International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-2, Issue-5,May 2016
ISSN: 2395-3470
www.ijseas.com

Research Issues and Challenges of Big Data –A Review

Samitha Sahu[1] Dr.V.Ilango[2] Dr.R.Chinnayan[3]
[1]*VTU-Research-Centre-MCA, New Horizon College of Engineering, Bangalore, India*
[2,3]*New Horizon College of Engineering, Bangalore, India*

Abstract- *Big data is the term for data sets so large and complicated that it becomes difficult to process using traditional data management tools or processing applications. This paper reveals most recent progress on big data analysis techniques and challenges. We have categorized reported efforts into four general categories. First, efforts related to classic big data technology such as components of big data, Importance of big data analysis, Steps and challenges of big data analysis and future of big data analysis are reported. Upon detailed summary and analysis, limitations of the proposed works as well as possible future research directions have been proposed.*

Keywords: Velocity, Volume, Varity, RFID, Massively Parallel Processing, Map Reduce.

I. Introduction

Big data is a term that describes the large volume of data – both structured and unstructured. But it's not the amount of data that's important. It's what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves. The concept gained momentum in the early 2000s when industry analyst Doug Laney [1] articulated the now-mainstream definition of big data as the three Vs as shown in the Figure.1 and Figure 2: *Volume:* Organizations collect data from a variety of sources, including business transactions, social media and information from sensor or machine-to-machine data. In the past, storing it would've been a problem – but new technologies (such as Hadoop) have eased the burden. 40 Zettabytes (43 Trillion Gigabytes) of data will be created by 2020. 300 Times increase from 2005. Most companies in the U.S have at least 100Tb of data. *Velocity:* Data streams in at an unprecedented speed and must be dealt with in a timely manner. RFID (radio frequency identification) tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time. NYSE (New York Stock Exchange) captures 1TB of trade information every day. The average modern car has over 100 sensors. *Variety:* Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions. We also consider two additional dimensions when it comes to big data. *Variability:* In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Is something trending in social media? Daily, seasonal and event-triggered peak data loads can be challenging to manage. Even more so with unstructured data. *Complexity:* Today's data comes from multiple sources, which makes it difficult to link, match, cleanse and transform data across systems. However, it's necessary to connect and correlate relationships, hierarchies and multiple data linkages or your data can quickly spiral out of control. Big Data includes huge volume, high velocity, and extensible variety of data. [1]The data in it will be of three types. *Structured data:* Relational data. *Semi Structured data:* XML data. *Unstructured data:* Word, PDF, Text, Media Logs.

II. Milestones of Big Data Technology

The information explosion is the rapid increase in the amount of published information or data and the effects of this abundance. As the amount (volume) of available data grows, the problem of managing the information becomes more difficult, which can lead to information overload. Already seventy years before people started thinking about information explosion. The major milestones in the history of sizing data volumes plus other "firsts" in the evolution of the idea of "big data" and observations pertaining to data or information explosion are also described in Table 1.

III. Importance of Big Data

The importance [11] of big data doesn't revolve around how much data we have, but what we do with it. We can take data from any source and analyze it to find answers that enable: Cost reductions, Time reductions, new product development and optimized offerings, Smart decision making. When we combine big data with high-powered analytics, we can accomplish business-related tasks such as: Determining root causes of failures, issues and defects in near-real time, Generating coupons at the point of sale based on the customer’s buying habits, Recalculating entire risk portfolios in minutes, Detecting fraudulent behavior before it affects your organization, Big data is really

IJSEAS

International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-2, Issue-5,May 2016
ISSN: 2395-3470
www.ijseas.com

