IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

Big Data – A Pool of Opportunities and Negotiations: Forseen and Unforseen

Anand P. Lad[1], Samarth A. Shah[2]

B.E, Computer Science & Engineering Department, Faculty of Technology & Engineering,

Maharaja Sayajirao University, Vadodara, Gujarat, India[1, 2]

Abstract: Big data is the term for any collection of large and complex data which becomes difficult to process using traditional data processing applications. The challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and privacy violations. Recent times have witnessed the generation and storage of large amount of vital data by various industries which is rapidly increasing on the internet and thus the data scientists are facing a lot of challenges for maintaining a huge amount of data as the fast growing industries require the significant information for enhancing the business and for predictive analysis of the information. The question of the hour is, how to develop a high performance platform that efficiently analyzes big data and how to design an adroit algorithm for mining the useful things from big data. Facilitation of information flows and mechanisms of learning and coordination by heterogeneous individuals is the primary role of big data in cities [4]. However, processes of self-organization in cities, as well as of service improvement and expansion must rely on general principles that enforce necessary conditions for cities to operate and evolve. Such ideas are the core a developing scientific theory of cities, which is itself enabled by the growing availability of quantitative data on thousands of cities worldwide, across different geographies and levels of development. Performing computation and database operations for massive amounts of data, remotely from the data owner's enterprise is the implication of Big Data. Since a key value proposition of big data is access to data from multiple and diverse domains, a very important role shall be played by Security and Privacy for the technology and implementations.

Keywords: Big data characteristics (Four Vs), big data analytics, big data application, Connectivity between Big data with IoT and cloud, big data limitations, Security and privacy preservations in big data.

I. INTRODUCTION

Saying again, "Big Data is the new gold" (Open Data Initiative). "Big Data" is a term containing the application of techniques and methods to capture, process, analyse and visualize potentially large datasets in a reasonable timeframe not accessible to standard IT technologies. By default, the platform, tools and software used for this purpose are collectively called "Big Data technologies". The concept of "database machine" came out of the mind, in the late 1970's, which is a technology specially used for storing and analyzing data [2]. The storage and processing capacity of a single mainframe computer system became inadequate as the data volume increased. To meet the demand of the increasing data volume, people proposed "share nothing," which was a parallel database system. The "share nothing" system architecture is based on the principle of use of cluster and an the fact that every machine has its own processor, storage, and disk. The first successful commercial parallel database system was Teradata system. Such database became very popular lately. On June 2, 1986, Teradata delivered the first parallel database system with the storage capacity of 1TB to Kmart to help the large-scale retail company in North America to expand its data warehouse which was a milestone event in the history of data analysis techniques . In the late 1990s, the advantage of parallel database was widely recognized in the field of data analysis [3]. The current buzz around big data is quite lagging in timeline compared to the starting of the story of how data became big. We encountered the first attempts to quantify the growth rate in the volume of data or what has popularly been known as the "information explosion" (a term first used in 1941, according to the Oxford English Dictionary) around 70 years ago. Presently, big data and analytics are "hot" topics in both the popular and business press. Articles in publications like the New York Times, Wall Street Journal and Financial Times, as well as books like Super Crunchers [Ayers, 2007], Competing on Analytics [Davenport and Harris, 2007], and Analytics at Work [Davenport, et al., 2010] have discussed and tried to showcase the potential value of big data and analytics. Big data is a buzzword, or catch-phrase, used to describe a massive volume of both structured and unstructured data that is so huge that it is complicated to be processed using traditional database and software techniques.

