BIG DATA-ISSUES AND CHALLENGES

A.Benusha Begam, B.Amsappriya, Dr.A.Anitha (PROF & HEAD)

Abstract

The amount of data in world is growing day by day. Data is growing because of use of internet and social network. Big data is a collection of data sets which is very large in size as well as complex. Size of the data is Petabyte and Exabyte. Traditional database systems is not able to capture, store and analyze this large amount of data. As the internet is growing, amount of big data continues to grow. Big data analytics provide new ways for business and government to analyze unstructured data. Now a days, Big data is one of the most hot topic in IT industry. Big data changes the way that data is managed and used. Some of the applications are in areas such as healthcare, traffic management, banking, retail, education and so on. Organizations are becoming more flexible and more open. The present paper highlights important concepts of Big Data. In this write up various aspects of big data are discussed. Big Data and the parameters of Big Data is defined. Although reports on big data success stories have been accumulating in the media most organizations dealing with the three V’s it still faces numerous challenges.Here the grand challenges as well as possible solutions to address these challenges are described.

Keywords— Big data, Petabyte, Exabyte, Database, velocity, volume, variety

I. INTRODUCTION

In recent years, big data has rapidly developed into a hotspot that attracts great attention from academia, industry, and even governments around the world. Nature and Science have published special issues dedicated to discuss the opportunities and challenges brought by big data. McKinsey, the well-known management and consulting firm, alleged that big data has penetrated into every area of today’s industry and business functions and has become an important factor in production. What is big data? So far, there is no universally accepted definition. In Wikipedia, big data is defined as “an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications”. From a macro perspective, big data can be regarded as a bond that subtly connects and integrates the physical world, the human society, and cyberspace. Here the physical world has a reflection in cyberspace, embodied as big data, through Internet, the Internet of Things, and other information technologies, while human society generates its big data-based mapping in cyberspace by means of mechanisms like human-computer interfaces, brain-machine interfaces, and mobile Internet. In this sense, big data can basically be classified into two categories, namely, data from the physical world, which is usually obtained through sensors, scientific experiments and observations (such as biological data, neural data, astronomical data, and remote sensing data), and data from the human society, which is often acquired from such sources or domains as social networks, Internet, health, finance, economics, and transportation. Big data is of great value, which is beyond all doubt. From the perspective of the information industry, big data is a strong impetus to the next generation of IT industry, which is essentially built on the third platform, mainly referring to big data, cloud computing, mobile Internet, and social business. IDC predicted that by 2020 the market size of the third IT platform will reach US$ 5.3 trillion; and from 2013 to 2020, 90% of the growth in the IT industry would be driven by the third IT platform. From the socio-economic point of view, big data is the core connotation and critical support of the so called second economy, a concept proposed by American economist W.B.Arthur in 2011 which refers to the economic activities running on a processor, connectors, sensors and executors. It is estimated that by 2030 the size of the second economy will approach that of the first economy (namely, the traditional physical economy). The main support of the second economy is big data, as it is an inexhaustable and constantly enriching resource. In the future, by virtue of big data, the competence under the second economy will no longer be that of labor productivity but of knowledge productivity.

II. GRAND CHALLENGES OF BIG DATA

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

A. Data complexity
The emergence of big data has provided us with unprecedented large-scale samples when dealing with computational problems, although we now have to face far more complex data objects. As aforementioned, the typical characteristics of big data are diversified types and patterns, complicated inter-relationships, and greatly varied data quality. The inherent complexity of big data (including complex types, complex structures, and complex patterns) makes its perception, representation, understanding and computation far more challenging and results in sharp increases in the computational complexity when compared to traditional computing models based on total data. Traditional data analysis and mining tasks, such as retrieval, topic discovery, semantic analysis, and sentiment analysis, become extremely difficult when using big data. At present, we do not have a good understanding on addressing the complexity of big data. For instance, we lack knowledge regarding the laws of distribution and association relationship of big data. We lack deep understanding on the inherent relationship between data complexity and computational complexity of big data, as well as domain-oriented big data processing methods. All these greatly confine our capacity to design highly efficient computational models and methods for solving problems using big data.

