves producing data in order of terabytes everyday and this amount of data is definitely difficult to be handled using the existing traditional systems.
- **Value**:
  User can run certain queries against the data stored and thus can deduct important results from the filtered data obtained and can also rank it according to the dimensions they require. These reports help these people to find the business trends according to which they can change their strategies.
  As the data stored by different organizations is being used by them for data analytics. It will produce a kind of gap inbetween the Business leaders and the IT professionals the main concern of business leaders would be to just adding value to their business and getting more and more profit unlike the IT leaders who would have to concern with the technicalities of the storage and processing.
- **Variability**:
  Variability considers the inconsistencies of the data flow. Data loads become challenging to be maintained especially with the increase in usage of the social media which generally causes peak in data loads with certain events occurring.

A. Big Data architecture

The big data architecture consists of big data framework which supports distributed computing and the programming model for processing the large datasets. The system is explained through the following components.

1) **Hadoop [14]** - The framework
   The Apache Hadoop project develops open-source software for reliable, scalable, distributed computing. The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using a simple programming model. It enables applications to work with thousands of computational independent computers and petabytes of data. Hadoop was derived from Google's MapReduce and Google File System (GFS).
   - **Hadoop Distributed File System** [14] a distributed file file system for efficient storage and retrieval of large datasets
     The Hadoop Distributed File System (HDFS) is a distributed file system providing fault tolerance and designed to run on commodity hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. Hadoop provides a distributed filesystem (HDFS) that can store data across thousands of servers, and a means of running work (Map/Reduce jobs) across those machines, running the work near the data. HDFS has master/slave architecture. Large data is automatically split into chunks which are managed by different nodes in the hadoop cluster [31].

     Figure 1 explains the steps involved in a Hadoop Distributed File System. The Big Data that is pooled from all the variety of source, e.g. textual data, graphic data, geographic data, data from sensors, cctv footages, etc., is unstructured in nature. This kind data cannot be used directly. The Master node chops of chunks of this data and distributes these chunks to various slave nodes for the of access purpose. The master node stores the meta-data, while the slave nodes (each capable of storing and processing data) store the chunks. A hdfs has one master/name node and many slave/data nodes. When the query arrives to hdfs, it is sent to the name node. The name node processes the query and dispatches jobs to slave nodes by "Map" process. The "Map" and "Reduce" steps are elaborated in MapReduce section. The intermediate steps are shown in 1.

2) **MapReduce [[15]]** MapReduce is a software framework introduced by Google in 2004 to support distributed computing on large data sets on clusters of computers . MapReduce is a programming model for processing and generating large data sets. Users specify a map function that processes a key/value pair to generate a set of intermediate key/value pairs and a reduce function that merges all intermediate values associated with the same intermediate key [16].
   As the name suggests **Map**:The master node takes the input, partitions it up into smaller sub-problems, and distributes them to worker nodes. A worker node may do this again in turn, leading to a multi-level tree structure. The worker node processes the smaller problem, and passes the answer back to its master node. Map takes one pair of data with a type in one data domain, and returns a list of pairs in a different domain [31]:

   Map $(k_1, v_1) \rightarrow list\_of\_pairs(k_n, v_n)$

   similarly, **Reduce**: The master node then collects the

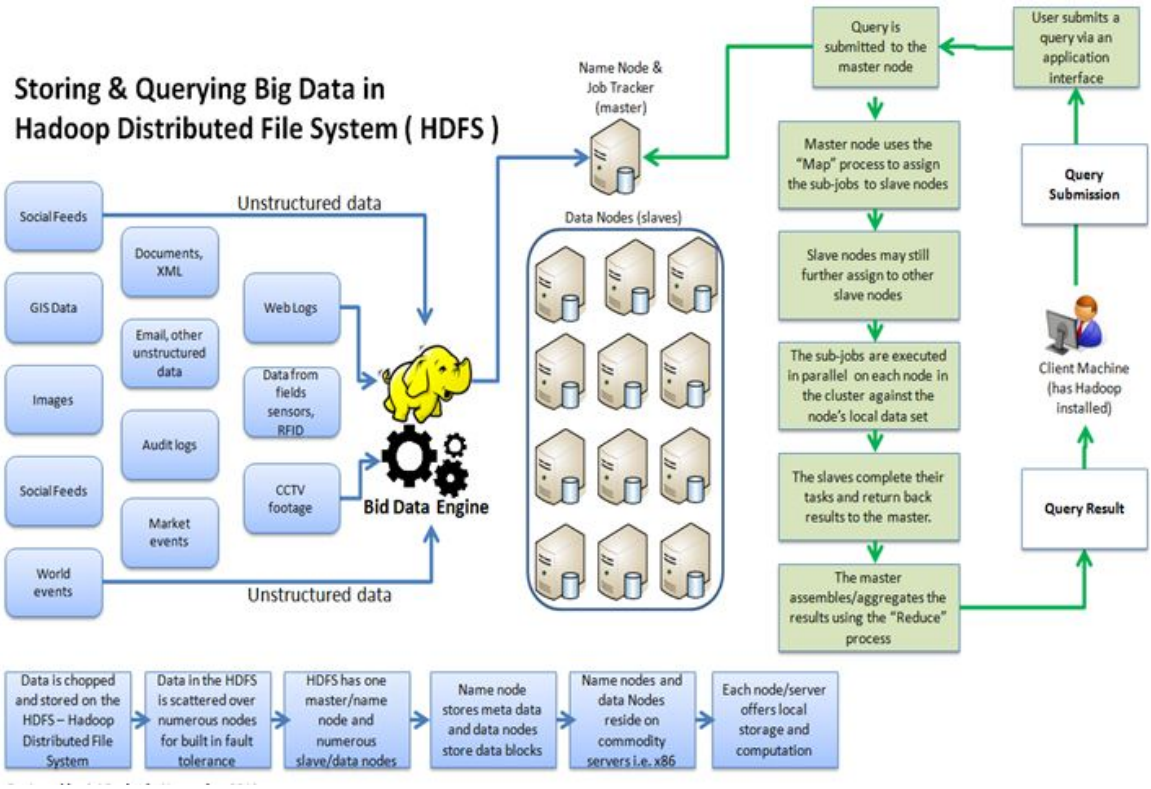

Fig. 1. The Hadoop Distributed File System (HDFS) workflow process architecture diagram [29].

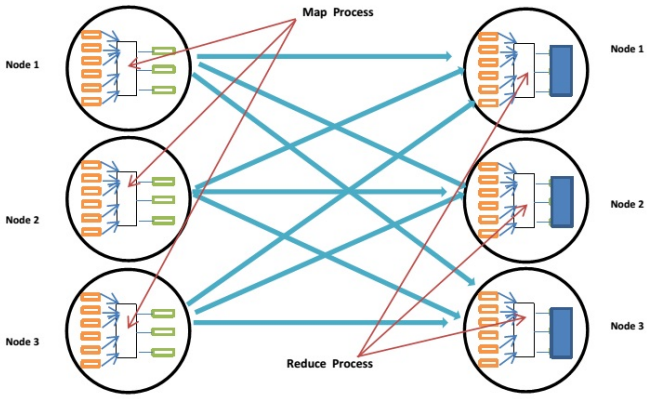

Fig. 2. The Map-Reduce process in Hadoop system.

answers to all the sub-problems and combines them in some way to form the output the answer to the problem it was originally trying to solve. The Reduce function, 2 is then applied in parallel to each group, which in turn produces a collection of values in the same domain [31]:

$$\text{Reduce } (k_n, list(v_n)) \rightarrow list(v_m)$$

B. Big Data techniques

Big data requires exceptional technologies to efficiently process large quantities of data within tolerable elapsed times. Technologies being applied to big data include massively parallel processing (MPP) databases, data mining grids, distributed file systems, distributed databases, cloud computing platforms, the Internet, and scalable storage systems. Real or near-real time information delivery is one of the defining characteristics of Big Data Analytics. Latency is therefore avoided whenever and wherever possible. A wide variety of techniques and technologies has been developed and adapted to aggregate, manipulate, analyze, and visualize big data [13]. These techniques and technologies draw from several fields including statistics, computer science, applied mathematics, and economics. This means that an organization that intends to derive value from big data has to adopt a flexible, multi-disciplinary approach.

