Big Data: Theory, Challenges and Application's Roadmap

Mrs. Shravani S. Ketkar
Department of Information Technology
Gogate Jogalekar College
Ratnagiri, Maharashtra, India
pundipat@yahoo.com

Mr. Dipesh D. Sawant
Department of Information Technology
Gogate Jogalekar College
Ratnagiri, Maharashtra, India
dipesh.sawant@hotmail.com

***Abstract*— Now a days the world has experienced the problems to store and manage a huge amount of data efficiently. Various sectors are facing those problems. To overcome those problems as well as the challenges, there is one solution to this is Big Data. Big data is a game changing thing. Big data is a game changing thing. Big data is going to use in wide areas to handle data produced by number of applications. It has received significant attention in recent years. To manage Big Data, it has some challenges. In this study, an attempt is made to review the basic theory and challenges of Big Data along with its usage in various sectors.**

Keywords*— *Big data; theory and challenges; sectors

I. Introduction

Big data is a term that describes the large volume of data both structured and unstructured that inundates a business on a day-to-day basis. In simplest terms, the phrase refers to the tools, processes and procedures allowing an organization to create, manipulate, and manage very large data sets and storage facilities. [1] But it's not the amount of data that's important. It's what organizations do with the data that matters.

Four dimensions (V's) of Big Data

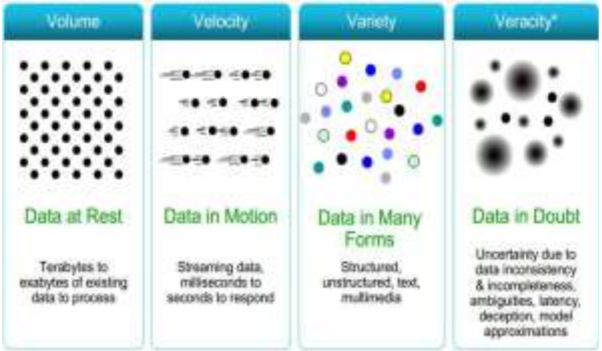
Volume
Velocity
Variety
Veracity*
Data at Rest
Terabytes to exabytes of existing data to process
Data in Motion
Streaming data, milliseconds to seconds to respond
Data in Many Forms
Structured, unstructured, text, multimedia
Data in Doubt
Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Figure 1: Four dimensions (V's) of Big Data

Big data described by the three dimensions occasionally referred to as the three V's but organizations need fourth V i.e. value to build big data job as below:

1. Volume = quantity, massive information sets that are command of size bigger than data managed in habitual storage and analytical results. Imagine petabytes rather than terabytes. [2]
2. Variety = structured, semi-structured, unstructured and complex, variable and heterogeneous data, which produced in different formats.
3. Velocity = Data created as a stable from batch processing to real-time queries for significant information to be present up on claim.
4. Veracity = quality, relevance, predictive value, meaningfulness i.e. Resulting insights that for trends and patterns, difficult analysis based on graph algorithms, machine learning and statistical modeling. These analytics overtake the results of querying, reporting and business intelligence. [3]

II. Big data architecture

In the economy fueled by the Internet and globalization, and amplified by social networks and mobile computing, Big Data is becoming an enterprise concern. The capability to capture, process and analyze Big Data can provide tremendous competitive advantage by increasing the firm's capability to respond to the dynamic market conditions and customer needs. In the

A Conceptual Cluster Architecture for Big Data

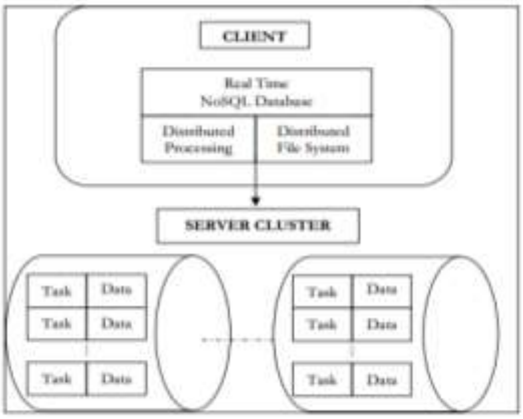

Figure 2: A Conceptual Cluster Architecture for Big Data

following, a client- server architecture for Big Data is presented (Figure 2).