Table: 1 Milestones of Big Data Analysis

Author	Year	Title	Description
John W Tukey	1947	*"Bit"*	Author coined the term bit which was used by Claude Shanon on his 1948 paper *"A Mathematical Theory of Communications."*
	1962	*"The Future of Data Analysis"*	The author has referred data analysis as intrinsically an *empirical science*
	1977	*"Exploratory Data Analysis"*	The author is emphasising the need on using data to suggest hypotheses to test and that Exploratory Data Analysis and Confirmatory Data Analysis should be proceeded side by side
Peter Naur	1974	*"Consise Survey of Computer Methods"*	The book in reference is a survey of contemporary data processing methods that are used in a wide range of applications, which is organised around the concept of data defined in the *IFIP "Guide to Concepts and Terms in Data Processing"*. Naur defines data science as "The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences."
Gregory Piatetsky-Sharpio	1989	*WorkShop on Knowledge Discovery in Databases*	The author organised and chaired the first Knowledge Discovery in Databases (KDD) workshop.
	1995	*Conference ACM SIGKDD*	Further a Conference on Knowledge Discovery and Data Mining (KDD) was conducted.
	1994	*Database Marketing* "Business Week"	This is a feature published about "Database Marketing" which explained about the data collected by the companies on their customers, to predict how likely they are to buy the product using the above knowledge, to craft a marketing message precisely calibrated to get to the customer. The feature was also explaining about the overwhelming experience of the companies by the sheer quantity of data to do anything useful with the information.
Usama Fayyad, Gregory Piatetsky-Shapiro,Padhraic Smyth	1996	*"From Data Mining to Knowledge Discovery in Databases"*	Established KDD as an overall process of discovering useful knowledge of data and data mining as a particular step in this process.
C F Jeff Wu	1997	Lecture	The author has inquired for the statistics to be renamed data science and statisticians to be renamed as data scientists
	1997	*The journal Data Mining and Knowledge Discovery*	The title reflects the ascendance of "data mining" as the most popular way to designate "extracting information from large databases."
Jacob Zahavi	1999	*"Mining Data for Nuggets of Knowledge"*	The author Jacob Zahavi explains the scalability issue in data mining and challenge is developing models that can do a better job analyzing data, detecting non-linear relationships and interaction between elements. The author was also pointing in developing Special data mining tools to address web-site decisions.
William S. Cleveland	2001	*"Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics."*	William S. Cleveland emphasizes on the need for computing and scientific involvement into the field of data and statistical analysis and thus pushes the need for expanding the technical expertise into the field.
Leo Breiman	2001	*"Statistical Modeling: The Two Cultures"*	The author was explaining about the two cultures used in statistical modelling one which is generated by a stochastic data model and other which uses algorithmic and treats the data mechanisms unknown

Thomas H. Davenport, Don Cohen, Al Jacobson	2005	*“Competing on Analytics,”*	The research was trying to show new form of data analytics and fact based decision making,how the companies are employing statistical and quantitative analysis as primary elements over traditional factors.
Yangyong Zhu and Yun Xiong	2009	*Introduction to Dataology and Data Science,* The Research Centre for Dataology and Data Science	Different form of natural sciences and social sciences are stated. The canter holds annual symposiums on Dataology and Data Science.
Hal Varian	2009		The author explains in his article about the ability on, extracting, visualising and, understanding and getting value from the data. He also emphasises that the managers need to be able to access and understand the data themselves.
Kirk D. Borne	2009	*“The Revolution in Astronomy Education: Data Science for the Masses*	Author proposes a Data Science education for masses by having the specialists inducted with newer data science technologies and the non-specialists educated on the information literacy, and thus form the 21st century work force
Mike Driscoll	2009	*The Three Sexy Skills of Data Geeks*	Mike Driscoll writes in “The Three Sexy Skills of Data Geeks”: “…with the Age of Data upon us, those who can model, munge, and visually communicate data—call us statisticians or data geeks—are a hot commodity.”
Nathan Yau	2009	*Rise of the Data Scientist*	Here the discussions are on the potentials and importance of the people who emerges out from the crowd in terms of their computational information design capabilities
Troy Sadkowsky	2009		Troy Sadkowsky creates the data scientists group on LinkedIn as a companion to his website, datasceintists.com (which later became datascientists.net)
Kenneth Cukier	2010	*”Data, Data Everywhere”,* The Economist	Kenneth Cukier discusses on emergence of new professional, the data scientist who combines the skills of software programmer, statistician and storyteller.
Mike Loukides	2010	*What is Data Science?*	Mike Loukides writes that the, Data scientists combine entrepreneurship with patience, the willingness to build data products incrementally, the ability to explore, and the ability to iterate over a solution. They are inherently interdisciplinary
Hilary Mason , Chris Wiggins	2010	*A Taxonomy of Data Science*	The author proposes one possible taxonomy of what a data scientist does, in chronological order of Obtain, Scrub, Explore, Model, and iNterpret.The author also describes data science as a blend of hakers,artsetc, so it requires creative decisions and open –mindness in a scientific context
Drew Conway	2010	*The Data Science Venn Diagram*	Rather enumerating texts and tutorials the author proposes Data Science Venn Diagram… hacking skills, math and stats knowledge, and substantive expertise”.
Pete Warden	2011	*“Why the term ‘data science’ is flawed but useful”*	Mr Warden Writes about the nature of work beyond the traditional categories that dominate the corporate and institutional world, handling everything from finding the data, processing it at scale, visualizing it and writing it up as a story. rather than the traditional scientist’s approach of choosing the problem first and then finding data to shed light on it.

