Big Data Management and Growth Enhancement

Debabrata Bhattacharya[1]

[1]M.Tech(IT), Institute of Engineering & Management, India

Abstract - Big Data means a huge volume of both structured and unstructured data that is so large it is difficult to process using traditional database and distributed databases are needed. In most enterprise the volume of data is huge or it moves too fast or it exceeds current processing capacity. Firms like Amazon, Google, LinkedIn, American Express and Face book uses big data. It is a collection of massive and complex data sets that include the huge quantities of data, social media analytics, data management capabilities, real time data etc. Big data analytics (systematic computational analysis of data) is the use of tools and processes to derive insights from large volumes of data. With the advent of the digital age, the amount of data being generated, stored and shared has risen exponentially. The huge amounts of data originating from various digital sources the importance of analytics has tremendously grown making the companies to tap the unused data that was considered useless. Industries, institutions, healthcare system, meteorological and environmental organizations and many other organizations, all of them use piles of data which are further used for creating reports in order to ensure continuity regarding the services that they have to offer. This article will be helpful to understand the processes of big data and limitations.

Key Words: Big Data, Big Data analytics, Structured data, Unstructured data, Data management.

1. INTRODUCTION

The term Big Data is very common and used almost everywhere in our daily life. The term Big Data refers to a wide range of large volume of data sets almost impossible to manage and process using traditional data management tools due to their size and their complexity within a tolerable time for its user. The processing applications of Big Data are analysis, capture, search, storage, transfer, visualization, querying and information privacy. Big Data can be seen in the finance and business where enormous amount of stock exchange, banking, online and onsite purchasing data flows through computerized systems every day and are then captured and stored for inventory monitoring, customer behavior and market behavior. It can also be seen in the life sciences where big sets of data such as genome sequencing (figuring out the order of DNA nucleotides), clinical data and patient data are analyzed and used to advance breakthroughs in science in research.

From data warehouses, web pages and blogs to audio video streams, all of these are sources of massive amounts of data. The result of this is the generation of massive amounts of data, which needs to be efficiently created, stored, shared and analyzed to extract useful information. This data has huge potential, ever increasing complexity, insecurity and risks, and irrelevance. The benefits and limitations of accessing this data are arguable in view of the fact that this analysis may involve access and analysis of medical records, social media interactions, financial data, government records and genetic sequences. The requirement of an efficient and effective analytics service, applications, programming tools and frameworks has given birth to the concept of Big Data processing and analytics. There is an immense need of architectures, platforms, tools, techniques and algorithms to handle Big Data. Big data analytics uses tools like Apache Hadoop, SAS, R – Programming, Knime, OpenRefine, Skytree, Talend and many other tools, which are more powerful than previously used rows and columns techniques. Big data analytics uses these tools to derive conclusions from both organized and unorganized data to provide insights that were previously beyond our reach. With the help of real time Big Data processing, companies can use data to enhance decision making. Big Data analytics can help companies use data to influence not only future decisions but present decisions as well. The data produced from several scientific explorations and business transactions often require tools to facilitate efficient data management, analysis, validation, visualization, and dissemination while preserving the intrinsic value of the data.

The Big Data consists of the three V's. The first V is volume which is the data must have large volume of data and it may not only refer to terabytes or petabytes but also can be measured by the number of files, records or transactions. The second V is variety where the data are in the many forms of format and can be organized in structured, semi structured or unstructured way. The last V is velocity refers at which speed the data can be generated.

The properties of Big Data are Volume, Variety, Velocity, Variability and Complexity are described below -

Volume – It is the amount of data produced by millions of sources. There has been an exponential growth in the volume of data day by day. Data includes text data, live streaming data, videos, music, large image files and many other types. Data is now stored in terms of Terabytes and even Petabytes in different enterprises. With the growth of the database, there is need to reevaluate the architecture and applications built to handle the data.

Variety - Variety of data is worth looking into in more detail, because it is important to note that there is structured and unstructured data. Data comes in all types of formats from traditional databases, text documents, videos, audios, emails, transactions etc. Data sources are extremely heterogeneous. The records comes in many layouts and of any type, it may be structured or unstructured such as text, audio, videos, log files and more. The variations are boundless and the data enters the network without having been quantified in any way.

Velocity - This means how fast the data is being produced and how fast the data needs to be processed to meet the demands and the challenges which lie ahead in the path of growth and development. The data comes at high speed. At times, one minute is too long, so Big Data is time sensitive. In some administrations data velocity is the central task. The social media posts and credit card trades done in millisecond create huge volume of data which gets stored in databases.