Eventough big data does not refer to any specific quantity, so this term is often used while speaking about petabytes and exabytes of data. An epitome of big data may be petabytes (1,024 terabytes) or exabytes (1,024 petabytes) of data consisting of billions to trillions of records of millions of people—all from different sources (e.g. Web, sales, customer contact center, social media, mobile data and so on). The report of IDC indicates that big data had a market of about $16.1 billion in 2014. Another report of IDC forecasts that by 2017 it will grow up to $32.4 billion. The reports further showed out that the marketing of big data will be $46.34 billion and $114 billion by 2018 and

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

2019, respectively. The sharp growth of data is endorsed by rapid development of the combination of cloud computing phase and the Internet of Things (IoT). Big data oriented cloud computing makes storage management easy as various applications and data resources are spread among the users of the worldwide in a distributed manner. A huge amount of data collection by various sensor nodes and transmission of the data over the cloud network for storing and further processing is included in the model of IoT (Internet of Things).

Big data characteristics:

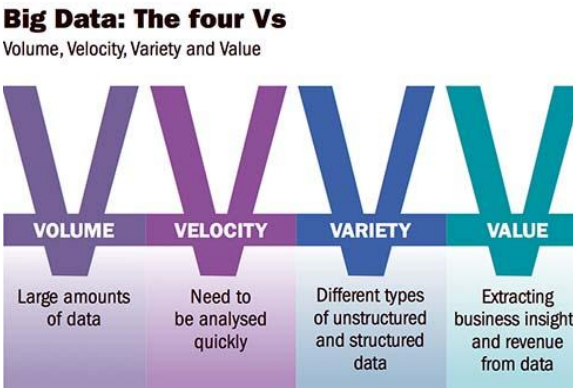

Figure1. The four Vs of Big Data

Volume: We currently see the exponential growth in the data storage as the data is now more than text data. There are videos, music and large images on our social media channels. It is very common to have Terabytes and Petabytes of the storage system for enterprises.

Variety: Today's data no longer fits into neat, easy to consume structures. New types include content, geo-spatial, hardware data points, location based, log data, machine data, metrics, mobile, physical data points, process, RFID etc. Data variety is a measure of the richness of the data representation – text, images video, audio, etc. From an analytic perspective, it is probably the biggest obstacle to effectively using large volumes of data. Incompatible data formats, non-aligned data structures, and inconsistent data semantics represents significant challenges that can lead to analytic sprawl.

Velocity: According to Gartner, velocity "means both how fast data is being produced and how fast the data must be processed to meet demand." The Velocity dimension represents the speed of processing of Big Data. In mathematical terms Dimension 2 is directly proportional to Dimension 1, i.e., the data volume must be processed at a velocity at which it flows. Financial Services Organizations such as Goldman Sachs, JP Morgan Chase, City Financial and others have coped with fast moving data to their advantage in the past. But the enormous increase in Volume of structured and unstructured data presents both opportunities and challenges for these organizations. Data Storage, the speed of processing data and the relationships between the structured and unstructured data are some of the challenges that must be addressed to deal with real-time analysis and discernment. A delay of even a few seconds could mean immeasurable financial losses when trade transactions are concerned. Velocity describes the frequency at which data is generated, captured and shared. Recent developments mean that not only consumers but also businesses generate more data in much shorter cycles.

Veracity: This refers to the uncertainty of the data available. Veracity isn't just about data quality, it's about data understandability. Veracity has an impact on the confidence data. Decision makers don't always trust the information they use to make decisions. Establishing trust in big data presents a challenge as the variety and number of sources grows. How do we define and keep useful data and discard unusable data? What data is precise and what is imprecise? The Big Data Dimension 4, Veracity deals with uncertain or imprecise data. When we consider social media such as Tweeter, Facebook, Linkedin, etc. we must consider the amount of trust that can be put in these data sources. If the data is not trustworthy, then we can perform validity checks, quality checks, amount for discrepancies and scrub it [1].However, both Velocity and Variety are detrimental to Veracity and our ability in cleansing it, presenting another huge challenge in controlling the quality of Big Data.