III. Big Data success stories

Examples in the literature are available in are astronomy, atmospheric science, genomics, biogeochemical, biological science and research, life sciences, medical records, scientific research, government, natural disaster and resource management, private sector, military surveillance, private sector, 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 [17]. Organizations in any industry have big data can benefit from its careful analysis to gain insights and depths to solve real problems [18].

McKinsey Global Institute specified the potential of big data in five main topics [19], [30]–[32]:

1) Healthcare: clinical decision support systems, individual analytics applied for patient profile, personalized

medicine, performance based pricing for personnel, analyze disease patterns, improve public health
2) Public sector: creating transparency by accessible related data, discover needs, improve performance, customize actions for suitable products and services, decision making with automated systems to decrease risks, innovating new products and services
3) Retail: in store behavior analysis, variety and price optimization, product placement design, improve performance, labor inputs optimization, distribution and logistics optimization, web based markets
4) Manufacturing: improved demand forecasting, supply chain planning, sales support, developed production operations, web search based applications
5) Personal location data: smart routing, geo-targeted advertising or emergency response, urban planning, new business models

Web provides kind of opportunities for big data too. For example; social network analysis such as understanding user intelligence for more targeted advertising, marketing campaigns and capacity planning, customer behavior and buying patterns also sentiment analytics. According to these inferences firms optimization their content and recommendation engine [20]. Some companies such as Google and Amazon publishing articles related to their work. Inspired by the writings published, developers are developing similar technologies as open source software such as Lucene, Solr, Hadoop and HBase. Facebook, Twitter and LinkedIn are going a step further thereby publishing open source projects for big data like Cassandra, Hive, Pig, Voldemort, Storm, IndexTank.

In 2012, Obama regime announced big data initiatives of more than $200 million in research and development investments for National Science Foundation, National Institutes of Health, Department of Defense, Department of Energy and United States Geological Survey. The investments were launchedto take a step forward instruments and methods for access, organize and collect findings from vast volumes of digital data [21].

Big Data is being put in practice by a major number of industries and/or organizations [32]. United States is the current largest consumer of Big Data. Various U.S. big data consuming organisations can be classified as follows:

1) **Private sector industries**

- Amazon.com handles millions of back-end operations every day, as well as queries from more than half a million third-party sellers. The core technology that keeps Amazon running is Linux-based and as of 2005 they had the worlds three largest Linux databases, with capacities of 7.8 TB, 18.5 TB, and 24.7 TB
- Walmart is estimated to store about more than 2.5 petabytes of data in order to handle about more than 1 million customer transactions every hour.
- FICO Falcon Credit Card Fraud Detection System protects 2.1 billion active accounts world-wide

2) **Public sector industries**

- The Obama administration project is a big initiative where a Government is trying to find the uses of the big data which eases their tasks somehow and thus reducing the problems faced. It includes 84 different Big data programs which are a part of 6 different departments.
- The Community Comprehensive National Cyber Security initiated a data center, Utah Data Center (United States NSA and Director of National Intelligence initiative) which stores data in scale of yottabytes. Its main task is to provide cyber security [32].

3) **Science & Technological organizations**

- The Large Hadron Collider (LHC) is the world's largest and highest-energy particle accelerator with the aim of allowing physicists to test the predictions of different theories of particle physics and high-energy physics. The data flow in experiments consists of 25 petabytes (as of 2012) before replication and reaches upto 200 petabytes after replication.
- The Sloan Digital Sky Survey is a multi-filter imaging and spectroscopic redshift survey using a 2.5-m wide-angle optical telescope at Apache Point Observatory in New Mexico, United States. It is Continuing at a rate of about 200 GB per night and has more than 140 terabytes of information.