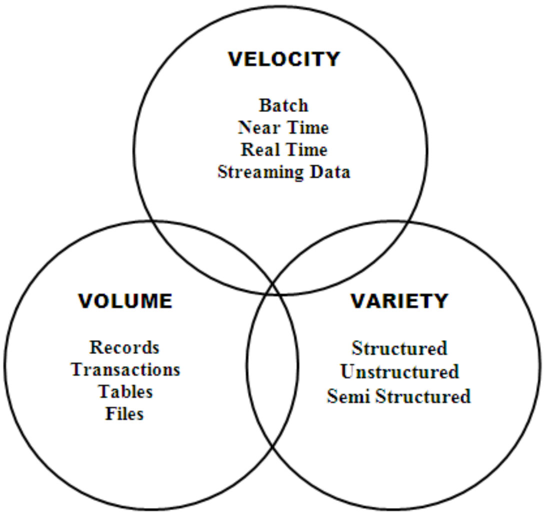

Fig - 1: Big Data 3 V's Model

We consider two other dimensions in Big Data -

Variability - In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Daily, seasonal and event triggered peak data loads is challenging to manage. Even more so with the unstructured data involved. This is a factor which can be a problem for those who are analyzing 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.
Complexity - Data comes from multiple sources and it is still an undertaking to link, match, cleanse and transform data across systems. However, it is necessary to connect and correlate relationships, hierarchies and multiple data linkages or your data can quickly spiral out of control. Complexity of the data also needs to be considered when the data is coming from multiple sources. The data must be linked, matched, cleansed and transformed into required formats before actual processing.

However, to extract and analyze the relevant information in large volume, varied and fast growing data is not an easy task. Analytics can be intended as intricate procedures running over large scale of data repositories as its main goal is that of mining useful knowledge kept in such repositories. Therefore, there are many analytical techniques are introduced in respective to gain as much as information from unmanageable large volume and varied data. Several of these techniques are association rule learning, data mining, cluster analysis, machine learning, text analytics and crowd sourcing.

Machine learning techniques have been found very effective and relevant to many real world applications in bioinformatics, network security, healthcare, banking and finance, and transportations. A major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data. For example, the U.S. Department of Homeland Security uses machine learning to identify patterns in cell phone and email traffic, as well as credit card purchases and other sources surrounding security threat. They use these patterns to try to identify future threats so they can handle them before they become large problems.

Big Data analytics refers to the process of collecting, organizing and analyzing large sets of data i.e. 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. Big data can be analyzed with the software tools commonly used as part of advanced analytics disciplines such as predictive analytics, data mining, text analytics and statistical analysis.

2. APPLICATIONS OF BIG DATA

2.1 Big Data Contributions to Education

The availability of big data as learning modes has stimulated the applications of big data technology in education. Education is one of the first places that were exposed to the idea of big data. After all, performance through school is in large part based on the data that teachers keep on throughout the school year. However, with the rise of Big Data, the term given to the ability to gather huge amounts of digital information and interact with it, schools may find themselves in a position to implement a great deal of Big Data motivated changes. Big Data in the education sector is likely to offer numerous benefits to students and educational institutions. It will revolutionize the way to manage education, in significant ways. Big Data in the education sector offers unprecedented opportunities for educators to reach out and instruct students with new ideas. It will give them a deeper understanding of student's education experience, and thereby help them evaluate the state of the education system. The overall idea of implementing Big Data within the educational system is to improve the student results. Currently, the only measurement of the performance of students is examination. However, during school life, each student generates a unique data trail. Analyzing this data trail in real time will help gain a better understanding of the individual behavior of students, and in creating an optimal learning environment for the students. With Big Data in the education sector, it is possible to monitor student actions, such as how long they take to answer a question, which sources they use for exam preparation, which questions they skip, etc. As Big Data in the education sector would help improve student results, dropout rates at schools and colleges would also reduce. Educational sectors can use predictive analytics on all the data that is collected to give them insights on future student outcomes. In fact, Big Data can also be used to monitor how students are performing in the job market after graduating from college. This would also help the future students in choosing the right college and course.

2.2 Big Data Contributions to Healthcare

The application of big data analytics in healthcare has a lot of positive and also life saving outcomes. Big data refers to the large volume of information created by the digitization of everything that gets consolidated and analyzed by specific technologies. Applied to healthcare, it will use specific health data of a population (or of a particular individual) and potentially help to prevent epidemics, cure disease, cut down costs and other facilities.

