International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Email:editorijoer@gmail.com http://www.ijoer.in

Vol.4., Issue.2., 2016
(Mar-Apr)

REVIEW ARTICLE

ISSN: 2321-7758

INNARDS OF BIG DATA

VINTI PARMAR[1], JYOTI[2], CHANDERKANT[3]

[1]Research Scholar, Indira Gandhi University, Rewari, India
[2,3]F.L.T.M.S.B.P. Govt. College For Women, Rewari, India

ABSTRACT

Data is the very crucial part for every organization, economy, business world and individual. Big data is the term for data sets so large and complicated that it becomes individual. 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. Big data is an amalgam of large and varieties of data sets including structured data, semi structured data and unstructured data so it's beyond the capability of traditional tools to capture, store, process and analysis of big data. This paper presents the theoretical overview of big data contents, challenges, data storage technologies, big data analytics.

Keywords : Big data, Challenges, Bigdata analytics, Hadoop.

1.INTRODUCTION

Now a days there is exponential growth in generation of data as compare to past years as everything on internet is recorded. Each and every activity of user on internet is generating data. For example: When you do any surfing, searching then your all clicks and search is recorded. Even when you do any online shopping then customer behaviour, buying pattern, items viewed by customer, items discarded by customer are recorded that is each and every details whether small or large is recorded. Such captured data through proper analysis helps in predicting trends, buying patterns so that appropriate decisions can be taken, strategies can be made for development. Google, Amazon, Twitter, Facebook, they all are the first who faced exponential growth of data at internet and solutions were also developed by them for dealing with that enormous growth of data[5]. So big data is collection of extreme large data set that helps in decision making through patterns, trends revealed through its proper and accurate analysis.

BIG DATA

Big data is used to describe a massive volume of data both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques. In most enterprise scenarios the data is too big or it moves too fast or it exceeds current processing capacity In 1997, NASA scientist used the term big data because they found difficulty with data set that is so large that they even do not get fit inside memory disk etc. They called that big data problem. It's an interesting challenge to manage such data. Then some American scientist in 2008 popularized the term big data because they predict that big data analysis can enable unlocking of valuable knowledge and helps in decision making in varieties of fields like medical, science, agriculture, industries etc. Big data cannot be considered in isolation .it is combination of various data management technologies that progresses over the time. Big data is all about large volume, diversified data that is captured from various data sources and analysed at high velocity to provide valuable insights. In the past years big data size was in terabytes but today it is in petabytes and soon it will be in Exabyte (millions of terabytes) So it's necessary to understand big data from all aspects and dimensions.Big data encompasses high velocity, high volume and variety of data like structured data, semi structured and unstructured data that have potential of unlocking new sources of development, providing valuable insights and decision making .

- *Structured data* – Structured big data is in proper formatting for use. For example numbers, date, sensor data, weblog data. Structured data can be machine generated data or human generated data. Machine generated data includes weblog data, sensor data, financial data etc. Click stream data is an example of human generated data that is each and every click on website by user is recorded for prediction of some pattern, association, events.

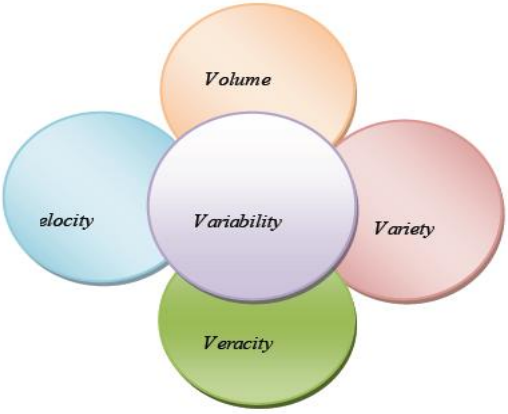

Figure 1. The five V's of Big data

- *Unstructured data* – Unstructured data is not in a specified format so can not fit properly in database.It is in same form in which it is collected and heterogenous in nature.For example: pdf,audio, ideo images emails. It is estimated that generation of unstructured data is more than structured data in an organization because analysis of unstructured data have potential of providing better accurate insights.Untstructured data can also be computer or human generated Satellite images,video,audio are machine generated data while mobile data,social media data is human generated data.
- S*emi structured data* – It is in-between structured and unstructured data that is to some extent processing is done on them. That data is not in proper organised form. For example: html, xml Since big data is collection of structured, unstructured and semi structured data so big data have enough potential to do tasks that were impossible earlier like disease management, crime management, providing new direction in business enterprises.[3]. Many areas or fields of science currently facing exponential growth in volume of data as compared to past years. It is true that big data revolutionized the research field but at same time challenges are faced in dealing with big data so there is need of emergence of new technology to utilize full potential of big data and addressing confronted challenges.As discussed by The Economist [2 panel panel] "Managed well, the data can be used to unlock new sources of economic value, provide fresh insights into science and hold governments to accounts"

Big data is characterized by the 5V that is volume, velocity and variety, veracity, variablity.