Big data Analytics: Data analytics is considerably difficult compared to data collection and storage. It is a serious challange to develop scalable and parallel machine learning algorithms for online analytics. Unstructured data is heterogeneous in nature and variable in nature and comes in many formats, including text, document, image, video and more. Structured data is slower than unstructured data. According to a 2011 IDC study, it will account for 91 percent of all data created in the next decade. As a new, relatively untapped source of insight, unstructured data analytics can reveal important internal relationships that were previously impossible to determine. Big data analytics is a technology-enabled strategy for gaining richer, deeper, wide and more accurate insights into customers, partners and the business and ultimately gaining competitive advantage. By processing a steady stream of real-time data, organizations can make time-sensitive decisions faster than ever and monitor emerging trends, course-correct rapidly and jump on new business opportunities faster and gently. Traditional data mining algorithms require loading of entire data in the main memory for mining, but for big data it is expensive to exchange data across various locations. Domain knowledge of applications is also very essential as data privacy and data sharing mechanisms can be varied based on the nature and requirement of the applications. Mining complex semantic relationship from Big Data improves performance of applications including search engines and recommendation systems and gives a deep insight into various social phenomena but it has become a great challenge due to the heterogeneity and huge volume of the data throughout the loop. One of the obstacles to widespread analytics adoption is a lack of understanding on how to use analytics to improve or rather withstand the business. The objects to be modeled and simulated are

DOI 10.17148/IJARCCE.2016.5160

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

complex and massive, and correspondingly the data is vast and distributed because of hetrogenity. At the same time, the modelling and simulation software solutions are expected to be simple and general, built on the solid foundations provided by a few robust computational paradigm shifts and naturally oriented towards distributed and parallel computing schemes. Hence, new methodologies and tools for data visualization and simulation are required extensively. Existing analytic frameworks are mostly transaction based which have been effectively used in business applications like customer segmentation and marketing, management of financial and accounting activities, while the perspective is being shifted towards ecosystem based analytic frameworks which mainly focus on integrated analysis of the less structured environments, compared to the isolated transaction analysis and development.

From the volume perspective, the deluge of input data is the very first thing that we need to face in hand because it may paralyze the entire data analytics. In addition, from the velocity perspective, real-time or streaming data bring up the problem of large bulk of data coming into the data analytics within a short duration but the device and system may not be able to handle these input data for the slot. This situation is similar to that of the network flow analysis for which we typically cannot mirror and analyze everything we can gather on the front. From the variety perspective, because the incoming data may use different types or have incomplete data, how to handle their amalgamation also brings up another issue for the input operators of data analytics.

This discussion of big data analytics is basically divided into input, analysis, and output for mapping the data analysis process of KDD. However, there still exist some new issues of the input and output that the data scientists need to confront but haven't been able to. A representative example in “Big data input” is that the bottleneck wills not only on the sensor or input devices, it may also appear in other places of data analytics. Similar situations also exist in the output portion. Although several measurements can be used to evaluate the performance of the frameworks, platforms, and even data mining algorithms, there still exist several new issues in the big data age, such as information fusion from different information sources or information accumulation from different times. Several studies attempted to present an efficient or effective solution from the perspective of system (e.g., framework and platform) or algorithm level. For the mining algorithm perspective, the clustering, classification, and frequent pattern mining issues play the vital role of these researches because several data analysis problems can be mapped to these essential issues that need to be addressed.

Big Data Application:

I. Targeting and Understanding Customers

This is one of the biggest and most publicized areas of big data utilization today. Here, big data is used to better understand customers and their preferences and behaviors. Companies are keen in expansion of their traditional data sets with social media data, browser logs as well as text analytics and sensor the data to get a more complete picture of their customers [11]. The big objective, in many cases, is the creation of predictive models. Using big data, Telecom firms can now better predict customer churn; the car insurance companies can understand how well their customers actually drive. The government election campaigns can also be optimized using big data analytics.