The most challenging parts for big data in healthcare are data privacy, data leakage, data security, efficient handling of large volumes of medical data, information confidentiality and security, wrong use of health data or failure to safeguard the healthcare information, and understanding unstructured clinical notes in the right context, extracting potentially useful information. Big data has a great prospective to progress healthcare management and transform healthcare industry to a high level.

Technology in recent years has made it possible not only to get data from the healthcare environment but also information from society itself (sensors, monitoring, Internet of Things (IoT) devices, social networks). The healthcare would benefit directly through the acquisition and analysis of the information generated in any kind of social environment such as social networks, forums, chats, social sensors, Internet of Things (IoT) devices, surveillance systems, virtual worlds, to name a few. These environments provide an incredible and rich amount of information that could be analyzed and applied to the benefit of public health. Combining information from informal (e.g. web based searches like Google) and syndrome surveillance and diagnostic data, for example, the next generation sequencing, can provide much earlier detection of disease outbreaks and detailed information for understanding links and transmission.

Now a day's people live longer, treatment models have changed and many of these changes are namely driven by data. Doctors want to understand as much as they can about a patient and as early in their life as possible, to pick up warning signs of serious illness as they arise and treating any disease at an early stage is far simpler and less expensive. With healthcare data analytics, prevention is better than cure and managing to draw a comprehensive picture of a patient will let insurance companies provide a good package.

Indeed, for years gathering huge amounts of data for medical use has been costly and time consuming. With today's improved technologies, it becomes easier not only to collect such data but also to convert it into relevant critical insights, which can then be used to provide better healthcare. This is the purpose of healthcare data analytics, using data driven findings to predict and solve a problem before it is too late, and also assess methods and treatments faster, keep better track of inventory, involve patients more in their own health and empower them with the tools to do so.

2.3 Big Data Contributions to Industry

In industry, Big Data enables us to better assess risks. Insurance companies are able to build more accurate profiles of their customers, and based on their whole database, they have a better idea of how probable it is for a customer to make an accident. Understanding the customer's behavior and provide them better services. In fact, by analyzing the data, they are having a deeper understanding of the customers and can therefore become more efficient in offering products and services that meet the client's needs. Big Data is used to increase productivity and to enhance supply chain management. Manufacturing companies use these analytical tools to ensure that are allocating the resources of production in an optimum manner which yields the maximum benefit.

Big data application comes from financial trading. High Frequency Trading (HFT) is an area where Big Data finds a lot of use today. Here, Big Data algorithms are used to make trading decisions. The majority of equity trading now takes place via data algorithms that increasingly take into account signals from social media networks and news websites to create buy and sell decisions in split seconds.

Computers are programmed with complex algorithms that scan markets for a set of customizable conditions and search for trading opportunities. The programs can be designed to work with no human interaction or with human interaction, depending on the needs and desires of the client. The most sophisticated of these programs are now also designed to change as markets change, rather than being hardcoded.

2.4 Big Data Contributions to Banking and Fraud detection

Big Data enables banks to get a better overview regarding their client's transactions and general behavior. This implies having Big Data insights about their spending habits and therefore we can have stronger customer segmentation. Banks now have access billions of customers' needs, and they can now use Big Data to cater to them in a more meaningful way. Cloud based analytics packages can sync in real time with your big data systems, creating actionable insight dynamically. Big Data will expand the banking industry in a way that will allow them to earn more revenue through cost reduction. And by cutting down on unnecessary costs, the banking industry can provide customers with exactly what they're looking for, instead of irrelevant information. Of course by building a stronger profile, they can tailor products and services to the needs of the client, adding value. Fraud detection has always been an important activity for banks. With Big Data, the process time of fraud detection is reduced and made much more efficient. But not only does Big Data come in handy in fraud detection, but also in risk management. Indeed, they can detect quite early risks, which will help prevent losses due to bad loans or failed investments. For monitoring financial markets through network activity monitors and natural language processors to reduce fraudulent transactions. Big data techniques are used to detect and prevent cyber attacks. Banking as a data intensive subject has been progressing continuously under the promoting influences of the era of Big Data. Exploring the advanced Big Data analytic tools like Data Mining (DM) techniques is the key for the banking sector, which aims to reveal valuable information from the overwhelming volume of data and achieve better strategic management and customer satisfaction.

There are various types of fraud, which are numerous and diverse as financial institutions and technology products. Several types of fraud are ATM and internet, transaction products credit and debit cards and checks.