Volume: It means quantity or amount of big data that is between terabyte and petabyte. So which data is big or not can to be said accurately. It depend on size of organisation. It is the size of the data which determines the value and potential of the data under consideration and whether it can actually be considered as Big name 'Big Data' itself contains Data or not.

Variety: Big data does not mean structured data only. It's a collection or amalgam of varieties of data that is structured, unstructured and semi structured data .Today's generation of unstructured data is more than structured data and analysis of such data reveals valuable information which is not possible to get from structured data only on which business world relied in past years so there is need of emergence of new technologies for handling and managing such different types of data. Combining different types of data from different sources provide valuable insight rather than isolated data.

Velocity: It is thespeed at which data is produced and processed, The term 'velocity' in the context refers to the speed of generation of data or how fast the data is generated and processed to meet the demands and the challenges which lie ahead in the path of growth anddevelopment..

Veracity: refers to the degree in which a leader trusts the used information in order to take decision. So getting the right correlations in Big Data is very important for the business future[7]. The quality of the data being captured can vary greatly. Accuracy of analysis depends on the veracity of the source data.

Variability : This is a factor which can be a problem for those who analyze the data. This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

So it's a challenging task to deal with different varieties of data captured from various sources that are arriving continuously at high speed. For cope up with that challenging task there is need of adoption of adoption of new technology that have potential of extracting meaningful information from large and diversified data that is arriving continuously.

2. Related Work

Over the last few years many research work has undergone on Big Data that enhances the applications of big data. Bo Li and Raj Jain [1] revealed most recent progress on big data networking and big data. They categorized reported efforts into four general categories. First, efforts related to classic big data technology such as storage, Software-Defined Network, data transportation and analytics are reported. Second, important aspects of big data in cloud computing such as recourse management and performances optimization are introduced. Lastly, they introduce interesting benchmarks and progress in both search engines and mobile networking.Bernice Purcell [2] used Hadoop to process unstructured and semi-structured big data, uses the map-reduce paradigm to locate all relevant data then select only the data directly answering the query. NoSQL, MongoDB, and TerraStore process structured big data. Sameera,Siddique and Deepa Gupta[3] provided an extensive survey of Big data analytics research, while highlighting the specific concerns in Big data world.They presented a taxonomy based on the key issues in this area, and discussed the different methods to tackle these issues. Based on this work, many midmarket organizations reported a need for tools ranging from realtime processing to predictive analytics, data cleansing, and data visualization .

3. Challenges of Big Data

Challenges with big data starts with very first phase of big data analysis pipeline that is data acquisition phase. It's a challenging task to determine what data to keep, what to discard and how to efficiently store the data. Other challenges are faced in data cleaning, integration and data analysis phase of big data analysis pipeline. Few major challenges of big data are as below-

1. Understanding the Unstructured Data : It takes a lot of understanding to get data in the right shape so that it can use visualization as part of data analysis. For example, if the data comes from social media content, it needs to know who the user is in a generalsense – such as a customer using a particular set of products – and understand what it is you're trying to visualize out of the data.

2. Capturing of important data : Another challenge of big data is how to capture data which is relevant and important. Many tools have been developed till now, but they are used in specific dimension only.Hadoop, used to process unstructured and semistructured big data, uses the map-reduce paradigm to locate all relevant data then select only the data directly answering the query.

3. Storing, Availability of data : Big Data does not arise out of a vacuum: it is recorded from some data generating source. Much of this data is of no interest, and it can be filtered and compressed by orders of magnitude. One challenge is to define these filters in such a way that they do not discard useful information. When you rely on big data to feed your essential, revenue-generating 24/7 business applications, even high availability is not high enough. Your data can never go down. A certain amount of downtime is built-in to RDBMS and other NoSQL systems.

4. Scalability : With big data you want to be able to scale very rapidly and elastically. Whenever and wherever you want. Across multiple data centers and the cloud if need be. You can scale up to the heavens or shard till the cows come home with your father's relational database systems and never get there. And most NoSQL solutions like MongoDB or HBase have their own scaling limitations.