II. OPTIMIZING AND UNDERSTANDING BUSINESS PROCESSES

Big data is also increasingly used in the optimization of business processes. Retailers are able to optimize their stock based on the basis of predictions generated from the social media data, web search trends and weather forecasts. A particular business process that has witnessed a lot of big data analytics is supply chain or delivery route optimization. HR business processes are also being improvised using big data analytics. This includes the optimization of talent acquisition too as well as the measurement of company culture and staff engagement using the tools of Bigdata.

III. PERFORMANCE OPTIMIZATION AND PERSONAL QUANTIFICATION

Big data is not limited to companies and governments but also for all of us individually. We can now benefit from the data generated from wearable devices such as smart watches and smart bracelets. Analyzing such bulky volumes of data will bring entirely new insights that it can feed back to individual users. The other area where we benefit from big data analytics is finding love virtually .Most online dating sites fooling the net apply big data tools and algorithms to find us the most appropriate matches.

IV. IMPROVING HEALTHCARE AND PUBLIC HEALTH

The computation power of big data analytics enables us to decode entire DNA strings in a few minutes and will allow us to find new cures and better understand and predict disease patterns. Big data techniques are already being used to monitor infants in premature and sick baby unit. By recording and analyzing every heart beat and breathing pattern of every baby, the unit was able to develop algorithms that can pre predict infections a day before any physical symptoms appear. That way, the team can come into action early and save fragile babies in an environment where every hour counts. Integration of data from medical records with social media analytics enables us to monitor flu outbreaks in real-time, simply by listening to what people have to say.

V. SPORTS PERFORMANCE IMPROVEMENT

Most elite sports are embracing big data analytics. We have the tool for tennis tournaments; we use video analytics to track the performance of every player in a soccer or baseball game, and sensor technology in sports equipment such as basket balls or golf clubs allows us to get feedback on our game and how to improvise it further.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

VI. IMPROVING RESEARCH AND SCIENCE

Science and research is currently being transformed by the new possibilities that have been brought by big data. Experiments to unlock the secrets of our universe – how it evolved and works - generate tremendous amounts of data. The CERN data center has 65,000 processors to analyze its 30 petabytes of data. Such computing capabilizes so many other areas of science and research.

VII. OPTIMIZING MACHINE AND DEVICE PERFORMANCE

Big data analytics help machines and devices become smarter and sufficiently autonomous. For example, big data tools are used to operate Google's self-driving car. Big data tools are also used in the optimization of energy grids using data from smart meters. We can even use big data tools to optimize the performance of computers and data warehouse units.

VIII. IMPROVING AND OPTIMIZING CITIES AND COUNTRIES

Big data is used to improve many aspects of our cities and countries. For example, it allows cities to optimize traffic flows based on real time traffic information as well as social media information and weather forecast. Currently piloting big data analytics with the aim of turning themselves into Smart Cities, where the transport infrastructure and utility processes are all joined up.

Connectivity between Cloud Computing and Big Data:
Big data is considered as an object of the computation oriented operations that cause the increase in the stress over various storage capacities of the cloud computing system. The main motive of the cloud computing is to handle a big amount of data applications with the efficient, fine-grained and low computational complexity with sufficient storage capacity and processing resources. The Cloud computing infrastructure development provides an ease of storage management, computing and processing of huge amount of data. Big Data is a data analysis methodology enabled by recent advances in technologies and architecture which support high velocity data capture, storage, and analysis. However, a huge commitment of hardware and software resources is essential, making adoption costs of big data technology prohibitive to small and medium sized businesses. Cloud computing offers the promise of big data implementation to small and medium sized businesses. A viable option for small to medium sized businesses considering the use of Big Data analytic techniques is data storage using cloud computing. Cloud computing is on-demand network access to computing resources which are often provided by an outside entity and requires just a little business interference.

