International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307

Big Data and Its Challenges

V. Maria Antoniate Martin*[1], Dr. K. David[2], A.Vignesh[3]
[1]Research Scholar, Department of Computer Science, Research and Development Centre, Bharathiar University, Coimbatore, Tamil Nadu, India
[2]Assistant Professor, Department of Computer Science, The Rajah's College, Pudukkottai, Tamil Nadu, India
[3]Student, Department of Information Technology, St. Joseph's College, Trichy, Tamil Nadu, India

ABSTRACT

Big Data is still a maturing and evolving discipline. Big data databases and files have scaled beyond the capacities and capabilities of commercial database management systems. Structured representations become a bottleneck to efficient data storage and retrieval. Gartner has noted four major challenges (the four Vs): increasing volume of data, increasing velocity (e.g., in/out and change of data), increasing variety of data types and structures, and increasing variability of data. A fifth V: value is suggested, which is the contribution big data has to decision making. Add to these the increasing number of disciplines and problem domains where big data is having an impact and one sees an increase in the number of challenges and opportunities for big data to have a major impact on business, science, and government.

Keywords: Big Data, Data Quality, Security

I. INTRODUCTION

What is big data? So far, there is no universally accepted definition. In Wikipedia, big data is defined as "an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications". Here the physical world has a reflection in cyberspace, embodied as big data, through Internet, the Internet of Things, and other information technologies, while human society generates its big data-based mapping in cyberspace by means of mechanisms like human–computer interfaces, brain–machine interfaces, and mobile Internet [1].

A. What Is Big Data?

Big data is data sets that are so voluminous and complex that traditional data processing application software are inadequate to deal with them. Big data challenges include capturing data, data storage, data analysis,
search, sharing, transfer, visualization, querying, updating and information privacy. There are five dimensions to big data known as Volume, Variety, Velocity and the recently added Veracity and Value. The term "big data" tends to refer to the use of predictive analytics, user behaviour analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set. "There is little doubt that the quantities of data now available are indeed large, but that's not the most relevant characteristic of this new data ecosystem [2].

B. What are The 5'vs Of Big Data?

1. Volume:

Volume is the V most associated with big data because, well, volume can be big. Here it is discussed about the quantities of data that reach almost incomprehensible proportions [3].

Example:

Just think of all the emails, Twitter messages, photos, video clips and sensor data that we produce and share every second. It is not about terabytes, but zettabytes or brontobytes of data. On Facebook alone one send 10 billion messages per day,billion times and upload 350 million new pictures each and every day. If all the data generated in the world between the beginning of time and the year 2000, it is the same amount it is now generated every minute! This increasingly makes data sets too large to store and analyze using traditional database technology. With big data technology, one can now store and use these data sets with the help of distributed systems, where parts of the data is stored in different locations, connected by networks and brought together by software.

2. Velocity:

The data growth and social media explosion have changed how data is looked upon. There was a time when we used to believe that data of yesterday is recent. The matter of the fact newspapers is still following that logic. However, news channels and radios have changed how fast one receives the news. Today, people reply on social media to update them with the latest happening. On social media sometimes a few seconds old messages (a tweet, status updates etc.) is not something interests users. They often discard old messages and pay attention to recent updates. The data movement is now almost real time and the update window has reduced to fractions of the seconds. This high velocity data represent Big Data. [4]

Example:

Just think of social media messages going viral in minutes, the speed at which credit card transactions are checked for fraudulent activities or the milliseconds it takes trading systems to analyze social media networks to pick up signals that trigger decisions to buy or sell shares. Big data technology now allows us to analyze the data while it is being generated without ever putting it into databases.

3. Variety

Variety of data produced by a multitude of sources like sensors, smart devices and social media in raw, semi-structured, unstructured and rich media formats is further complicating the processing and storage of data. Finally, the Velocity aspect describes how quickly the data is retrieved, stored and processed. [5]

Example:-

In the past, the focus was on structured data that neatly fits into tables or relational databases such as financial data (for example, sales by product or region).