A fundamental problem is how to formulate or quantitatively describe the essential characteristics of the complexity of big data. The study on complexity theory of big data will help understand essential characteristics and formation of complex patterns in big data, simplify its representation, get better knowledge abstraction, and guide the design of computing models and algorithms on big data. To do this, we will need to establish the theory and models of data distribution under multi-modal interrelationships. We will also need to sort out intrinsic connections between data complexity and spatiotemporal computational complexity. Moreover, by modeling and analyzing the intrinsic mechanisms of data complexity, we will be able to expound the principles and mechanisms for processing big data into a solid foundation for big data computing.

B. Computational complexity
Three of the key features of big data, namely, I. multi-sources II. huge volume III. fast-changing, make it difficult for traditional computing methods (such as machine learning, information retrieval, and data mining) to effectively support the processing, analysis and computation of big data. Such computations cannot simply rely on past statistics, analysis tools, and iterative algorithms used in traditional approaches for handling small amounts of data. New approaches will need to break away from assumptions made in traditional computations based on independent and identical distribution of data and adequate sampling for generating reliable statistics. When solving problems involving big data, we will need to re-examine and investigate its computability, computational complexity, and algorithms. New approaches for big data computing will need to address big data-oriented, novel and highly efficient computing paradigms, provide innovative methods for processing and analyzing big data, and support value-driven applications in specified domains. New features in big data processing, such as insufficient samples, open and uncertain data relationships, and unbalanced distribution of value density, not only provide great opportunities, but also pose grand challenges, to studying the computability of big data and the development of new computing paradigms.

To address the computational complexity of big data applications, we will need to focus on the whole life cycle of big data applications in order to study data-centric computing paradigms based on the characteristics of big data. We need to break away from traditional computing-centric paradigms and establish data-centric push-style computing paradigms and explore weak CAP network shared-data system model and its algebraic computational theory. We will need to develop algorithms for distributed and streaming computing and form a big data oriented computing framework where communication, storage, and computing are well
integrated and optimized. We will have to study non-deterministic algorithmic theory suitable for big data and de-part from the independent-and-identically-distributed assumption made in traditional statistical learning. We also need to explore existing reduction-based computing methods where big data is reduced on demand from being large enough to being just enough, and to being valuable enough. Finally, we will need to develop bootstrapping and sampling based local computation and approximation methods and propose novel theoretical basis for big data algorithms that are scalable to handling large amounts of data.

C. System complexity

Big data processing systems suitable for handling a diversity of data types and applications are the key to supporting scientific research of big data. For data of huge volume, complex structure, and sparse value, its processing is confronted by high computational complexity, long duty cycle, and real-time requirements. These requirements not only pose new challenges to the design of system architectures, computing frameworks, and processing systems, but also impose stringent constraints on their operational efficiency and energy consumption. The design of system architectures, computing frameworks, processing modes, and benchmarks for highly energy-efficient big data processing platforms is the key issue to be addressed in system complexity. Solving these problems can lay the principles for designing, implementing, testing, and optimizing big data processing systems. Their solutions will form an important foundation for developing hardware and software system architectures with energy-optimized and efficient distributed storage and processing.

The evaluation and optimization of energy efficiency of big data processing systems is a great research challenge. Not only do we need to untangle the relationship between complexity and computability of big data applications and between energy and energy consumption of processing systems, we will also need to comprehensively measure a variety of energy efficiency factors, including system throughput, parallel processing capabilities, job calculation accuracy, and energy consumption per unit. We also have to take actual workload conditions and scattered and repetitive resources into account. We will need to conduct fundamental research on performance evaluation, distributed system architecture, streaming computing framework, and online data processing, while taking into account features of value sparsity and weak access locality and the life cycle of big data applications. We will need to investigate validation tools, including benchmarks and system performance prediction methods. Through an iterative process of design, implementation, and validation, we will be able to develop big data processing systems with a high data acquisition rate of data generation is rising exponentially. With the introduction of smart devices which are used with a wide array of sensors continuously generate data. The Large Hadron Collider in Switzerland produces petabytes of data. Most of this data is not useful and can be discarded, however due to its unstructured form; selectively discarding the data presents a challenge. This data becomes more potent in nature when it’s merged with other valuable data and superimposed. Due to the interconnectedness of devices over the World Wide Web, data is increasingly being collated and stored in the cloud.

