A SURVEY ON BIG DATA ANALYSIS – AN OVERVIEW

Harshad Gupta[1] and Shilpa Nimbre[2]
Student[1] and Assistant Professor[2], Department of Information Technology and Mathematics, S.I.A College of Higher Education University of Mumbai, Dombivli (East)

ABSTRACT

A huge amount of data is generated each day from the modern informationsystems and digital technologies (Internet of Things i.e. IoT and Cloud Computing). Big data refers to the datasets that are not only big, but also high in variety and velocity, which makes them difficult to process and handle using tradition tools and techniques such as RDBMS system and other structured database management system. In these paper a detailed studyabout big data analysis, its basic concept, history, applications, techniques and tools are discussed. This paper mainly aims to analyze some of the different analytical methods and tools which can be applied to big data, as well as different opportunities provided by big data analysis in various decision domains. As a result, this paper provides a platform to explore big data at number of stages.

Keywords: Big data, structured data, analytics, decision making, knowledge discovery, etc.

1. OBJECTIVE

As data is growing day by day, the objective of the paper is to give insight of technologies which can be applied to big data and to provide a platform to explore big data at number of stages.

2. REVIEW OF LITERATURE

Dr. S. Vijayarani and Ms. S. Sharmila [1] explained the concept of big data and give information about various data visualization tools and technologies used with big data. They also gave advantages and disadvantages of each visualization tools.

D.P. Acharjya and Kauser Ahmed P.[2] presented techniques for analyzing big data with emphasis on three emerging tools namely MapReduce, Apache Spark, and Storm. As per their paper most of the available tools concentrate on batch processing, stream processing, and interactive analysis.

Nada Eglendy and Ahmed Elragal [3] presented the characteristics of big data and opportunities provided by the application of big data analytics in various decision domains. Neelam Singh, Neha Garg, Varsha Mittal [4] explained various big data taxonomies and challenges in handling Big Data.

Francis X. [5] gave thought on big data as a phenomenon and also as an emerging discipline.

3. INTRODUCTION AND STATEMENT OF PROBLEM

Consider a world without data storage, a place where every detail information about individual person or organization, every transaction performed on system, or every aspects which can be documented is lost directly after use. Thus, in this case the organization would lose the ability to extract valuable information and knowledge from those data [2].

In this growing digital world, data are generated at high speed from various sources and the fast transition from the digital technology has led to growth of big data. In general, big data refers to a collection of large and complex dataset which are difficult to process using tradition database management tools or data processing techniques. Big data are available in structured, semi-structured and unstructured formats in petabytes and beyond. The structure of big data can be explained using its 5Vs namely volume, velocity, variety, veracity and value.

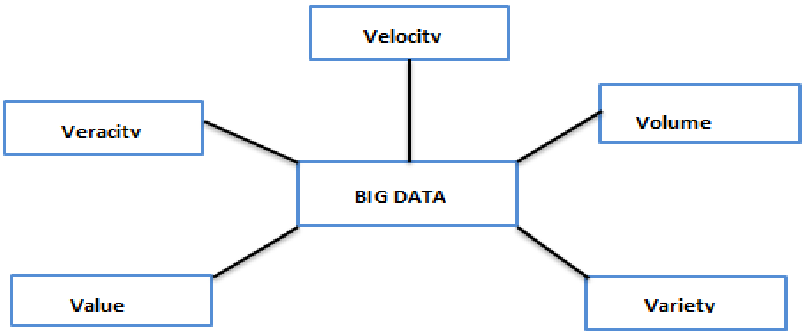

Figure 1: Structure of Big Data

Figure 1 explains the structure of big data which contains its five dimension or Vs. Volume describe the size of big data which mainly show how to handle large scalability database and high dimensional database and its processing needs. Velocity describes the speed at which the data is generated and the speed at which the data is processed. If the data is not processed at desire time then it loses its importance. Variety describes the different types of data which contribute to big data. The data coming from different sources does not follow any format or structure. Data can be vary from structured data to unstructured data and semi-structured data. Veracity include availability and accountability of data. The last v of big data refer to value which define the accuracy of data generated. Few typical characteristics of big data are the integration of structured, semi-structured and unstructured data. Big data addresses speed and measurability, quality, security, flexibility and stability. Another important advantage of big data is data analytics. Big data analytics refers to the process of collecting the raw data, organizing and analyzing large sets of data to discover patterns and other useful information from raw data.

4. RESEARCH METHODOLOGY

In order to gain better insight of the topic selected, various research papers were studied, which are mentioned in the reference.