In fact, 80 percent of the world's data is now unstructured and therefore can't easily be put into tables or relational databases—think of photos, video sequences or social media updates. With big data technology, one can now harness differed types of data including messages, social media conversations, photos, sensor data, and video or voice recordings and bring them together with more traditional, structured data.

4. Veracity

Veracity is not just about data quality, it is about data understand ability. Data governance initiatives have little sex appeal and most data stewards already have a primary job role. Too many users are waiting for data nirvana—perfectly clean data. Veracity is all about making sure the data is accurate, which requires processes to keep the bad data from accumulating in your systems. [6]

Example:-

Twitter posts with hash tags, abbreviations, typos and colloquial speech. Big data and analytics technology now allows us to work with these types of data. The volumes often make up for the lack of quality or accuracy.

But all the volumes of fast-moving data of different variety and veracity have to be turned into value! This is why value is the one V of big data that matters the most.

5. Value

Value starts and ends with the business use case. The business must define the analytic application of the data and its potential associated value to the business. Use cases are important both to define initial "Big Data" pilot justification and to build a road map for transformation.

The most important element of the big data we call the Sage Blue Book is value. Value that includes a large volume and variety of data that is easy to access and delivers quality analytics that enables informed decisions. Providing a fair market valuation on used technology - one piece or an entire portfolio at a time. This validates the investment return on investment (ROI) and promotes future funding [7].

Big data can deliver value in almost any area of business or society:

- ✓ It helps companies to better understand and serve customers: Examples include the recommendations made by Amazon or Netflix.
- ✓ It allows companies to optimize their processes: Uber is able to predict demand, dynamically price journeys and send the closest driver to the customers.
- ✓ It improves our health care: Government agencies can now predict flu outbreaks and track them in real time and pharmaceutical companies.
- ✓ It helps us to improve security:Government and law enforcement agencies use big data to foil terrorist attacks and detect cyber crime.

It allows sport stars to boost their performance: Sensors in balls, cameras on the pitch and GPS trackers on their clothes allow athletes to analyze and improve upon what they do.

II. BIG DATA ANALYTICS

Big data analytics refers to the process of collecting, organizing and analysing large sets of data ("big data") to discover patterns and other useful information. With the help of Big Data analytics, organizations use the large amounts of data made available to them to identify patterns and extract useful information. Big Data analysis not only helps us to understand the information contained in the data but also identify the information that is most important to the organization and future decisions. The most important goal of Big Data Analytics is to enable organizations to make better decisions. Data Scientists, predictive modellers and other analytics professionals deal with huge amounts of transactional data and use Big Data Analytics to tap this data that may be untapped by other, conventional Business Intelligence programs. Big data can be analysed with the software tools commonly used as part of advanced analytics disciplines such as predictive analytics, data mining, text analytics and statistical analysis. Due to the Volume and Velocity of Big Data, data warehouses are unable to handle the processing demands posed by data sets that are being updated in real time and continually, such as the movements on social media websites. The newer technologies involved in Big Data Analytics involve Hadoop and related tools such as YARN, MapReduce, Spark, Hive and Pig as well as NoSQL databases.

A. Difference and Challenges between big data and real-time big data:-

Big data is characteristic by multi-source heterogeneous data, widely distributed, dynamic growth, and "data mode after the data" [8] [9]. In addition to having all the characteristics with big data, real-time big data has its own characteristics. Compared with the big data, when it comes to data integration real-time big data has higher requirements in data acquisition devices, data analysis tools, data security, and other aspects. The following introduces from data integration, data

analysis, data security, data management and benchmarking.

B. Data Collect

With the development of internet of things [10] and Cyber Physical System (CPS) [11], the real time of data processing requires higher and higher. Under the big data environment, numerous sensors and mobile terminals disperse in different data management system which makes data collection itself a problem. In RTDP system, its real time data collection faced makes data integration facing many challenges.