III. Application domain

In today's world, big data is arguably important and often exploited for two main reasons. First, big data is utilized by companies for analytic purposes such as deriving useful insights about their businesses and supporting their higher level decision making. Second, big data enables the development of applications and real-time services that leverage massive amounts of electronic data in order to present customers with value (e.g., intelligent services, efficiency, and entertainment) that would not be possible without the availability of such data. All big data applications have to be equipped with a number of capabilities such as visualizing and personalizing data, integrating different sets of data, and exploring and analyzing data in a timely manner.

Table 1 lists the main Big Data origin domains and targeted use or application. We can assume high relevance of Big Data to business; this actually explains the current strong interest to Big Data from business which is actually becoming the main driving force in this technology domain.

Table 1. Big Data origin and target use domains

Big Data Origin	Big Data Target Use
1. Science 2. Telecom 3. Industry 4. Business 5. Living environment, Cities 6. Social media and networks 7. Healthcare	(a) Scientific discovery (b) New technologies (c) Manufacturing, process control, transport (d) Personal services, campaigns (e) Living environment support (f) Healthcare support

Science has been traditionally dealing with challenges to handle large volume of data in complex scientific research experiments, involving also wide cooperation among distributed groups of individual scientists and research organizations. Scientific research typically includes collection of data in passive observation or active experiments which aim to verify one or another scientific hypothesis. Scientific research and discovery methods are typically based on the initial hypothesis and a model which can be refined based on the collected data. The refined model may lead to a new more advanced and precise experiment and/or the previous data re-evaluation. The future Scientific Data and Big Data Infrastructure (SDI/BDI) needs to support all data handling operations and processes providing also access to data and to facilities to collaborating researchers. Besides traditional access control and data security issues, security services need to ensure secure and trusted environment for researcher to conduct their research.

The wealth of data available today offers the unprecedented opportunity to conduct science in ways never before envisioned: real-time science, real-world data, and new methods of scientific discovery. In embracing new approaches to scientific discovery that are not necessarily aligned with the scientific method, or maybe even contradict it, we must consider how to best employ and unify the new approaches with each other and with the traditional scientific method.

Understanding the environment requires collecting and analyzing data from thousands of sensors monitoring air and water quality and meteorological conditions, another example of eScience. These measurements can then be used to guide simulations of climate and groundwater models to create reliable methods to predict the effects of long-term trends, such as increased CO2 emissions and the use of chemical fertilizers.

Big Data in industry are related to controlling complex technological processes and objects or facilities. Modern computer-aided manufacturing produces huge amount of data which are in general need to be stored or retained to allow effective quality control or diagnostics in case of failure or crash. Similarly to eScience, in many industrial applications/scenarios there is a need for collaboration or interaction of many workers and technologists.

Big data in healthcare refers to electronic health data sets so large and complex that they are difficult (or impossible) to manage with traditional software and/ or hardware; nor can they be easily managed with traditional or common data management tools and methods [5]. Big data in healthcare is overwhelming not only because of its volume but also because of the diversity of data types and the speed at which it must be managed [5]. The totality of data related to patient healthcare and wellbeing make up "big data" in the healthcare industry. The increasing availability of health care data in the form of medical records, claims and cost data, R&D data from pharmaceutical companies, and other types of medical content has started to result in new types of big data applications, such as the applications that provide the analysis and aggregation of health care data as a service to third parties. These services can be used to improve clinal decision making, create preventive care programs, and help pharmaceutical and medical product companies in their R&D activities.