David Smith	2011	*'Data Science': What's in a name?*	Author tries to highlight general adoption of the term Data Science and Data Scientist. He comments well saying that "'Data Science' better describes what we actually do: a combination of computer hacking, data analysis, and problem solving."
Matthew J. Graham	2011	*"The Art of Data Science"*	Mr. Graham quotes "To flourish in the new data-intensive environment of 21st century science, we need to evolve new skills… We need to understand what rules [data] obeys, how it is symbolized and communicated and what its relationship to physical space and time is."
Harlan Harris	2011	*"Data Science, Moore's Law, and Moneyball"*	Author here puts his way of liking to be the idea that Data Science is defined by its practitioners, that it's a career path rather than a category of activities. Also goes on saying the prosperity of the career path.
D.J. Patil	2011	*"Building Data Science Teams"*	The author here reached to the term data scientist on the track of finding a suitable professional designation for those who use both data and science to create something new.
Tom Davenport D.J. Patil	2012	*"Data Scientist: The Sexiest Job of the 21st Century"*	Published in Harward Business Review.
Fremont Rider	1994	*The Scholar and the Future of the Research Library*	Mr Rider on his article was speculating the doubling size of American University libraries and resources needed for maintain the same.
Derek Price	1961	*Science Since Babylon*	According to Derek the amount of new journals has grown exponentially rather than linearly. It has been increasing by the factor of ten.
B. A. Marron ,P. A. D. de Maine	1967	*Automatic data compression*	The paper describes "a fully automatic and rapid three-part compressor which can be used with 'any' body of information by reducing external storage and increasing the rate of information through computer.
Arthur Miller	1971	*The Assault on Privacy*	Arthur states Too many information handlers seem to measure a man by the number of bits of storage capacity his database will occupy.
I.A. Tjomsland	1980	*Where Do We Go From Here*	The author on his talk was emphasizing the penalties on discarding the useful data by retaining out-dated data to the space available.
Ithiel de Sola Pool	1983	*Tracking the Flow of Information*	The author on his book mentions the yearly growth rate of words made available through the media which was 8.9% per year compared to 2.9% in the previous years. Author concludes by stating that the increasing growth rate of point to point media is much faster than broadcasting.
Hal B. Becker	1986	*Can users really absorb data at today's rates? Tomorrow's?*	Mr Becker estimates that the recoding density achieved by Gutenberg was approximately 500 symbols (characters) per cubic inch—500 time the density of clay tablets. By the year 2000, semiconductor random access memory should be storing 1.25X10^11 bytes per cubic inch.
Peter J. Denning publishes	1990	*Saving All the Bits*	Mr Denning was mentioning in his book on the difficulties of storage when it comes saving every bits of data. Monitoring, filtering and recognizing and predicting patterns without understanding the meaning requires such machines which are eventually fast enough to deal with large data streams in real time .With these machines it is possible to reduce the number of bits to be saved also by reducing the danger of losing hidden discoveries from burial in an immense database.