5. LIMITATION OF THE STUDY

In big data analysis, the data analyst use big data to tease out correlation when one variable is connected to another variable. But, sometimes not all correlations are important or meaningful. As with many technological barriers and endeavors, big data analytics is prone to data breach and lead to data compromise. Sometimes, the tools we are using for gathering the raw big data sets are imprecise. For big data, you need to know how to use it to your advantage in order for it to be useful.

I. NEED OF BIG DATA

The massive volume of data could not be handle by traditional database management systems and tools which mainly focused and handled structured data [4]. At the time of development of computers the amount of data stored in computers are very less due to its minimum storage capacity. After the invention of networking, the data stored in the computer systems are increased because the improved developments in hardware components. The arrival of internet creates a boom to store cast collection of data and it can be used for various purpose in various domains. This situation raised concerns about the need of new research related concepts like data mining, networking, image processing, grid computing, cloud computing, etc. which are used for analyzing the different types of data which are used in various domains. Many new technologies, algorithms and concepts have been proposed by the researchers for analyzing the static data sets. In this digital era, after the invention of mobiles and wireless technologies provides a new platform in which people may share their information through social media sites such as Facebook, twitter and Google [3]. In these places, the data may arrived continuously and it cannot be store in computer memory because the size of data is huge and it is considered as “Big Data”.

The term “Big Data” came into existence for first time in 1998 in a silicon graphics by John Mashey [1]. The growthof big data leads to increase the storage capacityand processing power. Frequently large amount of data are created through social networking. Big data analytics are used to examine this large amount of data and identifies the hidden patterns and unknown relationships. Two technologies are used in big data analytics which are NoSQL and Hadoop. NoSQL is a non-relation or non SQL database solution such as HBase, Cassandra and mongoDB [1]. Hadoop is an eco-software package which includes HDFS and MapReduce for analysis of data. Big data rely on structured, semi-structured and unstructured data to back up their decision.

Big data applications have introduced the large scale distribution applications which are work with large data sets.

II. BIG DATA TECHNOLOGIES Column-oriented databases

In column-oriented database, the data is stored in columns rather than in rows, which is used to compresses massive data and fast queries [5].

Schema-less databases

Schema-less databases are also called NoSQL databases. These databases provide different mechanism for storage and retrieval of data that is modeled in means other than the tabular relations used in structured database management systems. There are two types of database such as document stores and key-value stores which stores and retrieves massive amount of structured, semi-structured and unstructured data [5].

Hadoop

Hadoop is a popular open source tool for handling big data and implemented in MapReduce. It is java based programming framework which supports large data sets in distributed computing. Hadoop clusters use a master and slave structure. Distributed file system in hadoop system helps to transfer data in rapid rates. In case of some node failure distributed system allows the system to continue the normal operation. Hadoop has two main sub projects namely Map Reduce and HDFS (Hadoop Distributed File System) [2].

Map Reduce

This is a programming paradigm which allows execution scalability against thousands of servers and server clusters for large tasks. Map Reduce implementation consist of two tasks namely map task and reduce task. In the map task the input data sets are converted into key/value pairs or tuples. In reduce task several forms of output of map task is combined to form a reduced set of tuples.

HDFS

HDFS is a file system which extends all nodes in hadoop cluster for data storage. It links all the file system together on local node to make into single large file system. To overcome the node failure HDFS enhances the security by depicting data across multiple sources [1].

Hive

Hive is a data warehousing infrastructure toolwhich is built on hadoop. It has different storage types such as plain text, ORC, HBase, RC file etc. Built-in user-defined functions in Hive are used to handle dates, strings and other data mining tools.

HBase

HBase is a scalable distributive database which uses Hadoop distributed file system for storing raw data. It supports column-oriented database and structure data.

Chukwa

Chukwa is a systemwhich monitors large distributed systemand it adds required semantics for log collections and it uses end to end delivery model.

Storage Technology

To store huge volume of data in the system, efficient and effective techniques are required. The main focus of storage technologies is data compression and storage virtualization to effectively store raw data [1].