Innovative big data applications are also emerging in the retail industry. One type of application is price comparison services which offer pricing information on products from different retailers. Studies have shown that consumers can save an average of 10% when they shop using such services[4]. For example, RedLaser (www.redlaser.com) allows customers to scan the bar code of a product using their smart phones and obtain price comparisons for the product together with other product information.

Big data applications can create value for several other sectors including the public sector, manufacturing, recruitment, etc. For example, WhereDoesMyMoneyGo (www.wheredoesmymoneygo.org) provides a web site for the analysis and visualization of data about public spending in the United Kingdom. In the manufacturing sector, data obtained from sensors in a product may be used to create after-sales service offerings such as maintenance services for automobiles [4].

IV. Big data Challenges

Existing technologies has inadequate scalability and also facing challenges such as Volume, Velocity, Variety, Veracity and Value due to characteristics of Big Data, which gives the opportunities to work on efficient Big Data mining techniques which would results in success in this competitive world by overcoming Big Data challenges.

Challenges of big data include analysis, capture, data
curation, sharing, storage, search,transfer, visualization, querying and information privacy. The term often refers simply to the use of predictive analytics or certain other advanced methods to extract value from data, and seldom to a particular size of data set.

Some challenges of Big Data are as follows:

1. Heterogeneity and Incompleteness: When humans consume information, a great deal of heterogeneity is comfortably tolerated. In fact, the nuance and richness of natural language can provide valuable depth. However, machine analysis algorithms expect homogeneous data, and cannot understand nuance. In consequence, data must be carefully structured as a first step in (or prior to) data analysis. Consider, for example, a patient who has multiple medical procedures at a hospital. We could create one record per medical procedure or laboratory test, one record for the entire hospital stay, or one record for all lifetime hospital interactions of this patient.

2. Scale: The first thing anyone thinks of with Big Data is its size. For many decades, one challenging issue is managing large and rapidly increasing volumes of data. In the past, this challenge was mitigated by processors getting faster, following Moore’s law, to provide us with the resources needed to cope with increasing volumes of data. But, there is a fundamental shift underway now: data volume is scaling faster than compute resources, and CPU speeds are static.

3. Timeliness: The flip side of size is speed. The larger the data set to be processed, the longer it will take to analyze. The design of a system that effectively deals with size is likely also to result in a system that can process a given size of data set faster.

4. Privacy: The privacy of data is another huge concern, and one that increases in the context of Big Data. For electronic health records, there are strict laws governing what can and cannot be done. However, there is great public fear regarding the inappropriate use of personal data, particularly through linking of data from multiple sources. Managing privacy is effectively both a technical and a sociological problem, which must be addressed jointly from both perspectives to realize the promise of big data.

5. Human Collaboration: In the tremendous advances made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a hard time finding. Ideally, analytics for Big Data will not be all computational – rather it will be designed explicitly to have a human in the loop.

V. Conclusion

This paper has presented the basic theory about what is Big Data and its challenges. Also includes analysis of big data applications varying from scientific research to health care services. Overall, my analysis has showed the impact of big data on various application domain that requires the data processing. The arrival of Big Data in society has prompted business and government to take actions to exploit its value and applications.

References

[1] http://www.zdnet.com/blog/virtualization/what-is-big-data/1708.

[2] Bloem, J. Doorn, M. V. Duivestein, S. Manen & Ommeren, “ Creating clarity with Big Data”, Sogeti, 2012.

[3] Vinayak Borkar, Michael J. Carey, Chen Li, “Inside “Big Data Management”: Ogres, Onions, or Parfaits?”, EDBT/ICDT 2012 Joint Conference Berlin, Germany, 2012.

[4] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, “Big data: The next frontier for innovation, competition, and productivity,” McKinsey Global Institute, Tech. Rep., 2011.

[5] Frost & Sullivan:Drowning in Big Data? Reducing Information Technology Complexities and Costs for Healthcare Organizations.http://www.emc.com/collateral/analyst-reports/frost-sullivan-reducing-information-technologycomplexities-ar.pdf.