B.J. Truskowski	1996	*The Evolution of Storage Systems,*	According to Mr Morris and Mr Truskowshi , Digital storage became more cost effective for storing data than in paper.
Michael Cox ,David Ellsworth	1997	*Application-controlled demand paging for out-of-core visualization*	This was the first article in ACM (Association of Computing Machinery) to use the term “big data” .According to author when there are large data sets which do not fit in main memory or even in local disk the only solution will be acquiring more resources. The problem was called*problem of big data.*
Michael Lesk	1997	*How much information is there in the world?*	Mr Lest was discussing about the volume of information in the world, he concludes by saying that there may be thousand petabytes of information, by saving everything without throwing any information and in some case the information stored might be never looked back further.
John R. Masey	1998	*Big Data… and the Next Wave of Infrastress.*	John R Messey presented the paper at a USENIX meeting.
Steve Bryson, David Kenwright, Michael Cox, David Ellsworth, Robert Haimes	1999	*Visually exploring gigabyte data sets in real time*	This was the first CACM(communication of ACM) article to use the term “Big Data”. The article was discussing that powerful computers which have many fields of query are blessing and same time a curse since fast computation spew out massive amount of data. Understanding the data resulting from high-end computations is a significant endeavour.
Francis X. Diebold	2000	*`Big Data’ Dynamic Factor Models for Macroeconomic Measurement and Forecasting*	According to the author the big Data phenomenon has been forced to confront and has also been benefited to all whether physical, biological or social, Big data has caused unprecedented advancements in data recording and storage technology.
Doug Laney	2001	*3D Data Management: Controlling Data Volume, Velocity, and Variety.*	Mr Laney had defined the three dimensions of big data which later was accepted generally, but the term Volume, Velocity and Variety was not appeared in Laney’s note.
John F. Gantz, David Reinsel	2007	*The Expanding Digital Universe: A Forecast of Worldwide Information Growth through 2010*	This was the first study to estimate and forecast the amount of digital data created and replicated every year. The study estimates the increase in data from 161 Exabyte to 988 Exabyte and recently reaching to 1227 to 2837 Exabyte’s per year.
Randal E. Bryant, Randy H. Katz, Edward D. Lazowska	2008	*Big-Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science and Society*	According to the authors big Data was the biggest innovation in computing in the last decade and will transform the activities of companies, scientific researchers, medical practitioners, and nation’sdefence and intelligence operations.
Roger E. Bohn and James E	2009	*How Much Information? 2009 Report on American Consumers.*	The study was based on finding the amount of information consumed by Americans, which was about 12 gigabytes of information for the average worker or about 3 terabytes.
James Manyika,Michael Chui, Brad Brown, Jacques Bughin	2011	*Big data: The next frontier for innovation, competition, and productivity.*	The study estimates that 7.4 Exabyte’s of new data were stored by enterprises and 6.8 Exabyte’s by consumers in 2010.
Danahboyd, Kate Crawford	2012	*Critical Questions for Big Data” in Information, communication and Society.*	Danahboyd and Kate Crawford defines big Data in their paper as cultural technological and scholarly phenomenon that rests on the interplay of Technology, Analysis, Mythology.

critical to our life and its emerging as one of the most important Technologies in modern world. Follow are just few benefits which are very much known to all of us: Using the information [12] kept in the social network like Facebook, the marketing agencies are learning about the response for their campaigns, promotions, and other advertising mediums. Using the information in the social media like preferences and product perception of their consumers, product companies [3] and retail organizations are planning their production. Using the data regarding the previous medical history of patients, hospitals are providing better and quick service.