IV. Barriers in Big Data

In paper [22] the issues and challenges in Big data are discussed as the authors begin a collaborative research program into methodologies for Big data analysis and design. In paper [23] the author discusses about the traditional databases and the databases required with Big data concluding that the databases dont solve all aspects of the Big data problem and the machine learning algorithms need to be more robust and easier for unsophisticated users to apply. There is the need to develop a data management ecosystem around these algorithms so that users can manage and evolve their data, enforce consistency properties over it and browse, visualize and understand their algorithm results. In paper [24] architectural considerations for Big data are discussed concluding that despite the different architectures and design decisions, the analytics systems aim for Scale-out, Elasticity and High availability. In paper [25] all the concepts of Big data along with the available market solutions used to handle and explore the unstructured large data are discussed. The observations and the results showed that analytics has become an important part for adding value for the social business. This paper [26] proposes the Scientific Data Infrastructure (SDI) generic architecture model. This model provides a basis for building interoperable data with the help of available modern technologies and the best practices. The authors have shown that the models proposed can be easily implemented with the use of cloud based infrastructure services provisioning model. In paper [27] the author investigates the difference in Big data applications and how they are different from the traditional methods of analytics existing from a long time. In paper [28]

authors have done analysis on Flickr, Locr, Facebook and Google+ social media sites. Based on this analysis they have discussed the privacy implications and also geo-tagged social media; an emerging trend in social media sites. The proposed concept in this paper helps users to get informed about the data relevant to them in such large social Big data.

In order towards providing a better discrimination between the various factors repelling scientists from harnessing the full benefits of big data [32], we classify them in the following categories:

- IT security & Privacy:
  It is the most important issue with Big data which is sensitive and includes conceptual, technical as well as legal significance.
  - The personal information of a person when combined with external large data sets leads to the inference of new facts about that person and its possible that these kinds of facts about the person are secretive and the person might not want the Data Owner to know or any person to know about them.
  - Information regarding the users (people) is collected and used in order to add value to the business of the organization. This is done by creating insights in their lives which they are unaware of.
  - Another important consequence arising would be Social stratification where a literate person would be taking advantages of the Big data predictive analysis and on the other hand underprivileged will be easily identified and treated worse.
  - Big Data used by law enforcement will increase the chances of certain tagged people to suffer from adverse consequences without the ability to fight back or even having knowledge that they are being discriminated.
- *Data access, storage and processing* [32] :
  The storage available is not enough for storing the large amount of data which is being produced by almost everything: Social Media sites are themselves a great contributor along with the sensor devices etc.
  Because of the rigorous demands of the Big data on networks, storage and servers outsourcing the data to cloud may seem an option. Uploading this large amount of data in cloud doesnt solve the problem. Since Big data insights require getting all the data collected and then linking it in a way to extract important information. Terabytes of data will take large amount of time to get uploaded in cloud and moreover this data is changing so rapidly which will make this data hard to be uploaded in real time. At the same time, the cloud's distributed nature is also problematic for Big data analysis. Thus the cloud issues with Big Data can be categorized into Capacity and Performance issues.
  The transportation of data from storage point to processing point can be avoided in two ways. One is to process in the storage place only and results can be transferred or transport only that data to computation which is important. But both these methods would require integrity and provenance of data to be maintained. Processing of such large amount of data also takes large amount of time. To find suitable elements whole of data Set needs to be Scanned which is somewhat not possible .Thus Building up indexes right in the beginning while collecting and storing the data is a good practice and reduces processing time considerably.
  Moreover, by keeping data in one place, it occurs a target for attackers to sabotage the organization [30]–[32]. It required that big data stores are rightly controlled. To ensure authentication a cryptographically secure communication framework has to be implemented. Controls should be using principle of reduced privileges, especially for access rights, except for an administrator who have permission data to physical access.
- *Man-power expertise requirement*:
  Since Big data is at its youth and an emerging technology so it needs to attract organizations and youth with diverse new skill sets. There is huge demand of a skilled tasforce. These skills should not be limited to technical ones but also should extend to research, analytical, interpretive and creative ones. These skills need to be developed in individuals hence requires training programs to be held by the organizations. Moreover the Universities need to introduce curriculum on Big data to produce skilled employees in this expertise.
- *Technical & infrastructure challenges*:
  - **Quality of Data**: Collection of huge amount of data and its storage comes at a cost. More data if used for decision making or for predictive analysis in business will definitely lead to better results. Business Leaders will always want more and more data storage whereas the IT Leaders will take all technical aspects in mind before storing all the data. Big data basically focuses on quality data storage rather than having very large irrelevant data so that better results and conclusions can be drawn.
    This further leads to various questions like how it can be ensured that which data is relevant, how much data would be enough for decision making and whether the stored data is accurate or not to draw conclusions from it etc.
  - **Scalability**: The processor technology has changed in recent years. The clock speeds have largely stalled and processors are being built with more number of cores instead. Previously data processing systems had to worry about parallelism across nodes in a cluster but now the concern has shifted to parallelism within a single node. In past the techniques which were used to do parallel data processing across data nodes aren't capable of handling intra-node parallelism. This is because of the fact that many more hardware resources such as cache and processor memory channels are shared across a core in a single node. The scalability issue of Big data has lead towards cloud computing, which now aggregates