C. Extensive heterogeneity

In big data system, the data generated by mobile terminals, tablet computers, UPS and other terminals is often stored in cache, but in RTDP system it requires data synchronization which brings tremendous challenges to the wireless network transmission. When dealing with processing heterogeneity, big data system can use NoSQL technology and other new storage methods, such as Hadoop HDFS. But the real time requires low in this kind of storage technology, where the data is often stored once but read many times. However, this kind of storage technology is far from satisfying the requirement of real-time big data system that requires data synchronization. Due to extensive heterogeneity of big data, data conversion must be carried out during data integrations processing, however traditional data warehouse has obviously insufficient to meet the needs of time and scale that big data requires [12][13][14].

D. Data quality insurance

In the era of big data it is a phenomenon often appears that useful information is being submerged in a large number of useless information [15]. The data quality of Big Data has two problems: how to manage large-scale data and how to wash it. During the cleaning process, if the cleaning granularity is too small, it is easy to filter out the useful information; if the cleaning granularity is too coarse, it cannot achieve the real cleaning effect. Therefore, between the quantity and quality it requires careful consideration and weighed which is more evident in real-time big data system. On the one hand, it requires system to synchronize data in a very short time; on the other hand, it also requires the system to make a quick response to data in real time. The performance requirements of the speed of data transmission and data analysis are increasing. Moreover, the data may be filtered at a time node may become critical post processing data. Therefore, how to grasp the correlation between data and accurately determine the usefulness and effectiveness of data becomes a serious challenge.

III. GRAND CHALLENGES IN BIG DATA

There are many challenges in harnessing the potential of big data today, ranging from the design of processing systems at the lower layer to analysis means at the higher layer, as well as a series of open problems in scientific research. Among these challenges, some are caused by the characteristics of big data, some, by its current analysis models and methods, and some, by the limitations of current data processing systems. In this section, we briefly describe the major issues and challenges.

A. Data complexity

The study of **data complexity** metrics is an emergent area in the field of **data** mining and is focus on the analysis of several **data** set characteristics to extract knowledge from them. This information used to support the election of the proper classification algorithm

B. Computational complexity

Three of the key features of big data, namely, multi-sources, huge volume, and fast-changing, make it difficult for traditional computing methods (such as machine learning, information retrieval, and data mining) to effectively support the processing, analysis and computation of big data. Such computations cannot simply rely on past statistics,

analysis tools, and iterative algorithms used in traditional approaches for handling small amounts of data. New approaches will need to break away from assumptions made in traditional computations based on independent and identical distribution of data and adequate sampling for generating reliable statistics. When solving problems involving big data, we will need to re-examine and investigate its computability, computational complexity, and algorithms.

New approaches for big data computing will need to address big data-oriented, novel and highly efficient computing paradigms, provide innovative methods for processing and analyzing big data, and support value-driven applications in specified domains. New features in big data processing, such as insufficient samples, open and uncertain data relationships, and unbalanced distribution of value density, not only provide great opportunities, but also pose grand challenges, to studying the computability of big data and the development of new computing paradigms.

C. System complexity

Big data processing systems suitable for handling a diversity of data types and applications are the key to supporting scientific research of big data. For data of huge volume, complex structure, and sparse value, its processing is confronted by high computational complexity, long duty cycle, and real-time requirements. These requirements not only pose new challenges to the design of system architectures, computing frameworks, and processing systems, but also impose stringent constraints on their operational efficiency and energy consumption.

The design of system architectures, computing frameworks, processing modes, and benchmarks for highly energy-efficient big data processing platforms is the key issue to be address in system complexity. Solving these problems can lay the principles for designing, implementing, testing, and optimizing big data processing systems. Their solutions will form an important foundation for developing hardware and software system architectures with energy-optimized and efficient distributed storage and processing.

IV. CONCLUSION

Big data has made a strong impact in almost every sector and industry today. In this paper, we have briefly reviewed the grand challenges that big data brings us. We close by a few suggestions on how to make a big data project successful. It is no secret that in big data research and applications, industry is ahead of academia. The successful applications of big data in industry point to the following necessary conditions for a big data project to be successful. Firstly, there must be very clear requirements, regardless of whether they are technical, social, or economic. Secondly, to efficiently work with big data, we will need to explore and find the kernel structure or kernel data to be processed. Finding kernel data and structures, which are small enough and yet can characterize the behavior and properties of the underlying big data, is non-trivial because it is very domain-specific. Thirdly, a top-down management model should be adopted. Although a bottom-up approach may allow us to solve some niche problems, the isolated solutions often cannot be put together into a complete solution. Finally, the goal should be to solve the entire problem by an integrated solution, rather than striving for isolated successes in a few aspects.