3. BIGDATA CHALLENGES

Big Data is a broad term for large and complex datasets where traditional data processing applications are inadequate. The integration of this huge data sets is quite complex. There are several challenges during this integration such as analysis, capture, sharing, search, visualization, information privacy and storage. The core elements of the Big Data platform are to handle the data in new ways as compared to the traditional relational database. Heterogeneity, scale, timeliness, complexity, and privacy problems with Big Data impede progress at all phases of the pipeline that can create value from data. The problems start right away during data acquirement, when the data tsunami requires us to make decisions, currently in an ad hoc manner, about what data to maintain and what to reject, and how to store what we keep unfailingly with the right metadata. A great deal data today is not natively in structured format, for example, tweets and blogs are weakly ordered pieces of text, while images and video are structured for storage and display. Transforming content into a structured format for later analysis is a main test. The value of data explodes when it can be associated with other data. Thus data integration is a major creator of value. The majority data is directly generated in digital format today, the opportunity and the challenge both to influence the creation to facilitate later linkage and to automatically link before created data. Data analysis, organization, recovery, and modeling are other foundational challenges. Data analysis is a clear bottleneck in a lot of applications, both due to be small of scalability of the original algorithms and due to the complexity of the data that needs to be analyzed.

The Big Data challenges are -

3.1 Volume of data

The most obvious challenge associated with big data is simply storing and analyzing huge volume of information. In its digital report, IDC (International Data Corporation) estimates that the amount of information stored in the world's IT systems is doubling about every two years. By next 5 years, the total amount will be enough to fill a stack of tablets that reaches from the earth to the moon 10 times. And enterprises have responsibility or liability for about 80 percent of that information. The volume of data, especially machine generated data, is exploding, how fast that data is growing every year, with new sources of data that are emerging.

3.2 Talent gap

There's already a shortage of data scientists in the market. This includes a shortage of people who know how to labor well with large volumes of data and big data sets. Companies need the right merge of people to help make sense of the data streams that are coming into their organizations. This includes skills for applying prophetic analytics to Big Data. The new tools evolved in this sector can range from traditional relational database tools with some alternative data layouts designed to maximize access speed while reducing the storage footprints, NoSQL data management frameworks, in-memory analytics, and as well as the broad Hadoop ecosystem. The reality is that there is a lack of skills available in the market for big data technologies. The typical expert has also gained experience through tool implementation and its use as a programming model, apart from the big data management aspects.

3.3 Scalability

With big data, it's crucial to be able to scale up and down on-demand. Many organizations fail to take into account how quickly a big data project can grow and evolve. Constantly pausing a project to add additional resources cuts into time for data analysis. Big data workloads also tend to be huge, making it difficult to predict where resources should be allocated. The extent of this big data challenge varies by solution. A solution in the cloud will scale much easier and faster than an on-premises solution.

3.4 Securing big data

Security is also a big concern for organizations with big data stores. After all, some big data stores can be attractive targets for hackers or Advanced Persistent Threats (APTs). However, most organizations seem to believe that their existing data security methods are sufficient for their Big Data needs as well. In the IDG (International Data Group) survey, less than half of those surveyed (39 percent) said that they were using additional security measure for their Big Data repositories or analyses. Among those who do use additional measures, the most popular include identity and access control (59 percent), data encryption (52 percent) and data segregation (42 percent). Big Data adoption projects put security off till later stages. Big Data technologies do evolve, but their security features are still neglected, since it's hoped that security will be granted on the application level.

CONCLUSIONS

Big Data is changing the way we perceive our world. The impact big data has created and will continue to create can ripple through all facets of our life. We are living in the era of data deluge. The term Big Data had been coined to describe this age. This paper defines and characterizes the concept of Big Data. It gives a definition of this new concept and its characteristics. Here we have discussed the challenges of Big data (volume, variety, velocity, value, veracity) and various advantages and a disadvantage of these technologies. The main goal of this paper was to give a clear idea of Big Data handling techniques those handle a massive amount of data from different sources and improves overall performance of systems. The availability of Big Data, low cost commodity hardware, and new information management and analytic software has produced a unique moment in the history of data analysis. The convergence of these trends means that we have the capabilities required to analyze astonishing data sets quickly and cost-effectively. The data would be generated through a wide array of sensors that are continuously incorporating in our lives. A smart home today consists of an all encompassing architecture of devices that can interact with each other via the vast internet network. Bulbs that dim automatically aided by ambient light sensors and cars that can glide through heavy traffic using proximity sensors are examples of sensor technology advancements that we have seen over the years. Big Data is also changing things in the business world. Companies are using big data analysis to target marketing at very specific demographics. Scientific experiments and simulations can easily produce petabytes (2^{50} bytes) of data today. Building a viable solution for large and complex data is a challenge that companies in this field are continuously learning and implementing new ways to handle it. One of the biggest problems regarding Big Data is the infrastructure's high costs. Hardware equipment is very expensive for most of the companies, even if cloud solutions are available. Each Big data system requires massive processing power and stable and complex network configurations that are made by specialists. Besides hardware infrastructure, software solutions tend to have high costs if the beneficiary doesn't opt for open source software. Even if they chose open source, to configure there is still needed specialists with the required skills to do that. The downside of open source is that maintenance and support is not provided as is the case of paid software. So, all that is necessary to maintain a Big Data solution working correctly needs, in most cases, an outside maintenance team. A computer program can only do what is programmed to do, it cannot see grey areas and cannot learn or adapt to new types of information unless is programmed to handle it. Therefore, human capabilities are used to sort data with a set of tools which speed up the process. There is immense scope in Big Data and a huge scope for research and development.