multiple disparate workloads with varying performance goals into very large clusters. This requires a high level of sharing of resources which is expensive and also brings with it various challenges like how to run and execute various jobs so that we can meet the goal of each workload cost effectively. It also requires dealing with the system failures in an efficient manner which occurs more frequently if operating on large clusters. These factors combined put the concern on how to express the programs, even complex machine learning tasks.

There has been a huge shift in the technologies being used. Hard Disk Drives (HDD) are being replaced by the solid state Drives and Phase Change technology which are not having the same performance between sequential and random data transfer. Thus what kind of storage devices are to be used is again a big question for data storage.

– **Fault tolerance**: With the incoming of new technologies like Cloud computing and Big data it is always intended that whenever the failure occurs the damage done should be within acceptable threshold rather than beginning the whole task from the scratch. Fault-tolerant computing is extremely hard, involving intricate algorithms. It is simply not possible to devise absolutely foolproof, 100% reliable fault tolerant machines or software. Thus the main task is to reduce the probability of failure to an "acceptable" level [**?**]. Unfortunately, the more we strive to reduce this probability, the higher the cost.

Two methods which seem to increase the fault tolerance in Big data are as: First is to divide the whole computation being done into tasks and assign these tasks to different nodes for computation. One node is assigned the work of observing that these nodes are working properly. If something happens that particular task is restarted.

But sometimes it's quite possible that that the whole computation can't be divided into such independent tasks. There could be some tasks which might be recursive in nature and the input of the previous task is the input to the next computation. Thus restarting the whole computation becomes cumbersome process. This can be avoided by applying Checkpoints which keeps the state of the system at certain intervals of the time. In case of any failure, the computation can restart from last checkpoint maintained.

The Intel IT Center has also recognised the obstacles of big data as: security concerns, capital/operational expenses, increased network bottlenecks, shortage of skilled data science professionals, unmanageable data rate, data replication capabilities, lack of compression capabilities, greater network latency and insufficient CPU power [18].

In spite of potential barriers, challenges and obstacles of big data, it has great importance today and in the future.

V. Conclusion

In this article, an overview of big data's content, scope, samples, methods, advantages and challenges and discusses privacy concern have been reviewed. The results have shown that even if available data, tools and techniques available in the literature, there are many points to be considered, discussed, improved, developed, analyzed, etc. Besides, the critical issue of privacy and security of the big data is the big issue will be discussed more in future.

Although this paper clearly has not resolved the entire subject about this substantial topic, hopefully it has provided some useful discussion and a framework for researchers.

References

[1] K. Little., "Big Data Legal Rights and Obligations", http://www.kemplittle.com/Publications/WhitePapers/Big%20Data %20-%20Legal%20Rights%20and%20Obligations%202013.pdf, Jan. 2013.

[2] Francis, Matthew., "Future telescope array drives development of exabyte processing", from wikipedia.org, retrieved Oct. 2012.

[3] "Community cleverness required", "Community cleverness required". Nature Vol-455, doi:10.1038/455001a, Sept. 2008.

[4] "Sandia sees data management challenges spiral". HPC Projects. 4 August 2009.

[5] Reichman, O.J. Jones, M.B.; Schildhauer, M.P., "Challenges and Opportunities of Open Data in Ecology", Science 331, 2011.

[6] Hellerstein, Joe (9 November 2008). "Parallel Programming in the Age of Big Data". Gigaom Blog.

[7] Segaran, Toby; Hammerbacher, Jeff. Beautiful Data: The Stories Behind Elegant Data Solutions. O'Reilly Media. p. 257. ISBN 978-0-596-15711-1., 2009.