V. REFERENCES

[1]. T. Kalil:,Big data is a big deal available at: http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal (2012).

[2]. boyd, dana; Crawford, Kate (21 September 2011). "Six Provocations for Big Data". Social Science Research Network: A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society. doi:10.2139/ssrn.1926431.

[3]. David Gewirtz for DIY-IT | April 20, 2016 -- 12:47 GMT (18:17 IST) | Topic: Big Data Analytics

[4]. October 2, 2013Pinal DaveBig Data, SQL, SQL Server, SQL Tips and Tricks

[5]. \Laney, D.: 3D data managsement: controlling data volume, velocity and variety. Appl. Deliv. Strateg. File, 949 (2001)

[6]. https://www.impactradius.com/blog/7-vs-big-data

[7]. http://www.sagese.com/dosage/volume-velocity-value

[8]. QIN Xiong-Pai, WANG Hui-Ju, DU Xiao-Yong,WANG shan . Big Data Analysis-Compitition and Symbosis of RBDMS and MapReduce J]. Journal of Software. 2012, 23(1):32-45.

[9]. Tan Xiongpai, Wang Huiju,Li Furong,et al. New Landscape of Data Management Technologies J]. journal of Software.2013, 24(2):175-197.

[10]. CHEN Hai-Ming, CUI Li, XIE Kai-Bin. A Comparative Study on Architechtures and Implementation Methodologies of internet of Things J]. Chinese Journal of Computers 2013, 36(1): 168-188.

[11]. Lee E A, Seshia S A. Introduction to embedded systems: A cyber-physical systems approach M]. Lee & Seshia, 2011.

[12]. Thusoo A, Sarma J S, Jain n,et al. Hive-A Petabyte Scale data warehouse using Hadoop C]. Proc.of ICDE 2010. Piscataway,NJ: IEEE, 2010: 996-1005.

[13]. Abouzied A, Bajada-Pawlikowski K, Huang Jiewen. Hadoop DB in action: Building real world applications C]. Proc. of SIGMOD 2010, New York: ACM, 2010: 1111-1114.

[14]. Chen Songting, Cheetah: A high performance, custom data warehouse on top of MapReduce J]. PVLDB, 2010, 3(2): 1459-1468.

[15]. Jonathan T. Overpeck, Gerald A. Meehl, Sandrine Bony, and David R. Easterling. Climate Data Challenges in the 21st Century J]. Science, 2011, 331(6018): 700-702

VI. AUTHOR DETAILS

V. Maria Antoniate Martin is a Research Scholar in Computer Science at Bharathiar University, Coimbatore, Tamil Nadu, India. He is also working as an Assistant Professor in Department of Information Technology at St. Joseph's College, Tiruchirappalli, Tamil Nadu, India. He received his Bachelor of Science degree in Computer Science from Bharathidasan University in 2003, He completed his Masters in Science in Computer Science from the same University in 2006. He also completed his Masters in Philosophy in Computer Science from the same University in 2011.He has seven years of teaching experience. He has published seven research articles in reputed International Journals. He is also the co-author of a publication in a National Conference of importance. His area of research is Data Mining.

Dr. K. David is an Assistant Professor in the Department of Computer Science at H.H. the Rajah's College, Pudukkottai, TamilNadu, 622001. He has over fifteen years of teaching experience. He has published scores of papers in peer reviewed journals of National and International repute and is currently guiding seven Ph.D., scholars. His research interests include, UML, OOAD, Knowledge Management, Web Services and Software Engineering.

A.Vignesh, is a student of M.Sc. Computer Science, St. Joseph's College, Trichy-620002. She received her Bachelor of Science degree in Computer Science from Bharathidasan University in 2016.