Table-1: advantages and disadvantages of big data technologies

Technology	Advantages	Disadvantages
Column Oriented Databases	• Scalable and fast data loading for Big Data • Accessible by many third-party BI analytic tools • Simple systems administration • High performance on aggregation queries. • Highly efficient for data compression and/or partitioning	• Record updates and deletes reduce storage efficiency • Very difficult to design Effective partitioning/indexing schemes • Transactions are to be avoided or just not supported • Performance reduces because of Queries with table joins
Schema-less databases	• Flexibility • Can be more tolerant of variable Acidity and Consistency models • Ease of use and maintenance	• poor Integrity • Ambiguity
Hadoop	• Scalable • Cost effective • Flexible • Fast • Resilient to failure	• Security Concerns • Not Fit for Small Data • Potential Stability Issues
MapReduce	• Scalability • Flexibility • Security and Authentication • Cost-effective solution	• Data processing is Slow • Limited up to Batch processing • Caching is restricted • Latency
HDFS	• inexpensive • mature technology • provides very high performance for sequential reads and writes	
Hive	• Table structure are like tables in a relational database. • Hive-QL is used by multiple users to simultaneously query the data • more detailed data analysis can be performed by writing custom MapReduce framework processes to perform • Data extract/transform/load (ETL) can be done easily. • It provides the structure on a variety of data formats. • Converting diversity of format from to within Hive is simple.	• It's not designed for online transaction processing (OLTP), it is only used for the Online Analytical Processing (OLAP). • Hive do not support update and delete, but supports overwriting or apprehending data. • Sub-queries are not supported, in Hive
HBase	• Large data sets • Fast processing • Failover support and load sharing • Scalability • Schema-less	• It is not possible to implement any cross data operations and joining operations, in HBase. • When we want to migrate data from RDBMS external sources to HBase servers, HBase would require a new design.

III. CHALLENGES IN BIG DATA ANALYTICS

In the current year big data has been served in several domains like healthcare, public administration, retail, and bio-chemistry and other interdisciplinary scientific research [2]. Web-based applications encounter big data frequently, such as social computing, internet text and documents and internet search indexing. Social computing includes social network analysis, online communities, recommender systems, reputation systems and prediction markets. Consider the advantages of big data it provides new opportunities in the knowledge processing tasks for the upcoming researchers. However opportunities always follow some challenges.

To handle the challenges we need to know various computational complexities, information securityand computational methods to analyze big data. For example, many statistical methods perform well for small data sizes do not scale to voluminous data. Similarly, many computational techniques that perform well for small data face significant challenge in analyzing big data [2]. Here the challenges in big data are classified into four main categories namely data storage and analysis; knowledge discovery and computational complexities; scalability and visualization of data; and information security.

A. Data Storage and Analysis

In the recent years the size of the data has grown exponentially by various means such as mobile devices. Remote sensing, Aerial sensory technologies, radio frequency identification readers, etc. these are stored on spending more cost whereas they ignored or deleted finally because there is no enough space to store them. Hence, the first challenge in big data analysis is storage medium and higher input/output speed [1]. In such case, the accessibility of data in on the top priority for the knowledge discovery and representation.

Another challenge in big data analysis is attributed to diversity of data. With the growth of datasets, data mining tasks has significantly increased. Additional data reduction, data selection, feature selection is an essential task especially when dealing with large datasets. It is because, existing algorithms may not always respond in adequate time when dealing with these high dimensional datasets. Automation of this process and developing new machine learning algorithms to ensure consistency is major challenge in recent years. The major challenge in this case is to take more attention for designing storage systems and to elevate efficient data analysis tool that provide guarantees on the output when the data comes from different sources [2].

B. Knowledge Discovery and Computational Complexities

Knowledge discovery and representation is prime issue in big data analysis. It includes a number of subfields such as authentication, authorization, archiving, management, preservation, information retrieval and representation. There are several tools for knowledge discovery and representation such as fuzzy set, rough set, soft set, near set, formal concept analysis, etc. Since the size of big data keeps increasing the available tools may not be sufficient to process the data for obtaining meaningful information. The most popular approach in case of large dataset management is data warehouses and data marts. Data warehouse mainly used to store the data that are sourced from operational systems whereas data marts is based on a data warehouse and facilities analysis.

However, current big data analytic tools have poor performance in handling computational complexities, uncertainty and inconsistence. It leads to great challenge to develop techniques and technologies that can deal computational complexity, uncertainty and inconsistency in an effective manner [2].

C. Scalability and Visualization of Data

The most important challenge for big data analysis techniques is its scalability for future work and security. Incremental techniques have good scalability property in the aspect of big data analysis. As the size of the data is scaling much faster than CPU speeds, there is a natural dramatic shift in technology of processor being embedded with increasing number of cores.

The objective of visualizing data is to present them more adequately and efficiently using some techniques of graph theory. Graphical visualization provide link between data with proper interpretation. We can see that big data have produced many challenges for the development of hardware and software which leads to parallel computing, cloud computing, distributed computing, visualization process and scalability [2]. To overcome this issue, we need to correlate more mathematical models to computer science.