IV. Technologies for Big Data

Big data technologies are important in providing more accurate analysis, which may lead to more concrete decision-making resulting in greater operational efficiencies, cost reductions, and reduced risks for the business. To harness the power of big data, we would require an infrastructure that can manage and process huge volumes of structured and unstructured data in real-time and can protect data privacy and security. There are various technologies in the market from different vendors including Amazon, IBM, Microsoft, etc., to handle big data. While looking into the technologies that handle big data, we examine the following two classes of technology: *Operational Big Data:* This includes [13] systems like MongoDB that provide operational capabilities for real-time, interactive workloads where data is primarily captured and stored. NoSQL Big Data systems are designed to take advantage of new cloud computing architectures that have emerged over the past decade to allow massive computations to be run inexpensively and efficiently. This makes operational big data workloads much easier to manage, cheaper, and faster to implement. Some NoSQL systems can provide insights into patterns and trends based on real-time data with minimal coding and without the need for data scientists and additional infrastructure. *Analytical Big Data:* This includes systems like Massively Parallel Processing (MPP) database systems and Map Reduce [8] that provides analytical capabilities for retrospective and complex analysis that may touch most or all of the data. Map Reduce provides a new method of analyzing data that is complementary to the capabilities provided by SQL, and a system based on Map Reduce that can be scaled up from single servers to thousands of high and low end machines. These two classes of technology are complementary and frequently deployed together as shown in Table 2.

V. Challenges in Big Data Analysis

The major[4] challenges associated with big data are as follows: Heterogeneity, Scale, Timelines, Privacy, Human Collaboration, Capturing data, Storage, Searching, Sharing, Transfer, Analysis, and Presentation. When humans consume information, a great deal of heterogeneity is comfortably tolerated. In fact, the nuance and richness of natural language can provide valuable depth. However, machine analysis algorithms expect homogeneous data, and cannot understand nuance. In consequence, data must be carefully structured as a first step in (or prior to) data analysis. Of course, the first thing anyone thinks of with Big Data is its size. After all, the word “big” is there in the very name. Managing large and rapidly increasing volumes of data has been a challenging issue for many decades. In the past, this challenge was mitigated by processors getting faster, following Moore’s law, to provide us with the resources needed to cope with increasing volumes of data. But, there is a fundamental shift underway now: data volume is scaling faster than compute resources, and CPU speeds are static. The flip side of size is speed. The larger the data set to be processed, the longer it will take to analyze. The design of a system [9]that effectively deals with size is likely also to result in a system that can process a given size of data set faster. However, it is not just this speed that is usually meant when one speaks of Velocity in the context of Big Data [6]. The privacy of data is another huge concern, and one that increases in the context of Big Data. For electronic health records, there are strict laws governing what can and cannot be done. For other data, regulations, particularly in the US, are less forceful. However, there is great public fear regarding the inappropriate use of personal data, particularly through linking of data from multiple sources. Managing privacy is effectively both a technical and a sociological problem, which must be addressed jointly from both perspectives to realize the promise of big data. In today’s complex world, it often takes multiple experts from different domains to really understand what is going on. A Big Data analysis system must support input from multiple human experts, and shared exploration of results. These multiple experts may be separated in space and time when it is too expensive to assemble an entire team together in one room. The data system has to accept this distributed expert input, and support their collaboration.

VI. Future of Big Data

Analyzing big data has been on the tip of many a technologist’s [15] tongue for the past couple of years. This analysis is described as the future for enterprises looking to gain insights into business operations and find patterns between sales

Table:2. Classes of Technologies in Big Data

	Operational	Analytical
Latency	1 ms - 100 ms	1 min - 100 min
Concurrency	1000 - 100,000	1 - 10
Access Pattern	Writes and Reads	Reads
Queries	Selective	Unselective
Data Scope	Operational	Retrospective
End User	Customer	Data Scientist
Technology	NoSQL	Map Reduce, MPP Database

and marketing activity against revenue. Many organizations have used it to good effect. Camden Council uses IBM big data analytics [5] to create a database that consolidates resident's data to reduce fraud and costs, while Expedia consumes big data to better understand what its customers are buying. Open source frameworks like Hadoop [2] make the storage of data more cost effective and, with numerous analytics tools on offer, the promised big data future is here. But it is set to change. Much of the analysis of large data sets is currently a process of looking at what is happening or has happened across an organization. This data is analyzed into insightful information that highlights sales opportunities or problems in a supply or manufacturing chain. This is often used to make an organization more effective, but cloud computing, machine learning and in-memory technologies are creating the foundations for a big data future where looking forward is the objective.