B. Data Extraction:

All of the data generated and acquired is not of use. It contains a large amount of redundant or unimportant data. For instance, a simple CCTV camera, constantly polls sensor to gather information of the user’s movements. However, when the user is in a state of inactivity, the data generated by the activity sensor is redundant and of no use. The challenges presented in data extraction are twofold: firstly, due to nature of data generated, deciding which data to keep and which to discard increasingly depends on the context in which the data was initially generated. For instance, footage of a security camera with the same frames may be discarded however it is important not to discard similar data in a case where it is being generated by a heart-rate sensor.

Secondly, a lack of a common platform presents its own set of challenges. Due to wide variety of data that exists, bringing them under a common platform to standardize data extraction is a major challenge.
C. Data Collation
Data from a singular source often is not enough for analysis or prediction. More than one data sources are often combined to give a bigger picture to analyze. For example a health monitor application often collects data from the heart-rate sensor, pedometer, etc. to summarize the health information of the user. Likewise, weather prediction software take in data from many sources which reveal the daily humidity, temperature, precipitation, etc. In the scheme of Big Data convergence of data to form a bigger picture is often considered a very important part of processing.

D. Data Structuring
Once all the data is aggregated, it is very important to present and store data for further use in a structured format. The structuring is important so queries can be made on the data. Data structuring employs methods of organizing the data in a particular schema. Various new platforms, such as NoSQL, can query even on unstructured data and are being increasingly used for Big Data Analysis. A major issue with big data is providing real time results and therefore structuring of aggregated data needs to be done at a rapid pace.

E. Data Visualization
Once the data is structured, queries are made on the data and the data is presented in a visual format. Data Analysis involves targeting areas of interest and providing results based on the data that has been structured. For instance, data containing average temperatures are shown alongside water consumption rates to calculate a relation in between them. This analysis and presentation of data makes it ready for consumption for users. Raw data cannot be used to gain insights or for judging patterns, therefore “humanizing” the data becomes all the more important.

F. Data Interpretation:
The ultimate step in Big Data processing includes interpretation and gaining valuable information from the data that is processed. The information gained can be of two types: Retrospective Analysis includes gaining insights about events and actions that have already taken place. For instance, data about the television viewership for a show in different areas can help us judge the popularity of the show in those areas. Prospective Analysis includes judging patterns and discerning trends for future from data that is already been generated. Weather Prediction using big data analysis is an example of prospective analysis. Problems accruing from such interpretations pertain to fallacious and misleading trends being predicted. This is particularly dangerous due to an increasing reliance on data for key decisions. For example, if a particular symptom is plotted against the likelihood of being diagnosed with a particular disease, it might lead to misinformation about the symptom being caused due to the particular disease itself. Insights gained from data interpretation are therefore very important and the primary reason for processing big data as well.

IV. CONCLUSION AND FUTURE SCOPE
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. Global Data is on the rise, by 2020, we would have quadrupled the data we generate every day. This data would be generated through a wide array of sensors we are continuously incorporating in our lives. Data collection would be aided by what is today dubbed as the “Internet of Things”. Through the use of smart bulbs to smart cars, everyday devices are generating more data than ever before. These smart devices are incorporated not only with sensors to collect data all around them but they are also connected to the grid which contains other devices. 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. Focus Groups are becoming increasingly redundant as analytics firms such as McKinsey are using analysis on very large sample bases that have today been made possible due to advancements in Big Data. The potential value of global personal location data is estimated to be $700 billion to end users, and it can result in an up to 50% decrease in product development and assembly costs, according to a recent McKinsey report. Big Data does not arise out of a vacuum: it is recorded from some data generating source. For example, consider our ability to sense and observe the world around us, from the heart rate of an elderly citizen, and presence of toxins in the air we breathe, to the planned square kilometer array telescope, which will produce up to 1 million terabytes of raw data per day. Similarly, scientific experiments and simulations can easily produce petabytes of data today. Much of this data is of no interest, and it can be filtered and compressed by orders of magnitude. There is immense scope in Big Data and a huge scope for research and Development.

References