D. Information Security.

In big data analysis massive amount of data are correlated, analyzed and mined to find meaningful patterns. All organization has different policies to safe guard their sensitive information. Preserving sensitive information is major issue in big data analysis. There is huge risk associated with big data. Hence, information security is becoming a big data analytics problem. Security of big data can be enhanced using techniques of authentication, authorization and encryption [1].

Various security measures that big data applications face are the scale of network, variety of different devices, real time security monitoring system and lack of intrusion system. The security challenge caused by big data has attracted the attention of information security in big data. Hence, attention has to be given to develop a multi-level security policy model and prevention system [2].

IV. OPEN RESEARCH ISSUES IN BIG DATA ANALYTICS

Big data analytics and data science are becoming the research focal point in industry and academia. Data sciences are aims at researching big data and knowledge extraction for data [2]. Effective integration of different technologies and analysis will result in predicting the future drift of events. The research issues pertaining to big data analysis are classified into three broad categories namely Internet of Things (IoT), cloud computing, bio-inspired computing and quantum computing. However it is not limited to these issues.

IoT for Big Data Analytics

Currently, machines are getting in on the act to control innumerable autonomous gadgets via internet and create Internet of Things. Hence, the appliances are becoming the user of the internet, just like humans with the web browsers. IoT is attracting the attention of researchers for its most promising opportunities and problems [2].

Knowledge acquisition from IoT data is the biggest challenge that big data professionals are facing. Hence it is essential to develop infrastructure to analyze the IoT data. IoT devices generate continuous streams of data and the researchers can develop tools to extract meaningful information from these data using machine learning techniques.

Cloud Computing for Big Data Analytics

The development of virtualization technologies have made supercomputing more accessible, securable and affordable. Computing infrastructure that are hidden in virtualization software make systems to behave like a true computer, but with the flexibility of specification details such as number of processor, disk space, memory and operating system.

Big data application using cloud computing should support data analytic development. The cloud environment should provide tools that allow data scientists and business analyst to interactively and collaboratively explore knowledge acquisition data for further processing and extracting faithful information [2].

Bio-inspired Computing for Big Data Analytic

Bio-inspired computing is a technique inspired by b\nature to address complex real world problems. A bio-inspired cost minimization mechanism search and find the optimal data service on considering cost of data management and service management [2].

Bio-inspired computing techniques serve as a key role in intelligent data analysis and its application in big data analysis. These algorithms help to perform data mining for large datasets due to its optimization application.

Quantum Computing for Big Data Analytics

A quantum computer has memory that is exponentially larger than its physical size and can manipulate an exponential set of inputs simultaneously [2]. If a real quantum computer is available now, it could have solved problems that are exceptionally difficult on recent computer system, of course today’s big data problems.

The main technical difficulty in building quantum computer could as soon as be possible. Quantum computing provides a way to merge the quantum mechanics to process the information. In traditional computer, information is presented by long strings of bits which encoded in either a zero or a one. On the other hand a quantum computer uses another bit known as quantum bits or qubits [2].

6. SCOPE OF THE STUDY

This paper gives light on various aspects such as information about big data, technologies used with big data, challenges, and research issues. This paper is going to be useful for everyone either academician, researcher etc. who want to explore or to get basic idea about big data.

7. CONCLUSION

In recent years data are generated at a dramatic speed. Analyzing these data is challenging for a general man and software. To this end in this paper, we survey the various research issues, challenges, and tools used to analyze these raw big data. From this survey, it is understood that every big data platform has its individual focus and features. Some of them are designed for batch processing whereas some are good at real-time analytic and visualization. Each big data platform also has specific functionality. Different techniques used for the analysis include statistical analysis, machine learning, data mining, intelligent analysis, cloud computing, quantum computing, and data stream processing. We believe that in future researchers will pay more attention to these techniques to solve problems of big data effectively and efficiently.

REFERENCES

[1] Dr. S. Vijayarani and Ms. S. Sharmila, “Research in Big Data – An Overview”, Informatics Engineering, an International journal (IEIJ), Vol.4, No.3, September 2013.

[2] D.P. Acharjya and Kauser Ahmed P., “A survey on Big Data Analytics: Challenges, Open Research Issues and Tools”, International journal of Advanced Computer Science and Application (IJACSA), Vol.7, No.2, 2016.

[3] Nada Eglendy and Ahmed Elragal, “Big Data Analytics: A Literature Review Paper”, September 2014.

[4] Neelam Singh, Neha Garg, Varsha Mittal, “Data – insights, motivation and challenges”, Volume 4, Issue 12, December-2013, 2172, ISSN 2229-5518 2013.

[5] Francis X. “On the Origin(s) and Development of the Term \Big Data"_ Francis X., 2012