VII. Conclusion

We have entered an era of Big Data. In this work, we have done in- depth reviews on recent efforts dedicated to big data analysis. We have reviewed the progresses in fundamental big data technologies such as storage and warehousing. Important aspects of big data analysis challenges and opportunities, resource management and performance optimizations are also introduced and discussed with independent viewpoints. Moreover, software-oriented studies also need to systematically explore cross-layer, cross-platform tradeoffs and optimizations. Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises. However, many technical challenges described in this paper must be addressed before this potential can be realized fully. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. These technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone. Furthermore, these challenges will require transformative solutions, and will not be addressed naturally by the next generation of industrial products. We must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data.

VII. Reference:

1. "An Enterprise Architect's Guide to Big Data Reference Architecture Overview" Oracle Enterprise Architecture White Paper | February 2015.
2. Agrawal, Divyakant, Sudipto Das, and Amr El Abbadi. "Big data and cloud computing: current state and future opportunities." Proceedings of the 14th International Conference on Extending Database Technology. ACM, 2011.http://www.edbt.org/Proceedings/2011-Uppsala/papers/edbt/a50- agrawal.pdf.
3. Bakshi, Kapil. "Considerations for big data: Architecture and approach." Aerospace Conference, 2012 IEEE.IEEE, 2012. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6187357.
4. Dai, Jinquan, et al. "Hitune: dataflow- based performance analysis for big data cloud." Proc. of the 2011 USENIXATC (2011): 87- 100. https://www.usenix.org/legacy/event/atc11/tech/final_files/Dai.pdf.
5. Ferguson, Mike. "Architecting A Big Data Platform for Analytics." A Whitepaper Prepared for IBM (2012).http://public.dhe.ibm.com/common/ssi/ecm/en/iml14333usen/IML14333USEN.PDF.
6. Fisher, Danyel, et al. "Interactions with big data analytics." interactions 19.3 (2012): 50-59.http://research.microsoft.com/pubs/163593/inteactions_big_data.pdf.
7. Gao, Wanling, et al. "Bigdatabench: a big data benchmark suite from web search engines." arXiv preprintarXiv:1307.0320 (2013), http://arxiv.org/pdf/1307.0320.pdf.
8. He, Yongqiang, et al. "RCFile: A fast and space- efficient data placement structure in MapReduce- based warehousesystems." Data Engineering (ICDE), 2011 IEEE 27th International Conference on. IEEE, 2011. http://www.cse.ohiostate.edu/hpcs/WWW/HTML/publications/papers/TR-11- 4.pdf.
9. Herodotou, Herodotos, et al. "Starfish: A Self-tuning System for Big Data Analytics." CIDR. Vol. 11.2011. http://www.cs.duke.edu/~gang/documents/CIDR11_Paper36.pdf.

10. Jeffrey Dean and Sanjay Ghemawat, “MapReduce: Simplied Data Processing on Large Clusters”.
11. Keim, Daniel, Huamin Qu, and Kwan- Liu Ma. "Big- Data Visualization." Computer Graphics and Applications, IEEE 33.4 (2013): 20- 21. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6562707.
12. Laurila, Juha K., et al. The mobile data challenge: Big data for mobile computing research. Proceedings of the Workshop on the Nokia Mobile Data Challenge, in Conjunction with the 10th International Conference on Pervasive Computing. 2012. https://research.nokia.com/files/public/MDC2012_Overview_LaurilaGaticaPerezEtAl.pdf.
13. Madden, Sam. From databases to big data. Internet Computing, IEEE 16.3 (2012): 4-6.http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6188576.
14. Orion Gebremedhin, “Big Data Processing:Past, Present and Future” Orion.Gebremedhin@Neudesic.COM.
15. Silberstein, Adam, Ashwin Machanavajjhala, and Raghu Ramakrishnan. "Feed following: the big data challenge in social applications." Databases and Social Networks. ACM, 2011.