REFERENCES

[1] Raghav Toshniwal, Kanishka Ghosh Dastidar and Asoke Nath, "Big Data Security Issues and Challenges", International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Issue 2, Volume 2 (February 2015).

[2] Bakshi Rohit Prasad and Sonali Agarwal, "Comparative Study of Big Data Computing and Storage Tools: A Review", International Journal of Database Theory and Application Vol.9, No.1 (2016), pp.45-66.

[3] Ms. Vibhavari Chavan and Prof. Rajesh. N. Phursule, "Survey Paper On Big Data", ISSN: 0975-9646.

[4] Senthilkumar SA, Bharatendara K Rai, Amruta A Meshram, Angappa Gunasekaran and Chandrakumarmangalam S, "Big Data in Healthcare Management: A Review of Literature" American Journal of Theoretical and Applied Business ISSN: 2469-7834 (Print); ISSN: 2469-7842 (Online).

[5] Sanskruti Patel and Atul Patel, "A BIG DATA REVOLUTION IN HEALTH CARE SECTOR: OPPORTUNITIES, CHALLENGES AND TECHNOLOGICAL ADVANCEMENTS" International Journal of Information Sciences and Techniques (IJIST).

[6] Angkoon Phinyomark, Esther Ibanez-Marcelo and Giovanni Petri, "Resting-State fMRI Functional Connectivity: Big Data Preprocessing Pipelines and Topological Data Analysis", IEEE TRANSACTIONS ON BIG DATA, VOL. 3, NO. 4, OCTOBER-DECEMBER 2017.

[7] Dr. S.Vijayarani1 and Ms. S.Sharmila, "RESEARCH IN BIG DATA – AN OVERVIEW", Informatics Engineering, an International Journal (IEIJ), Vol.4, No.3, September 2016.

[8] Dr. Puneet Goswami, "A Survey on Big Data & Privacy Preserving Publishing Techniques", Advances in Computational Sciences and Technology ISSN 0973-6107 Volume 10, Number 3 (2017) pp. 395-408.

[9] Marcos D. Assunçao, Rodrigo N. Calheiros, Silvia Bianchi, Marco A.S. Netto and Rajkumar Buyya, "Big Data computing and clouds: Trends and future directions", J Parallel Distrib. Comput. 79–80 (2015) 3–15.

[10] Dr.M.Padmavalli, "Big Data: Emerging Challenges of Big Data and Techniques for Handling",
IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 6, Ver. IV (Nov.-Dec. 2016), PP 13-18.

[11] Samiddha Mukherjee and Ravi Shaw, "Big Data – Concepts, Applications, Challenges and Future Scope", International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016 ISSN (Online) 2278-1021 ISSN (Print) 2319 5940.

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

[13] Mrs. Mereena Thomas, "A Review paper on BIG Data", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056 Volume: 02 Issue: 09 Dec-2015 p-ISSN: 2395-0072.

[14] Neha D. Patil and Dr. D. S. Bhosale, "Providing highly accurate service recommendation for semantic clustering over big data", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056 Volume: 04 Issue: 02 Feb -2017 p-ISSN: 2395-0072.

[15] Anurag Agrahari and Prof D.T.V. Dharmaji Rao, "A Review paper on Big Data: Technologies, Tools and Trends", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 10 Oct -2017 p-ISSN: 2395-0072.

[16] Mansi H Shah, Shreya P Shah and Prof. Nidhi Barot, "Big Data Concern with Web Applications: The Growth Enhancement for Free Venture Proposal", IJIRST National Conference on Latest Trends in Networking and Cyber Security March 2017.