5. Privacy and security : The biggest risk that anyone familiar with big data knows is privacy concerns and security issues that emerge from such concerns. Privacy, in the big data world, indicates any or all "identifiable information blocks" that may be used to establish an individual's identity. In big data analytics, very large volumes of data involving many variables have a high probability of displaying bogus patterns or correlations, thereby establishing relationships between variables by the sheer volume of sample data, where such relationships do not exist. These types of spurious results will mislead and misguide decision makers.

Inspite of various challenges there are myriads of Opportunities in Big Data. Some major Big Data opportunities are listed below:

- Rising customer demands for smarter products, higher individualization, and mass customization.
- Better use of freely available data.
- Levelling the playing field by giving access to formerly very demanding analytical tools through commoditization.
- Developing new products and services enhanced with Big Data analytics and privacy by design, developing products adapted to European privacy standards .

The power and opportunity of big data applications used well, big data analysis can boost economic productivity, drive improved consumer and government services, thwart terrorists, and save lives. Examples include:

- Big data and the growing "Internet of Things" have made it possible to merge the industrial and information economies. Jet engines and delivery trucks can now be outfitted with sensors that monitor hundreds of data points and send automatic alerts when maintenance is needed. [2]This makes repairs smoother, reducing maintenance costs and increasing safety.
- The Centers for Medicare and Medicaid Services have begun using predictive analytics software to flag likely instances of reimbursement fraud before claims are paid. The Fraud Prevention System helps identify the highest risk health care providers for fraud, waste and abuse in real time, and has already stopped, prevented or identified $115 million in fraudulent payments—saving $3 for every $1 spent in the program's first year.14
- During the most violent years of the war in Afghanistan, the Defense Advanced Research Projects Agency (DARPA) deployed teams of data scientists and visualizers to the battlefield. In a program called Nexus 7, these teams embedded directly with military units and used their tools to help commanders solve specific operational challenges. In one area, Nexus 7 engineers fused satellite and surveillance data to visualize how traffic flowed through road networks, making it easier to locate and destroy improvised explosive devices.

4.BIG DATA ANALYTICS

Big data analytics is where advanced analytic techniques operate on big data sets. Hence, big data analytics is really about two things—big data and analytics—plus how the two have teamed up to create one of the most profound trends in business intelligence (BI) today. Big data analytics enables organization to analyze a mix of structured, semistructured and unstructured data in search of valuable business information and insights. Big data analytics refers to the process of collecting, organizing and analyzing large sets of data (big data) to discover patterns and other useful information. Big data analytics will help organization to better understand the information contained within the data and will also help identify the data that is most important to the business and future business decisions. Big data analytics basically want the knowledge that comes from analyze the data.Big data can be analyzed with software tools commonly used as a part of advanced analytics displines such as predictive analytics, data mining, text analytics and statistical analysis. Mainstream BI software and data visualization tools can also play a role in the analysis process. But the semi structured and unstructured data may not fit well in traditional data warehouse based on relational databases.Furthermore, data warehouses may not be able to handle the processing demands posted by sets of big data that need to be updated frequently or even continually. For example real time data on the performance of mobile applications.

5.Tools for Analyzing Big Data

There are five key approaches to analyzing big data and generating insight:

- **Discovery tools** are useful throughout the information lifecycle for rapid, intuitive exploration and analysis of information from any combination of structured and unstructured sources. These tools permit analysis alongside traditional BI source systems. Because there is no need for up-front modeling, users can draw new insights, come to meaningful conclusions, and make informed decisions quickly.
- **BI tools** are important for reporting, analysis and performance management, primarily with transactional data from data warehouses and production information systems. BI Tools provide comprehensive capabilities for business intelligence and performance management, including enterprise reporting, dashboards, ad-hoc analysis, scorecards, and what-if scenario analysis on an integrated, enterprise scale platform.
- **In-Database Analytics** include a variety of techniques for finding patterns and relationships in your data. Because these techniques are applied directly within the database, you eliminate data movement to and from other analytical servers, which accelerates information cycle times and reduces total cost of ownership.
- **Hadoop** is useful for pre-processing data to identity macro trends or find nuggets of information, such as out of range values. It enables businesses to unlock potential value from new data using inexpensive commodity servers. Organizations primarily use Hadoop as a precursor to advanced forms of analytics.
- **Decision Management** includes predictive modeling, business rules, and self-learning to take informed action based on the current context. This type of analysis enables individual recommendations across multiple channels, maximizing the value of every customer interaction. Oracle Advanced Analytics scores can be integrated to operationalize complex predictive analytic models and create realtime decision processes.All of these approaches have a role to play uncovering hidden relationships.