[8] "IBM What is big data? Bringing big data to the enterprise". from "www.ibm.com/big-data/us/en/"

[9] Oracle and FSN, "Mastering Big Data: CFO Strategies to Transform Insight into Opportunity", Dec. 2012.

[10] Global Research Data Infrastructures: Towards a 10-year vision for global research data infrastructures. Final Roadmap, Mar. 2012.

[11] Riding the wave: How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data. Oct. 2010.

[12] C. Tankard, "Big Data Security", Network Security Newsletter, Elsevier, ISSN 1353-4858, July 2012.

[13] McKinsey Global Institute, 2011, Big Data: The next frontier for innovation, competition, and productivity, from "www.mckinsey.com/ /media/McKinsey/dotcom/Insights 0Innovation/Big

[14] Apache Software Foundation. Official apache hadoop website, http://hadoop.apache.org/, Aug, 2012.

[15] The Hadoop Architecture and Design, from: http://hadoop.apache.org/common/docs/r0.16.4/hdfs_design.html, Aug, 2012

[16] Hung-Chih Yang, Ali Dasdan, Ruey-Lung Hsiao, and D. Stott Parker from Yahoo and UCLA, "Map-ReduceMerge: Simplified Data Processing on Large Clusters", paper published in Proc. of ACM SIGMOD, pp. 1029 1040, 2007.

[17] http://en.wikipedia.org/wiki/Big_data, Mar. 2013.

[18] Intel IT Center, "Planning Guide: Getting Started with Hadoop", Steps IT Managers Can Take to Move Forward with Big Data Analytics, June 2012.

[19] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh and A.H. Byers, "Big data: The next frontier for innovation, competition, and productivity", McKinsey Global Institute, 2011.

[20] A. Vailaya, "Whats All the Buzz Around Big Data?", IEEE Women in Engineering Magazine, pp. 24-31, Dec. 2012.

[21] R. Weiss and L.J. Zgorski, "Obama Administration Unveils Big Data Initiative:Announces $200 Million in new R&D Investments", Office of Science and Technology Policy Executive Office of the President, March 2012.

[22] Stephen Kaisler, Frank Armour, J. Alberto Espinosa, William Money, "Big Data: Issues and Challenges Moving Forward", IEEE, 46th Hawaii International Conference on System Sciences, 2013.

[23] Sam Madden, "From Databases to Big Data", IEEE, Internet Computing, May-June 2012.
[24] Kapil Bakshi, "Considerations for Big Data: Architecture and Approach", IEEE , Aerospace Conference,2012
[25] Sachchidanand Singh, Nirmala Singh, "Big Data Analytics", IEEE, International Conference on Communication, Information & Computing Technology (ICCICT),Oct. 19-20, 2012.
[26] Yuri Demchenko, Zhiming Zhao, Paola Grosso, Adianto Wibisono, Cees de Laat, "Addressing Big Data Challenges for Scientific Data Infrastructure", IEEE , 4th International Conference on Cloud Computing Technology and Science, 2012.
[27] Martin Courtney, "The Larging-up of Big Data", IEEE, Engineering & Technology, September 2012.
[28] Matthew Smith, Christian Szongott, Benjamin Henne, Gabriele von Voigt, "Big Data Privacy Issues in Public Social Media", IEEE, 6th International Conference on Digital Ecosystems Technologies (DEST), 18-20 June 2012.
[29] S. Prakash., "Storing and querying big data in HDFS", from www.sethsidhhart.com., Nov. 2012.
[30] Sagiroglu, S.; Sinanc, D., "Big data: A review," Collaboration Technologies and Systems (CTS), pp.42-47, May 2013
[31] A. Patel, M. Birla, U. Nair., "Addressing big data problem using hadoop and marreduce", Nirma University Iinternational Conferene on Engineering, (NUiCONE)., DEC. 2012.
[32] Katal, A.; Wazid, M.; Goudar, R.H., "Big data: Issues, challenges, tools and Good practices," Contemporary Computing (IC3), 2013 Sixth International Conference on , vol., no., pp.404,409, 8-10 Aug. 2013.