Connectivity between IoT and Big Data:
The Internet of things involves an enormous amount of sensor nodes entrenched into various devices and machines for accomplishing a particular job. Big Data capacity is, in essence, a prerequisite to tapping into the Internet of Things. Without an optimized data storehouse,

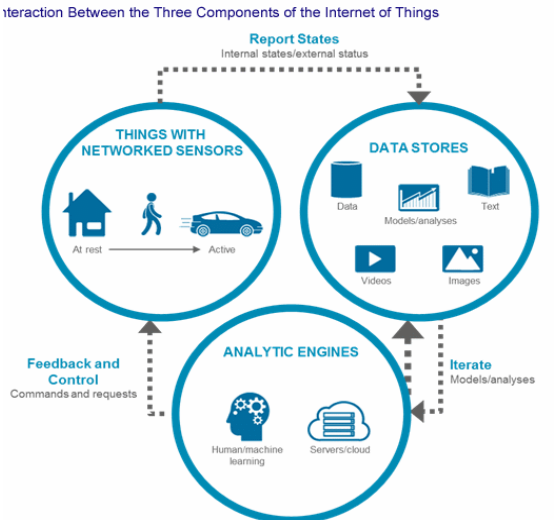

Fig2. IoT Interaction

it'll be impossible for businesses to sort through all the information flowing in from embedded sensors. What that means is that, without Big Data, the Internet of Things can offer an enterprise closer to noise than anything else.

Big Data Limitations:
Various limitations of Big Data program are highlighted below. Most of these are concerned with data achievement, Storage management and analysis.

I. Data Representation: Characterization of various types of datasets where it is found that a certain levels of heterogeneity in type are present such as structure, semantics, organization, granularity, and accessibility is really difficult. The main objective of the representation of data is to provide meaningful and structured data for analysis and user interpretation and access.

II. Redundancy Optimization and Data Compression: It is also observed in the area of Big Data analytics that presence of high level complexities in the datasets makes the exploration of data analysis very challenging. With the Redundancy optimization and data looseness techniques, it is easy to optimize the indirect cost of the entire system without affecting the possible standards of the data.

III. Data Life Cycle Management: Data life cycle management includes various challenges associated with the slow advances of storage management systems where a lot of processing challenges with current storage system faces scalability disorders.

IV. Analytical Mechanism: In The traditional RDBMS concept designing, scalability and expandability considered as two main issues thus the investigative arrangement of big data executes diverse data within a limited time.

V. Confidentiality: In the recent times most of the big data service providers are not able to efficiently maintain and analyze the big datasets as they have very limited computing capability and resources. The big organizations must depend on the other professional or services for

DOI 10.17148/IJARCCE.2016.5160

analyzing the data which increases the possibility of the vulnerable attack on data and poses safety risks.

VI. Energy Management: The energy depletion issues of mainframe computing systems grab the attention of the researchers from both the economy and environmental perspective as the processing, storage management and transmission of enormous amount of data consumes more energy. Big data may have the potential to yield more insights than smaller data, but it will take much more time, consideration, and technical ability in order to extract them. Meanwhile, there should be plenty of room to gain learnings and improve campaign results using less granular data.

Security and Privacy Preservations in Big data:

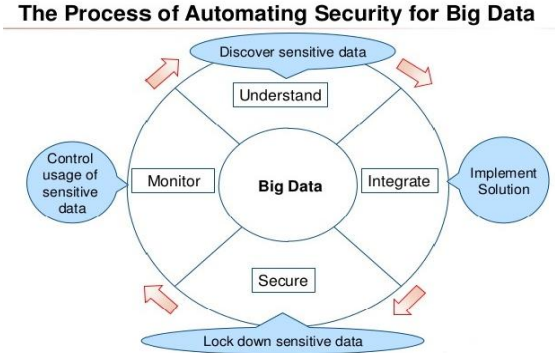

Figure3. Big Data and Security

Conventional security mechanisms fail to handle big data due its larger volume, variety and high velocity. Among various security aspects of big data, privacy is one of the major issues. This is because big data generally comprises of information specific to a person. The use of data for security tasks is however raising major privacy concerns across the globe. Collected data, even on being anonymized by removing identifiers such as personal names or social security numbers, when linked with other data lead to re-identify the individuals to which specific data items are related to, in original. Also, as organizations like governmental agencies, often need to collaborate on security tasks, data sets are exchanged across different organization domains, resulting in these data sets being available to multiple parties. Data anonymization or de-identification is also helpful in hiding personal information and narrowing the identification. It is the process of changing data that will be used and published in a way that prevents the identification of key information. There are basically three data anonymization methods that are used in the jaunt of preserving big data privacy. They are: K-Anonymity, L-Diversity, TCloseness. Differential privacy is a separate big data privacy preservation method that is being used. It is a method enabling analysts to extract answers from databases containing personal information while maintaining strong individual privacy protections. These reports released by the US government declines encryption as a perfect solution for privacy preservation and points to the inadequacies of data anonymization and de-identification techniques currently implemented.

IX. CONCLUSION

Challenges for Big Data processing and analysis are numerous. As all the data is currently visualized through computers, it leads to difficulties in the extraction of data, which is followed by its perception and cognition. Those tasks are time-consuming and seldom provide correct or acceptable results [8]. Big data is the "new" business and social science frontier or rather saviour. The amount of information and knowledge that can be extracted from the digital universe is continuing to grow in leaps as users come up with new ways to massage and process data. Big data is just the beginning solution. Technology evolution and placement guarantees that in a few years, a lot more data will be available in a year than has been collected since the dawn of man. We as a global society are evolving from data-centric towards a knowledge centric community. Our knowledge is widely distributed and equally widely accessible non limitingscope[12]. The future research shall concentrate more on developing a complete understanding of the issues associated with big data, and those factors that may contribute to a need for a big data analysis and design methodology towards consistency. Although it is viewed as an enabler of breakthroughs in key sectors of society, such as healthcare, science, business, law enforcement and national security, big data analytics entail important privacy, security and ethical challenges that technologists, regulators, business and the society at large have yet to address. The challenges include not just the obvious severe issues of scale, also heterogeneity, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. These technical and non technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of a single domain.

REFERENCES

[1] L. Douglas, "3d data management: Controlling data volume, velocity and variety," Gartner, Retrieved 6 (2001).
[2] IBM what is big data? - Bringing big data to the enterprise. http://www-01.ibm.com/software/in/databig data /, Accessed on Sept. 20, 2015.
[3] M. A. Beyer and L. Douglas, "The importance of big data: A definition," Stamford, CT: Gartner, 2012.
[4] J. S. Ward and A. Barker, "Undefined By Data: A Survey of Big Data Definitions," http://arxiv.org/abs/ 1309 .5821v1.
[5] IBM developer Works, http://www.ibm.com/developerworks/ library/l-hadoop-1 /, Accessed on Sept. 20, 2015.
[6] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes and R. E. Gruber, "Bigtable: A distributed storage system structured data," in ACM Transactions on Computer Systems (TOCS), vol. 26, no.2, (2008): 4.
[7] Yong Yu, Yi Mu and Giuseppe Ateniese, "Recent advances in security and privacy in big data," (2015): 365.
[8] A. C. Mora et. al, "Top ten big data security and privacy challenges," Cloud Security Alliance (2012).
[9] InfoWorld, http://www.infoworld.com/article/2613587/big-data /thereal-story-of-how-big-data-analytics-helped-obama-win.html, Accessed on Sept. 20, 2015.
[10] Stata: Data Analysis and Statistical Software, http://www.stata.com/
[11] Bernard Marr, The Awesome Ways Big Data Is Used Today to Change Our World, https://www.linkedin.com/pulse/20131113 065157-64875646-the awesome-ways-big-data-is-used-today-to-change- our world# notifications, Accessed on Sept. 20, 2015.
[12] Venkata Narasimha Inukollu , Sailaja Arsi and Srinivasa Rao Ravuri "Security issues associated with big data" International Journal of Network Security & Its Applications.