How Big Data Analytics Work

Big data is generated and collected from different sources which can be in the form of transactions, log data,events, emails, social media, free-form text, sensors, external feeds, Radio Frequency Identification (RFID) scans, Point of Sale (POS) data, geospatial and multimedia data (audio, images, videos). The sources of big data can come from different industries like financial, healthcare, public sector, media, entertainment as well as IT. Once big data (can be in single large file or multiple smaller files) is input into the big data analytical system, it will load into respective data stores, staging repositories, electronic data warehouses (EDW) or Hadoop Distributed File System (HDFS) depends on the data types like structured, semi-structured and unstructured data. Big data analytical system consists of few components which include: HDFS, Mapper and Reducer, NoSQL, data stores, content repositories as well as some processing models. HDFS, Mapper and Reducer formed the key components of the big data analytical system called Hadoop. Hadoop consists of one master node and many slave nodes (computer servers) which can range from few to thousands. When semi-structured and unstructured big data is fetched into the HDFS of the master node, it will divide the file(s) into multiple blocks and each block will be duplicated for several times (for fault tolerance and automatic recovery purpose) before distributed to different slave nodes. Once a processing job is issued by a user of the big data analytical system to the master node, the job will be translated into number of map and reduce tasks before fetching them to run and complete on all the slave nodes. Map task is to analyze and breakdown the data block further before more granular data can be sorted, shuffled and transformed. Once the map task is completed, reduce task will synthesize and aggregate them into meaningful new output. New output from all the slave nodes will be further aggregated up the master node as a more meaningful final output.The final output will export into different transactional and analytical systems like On-line Transaction Processing (OLTP), On-line Analytical Processing (OLAP) and Real-time Analytical Processing (RTAP) systems. Then the user will use these systems to search, extract and generate required information, insights or intelligence for business use. In order to apply big data analytics in business or IT, users need to know the four possible applications of big data analytics which include: recommendation, clustering, classification and frequent mining pattern [9].

6.CONCLUSIONS

We conclude that proper analysis of big data reveals valuable actionable knowledge which proves to be very useful in decision making in various areas like medical, scientific research, agricultural, organisation etc. Big data analysis give rise opportunities in designing of competitive offer packages for customers, configuring network to provide reliable services but analysis must be accurate and timely for successful decision making. review of various challenges faced with big data has been outlined in this paper and these challenges must be addresses in order to realize full potential of big data. Also all the challenges outlined are not domain specific, they are common across varieties of domain. In future research must be done to address outlined challenges.

REFERNCES

[1] Bo Li, Prof. Raj Jain "Survey of Recent Research Progress and Issues in Big Data", 2013.

[2] Bernice Purcell, "Emergence of big data technology and analytics",Journal of technology research, 2012.

[3] Sameera, Siddique and Deepa Gupta, "Big data process analytics",International Journal of Emerging Research in Management & Technology, 2014.

[4] Tawny Schlieski and Brain David Johnson, "Entertainment in the age of big data", Proceedings of the IEEE, Vol. 100, May 13th, 2012.

[5] Divyakant Agrawal, Philip Bernstein, Elisa Bertino, Susan Davidson, Umeshwas Dayal, "Challenges and Opportunities with Big Data", 2011.

[6] "Big Data A new World of Opportunities",NESSI White Paper, December 2012.

[7] Xindong Wu , Xingquan Zhu ; Gong-Qing Wu ; Wei Ding, "Data Mining with Big Data", IEEE Computer Society, Volume:26, Issue:1, P 97 – 107, 2014.

[8] Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D. & Tufano, P., " Analytics: The Real-World Use of Big Data – How Innovative Enterprises Extract Value from Uncertain Data", Executive Report, IBM Institute for Business Value, 2012.

[9] Eltabakh, M. "Special Topics in DBs Large-Scale Data Management:Advanced Analytics on Hadoop",Retrieved from web.cs.wpi.edu/~cs525/s13-MYE/lectures/6/HadoopAnalytics.pptx, 2013.

[10] LaValle "Big Data, Analytics and the Path From Insights to Value", Dec 2010.

[11] McKinsey Global Institute, "Big Data: The next frontier for innovation,competition and productivity", June 2011 VIII.