Big Data Analysis

Abhijeet Bhattacharya
Arka Jain University Jamshedpur, Jharkhand, India

Abstract— In the information era, enormous amounts of data have become available on hand to decision makers. Big data refers to datasets that are not only big, but also high in variety and velocity, which makes them difficult to handle using traditional tools and techniques. Due to the rapid growth of such data, solutions need to be studied and provided in order to handle and extract value and knowledge from these datasets. Furthermore, decision makers need to be able to gain valuable insights from such varied and rapidly changing data, ranging from daily transactions to customer interactions and social network data. Such value can be provided using big data analytics, which is the application of advanced analytics techniques on big data. This paper aims to analyse some of the different analytics methods and tools which can be applied to big data, as well as the opportunities provided by the application of big data analytics in various decision domains.

Keywords: Big Data, Data Mining, Analytics, Decision Making

I. INTRODUCTION

A collection of large and complex data sets which are difficult to process using common database management tools or traditional data processing applications. Big data is not just about size. Finds insights from complex, noisy, heterogeneous, longitudinal, and voluminous data. It aims to answer questions that were previously unanswered. Big Data constantly facing significant challenges like outsized, heterogeneity, noisy labels, non-stationary distribution. Capturing, storing, searching, sharing & analysing. The four dimensions (V's) of Big Data It is important to recognize the full potential of Big Data by addressing these technical challenges with new ways of thinking and transformative solutions. If these challenges are resolved on time, there will be a plenteous opportunities to provide major advancement in science, medicine and business. While there is clearly an important research space examining the fundamental methods and technologies for big data analytics, it is vital to acknowledge that it is also necessary to fund domain targeted research that allows specialized solutions to be developed for specific applications. Healthcare, in general, deserves to be a natural candidate for this kind of evaluation. Above diagrammatic representation explains the advantage of the massive amounts of data which provide right intervention to the right patient at the right time. Personalized care to the patient that potentially benefit all the components of a healthcare system i.e., provider, payer, patient, and management.

A. *Big Data Analytics*

The term "Big Data" has recently been applied to datasets that grow so large that they become awkward to work with using traditional database management systems. They are data sets whose size is beyond the ability of commonly used software tools and storage systems to capture, store, manage, as well as process the data within a tolerable elapsed time. Big data sizes are constantly increasing, currently ranging from a few dozen terabytes (TB) to many petabytes (PB) of data in a single data set. Consequently, some of the difficulties related to big data include capture, storage, search, sharing, analytics, and visualizing. Today, enterprises are exploring large volumes of highly detailed data so as to discover facts they didn't know before. Hence, big data analytics is where advanced analytic techniques are applied on big data sets. Analytics based on large data samples reveals and leverages business change. However, the larger the set of data, the more difficult it becomes to manage. In this section, we will start by discussing the characteristics of big data, as well as its importance. Naturally, business benefit can commonly be derived from analysing larger and more complex data sets that require real time or near-real time capabilities; however, this leads to a need for new data architectures, analytical methods, and tools. Therefore the successive section will elaborate the big data analytics tools and methods, in particular, starting with the big data storage and management, then moving on to the big data analytic processing. It then concludes with some of the various big data analyses which have grown in usage with big data.

B. *Characteristics of Big Data*

Big data is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures, analytics, and tools in order to enable insights that unlock new sources of business value. Three main features characterize big data: volume, variety, and velocity, or the three V's. The volume of the data is its size, and how enormous it is. Velocity refers to the rate with which data is changing, or how often it is created. Finally, variety includes the different formats and types of data, as well as the different kinds of uses and ways of analysing the data. Data volume is the primary attribute of big data. Big data can be quantified by size in TBs or PBs, as well as even the number of records, transactions, tables, or files. Additionally, one of the things that make big data really big is that it's coming from a greater variety of sources than ever before, including logs, clickstreams, and social media. Using these sources for analytics means that common structured data is now joined by unstructured data, such as text and human language, and semi-structured data, such as extensible mark-up Language (XM) or Rich Site Summary (RSS) feeds. There's also data, which is hard to categorize since it comes from audio, video, and other devices. Furthermore, multi-dimensional data can be drawn from a data warehouse to add historic context to big data. Thus, with big data, variety is just as big as volume. Moreover, big data can be described by its velocity or speed. This is basically the frequency of data generation or the frequency of data delivery. The leading edge of big data is streaming data, which is collected in real-time from the websites. Some researchers and organizations have discussed the addition of a fourth V, or veracity. Veracity focuses on the quality of the data. This characterizes big data quality as good, bad, or undefined due to data inconsistency, incompleteness, ambiguity, latency, deception, and approximations.

II. Analysis

With the evolution of technology and the increased multitudes of data flowing in and out of organizations daily, there has become a need for faster and more efficient ways of analysing such data. Having piles of data on hand is no longer enough to make efficient decisions at the right time. Such data sets can no longer be easily analysed with traditional data management and analysis techniques and infrastructures. Therefore, there arises a need for new tools and methods specialized for big data analytics, as well as the required architectures for storing and managing such data. Accordingly, the emergence of big data has an effect on everything from the data itself and its collection, to the processing, to the final extracted decisions. Consequently, proposed the Big – Data, Analytics, and Decisions (B-DAD) framework which incorporates the big data analytics tools and methods into the decision making process. The framework maps the different big data storage, management, and processing tools, analytics tools and methods, and visualization and evaluation tools to the different phases of the decision making process. Hence, the changes associated with big data analytics are reflected in three main areas: big data storage and architecture, data and analytics processing, and, finally, the big data analyses which can be applied for knowledge discovery and informed decision making. Each area will be further discussed in this section. However, since big data is still evolving as an important field of research, and new findings and tools are constantly developing, this section is not exhaustive of all the possibilities, and focuses on providing a general idea, rather than a list of all potential opportunities and technologies. Big Data Storage and Management One of the first things organizations have to manage when dealing with big data, is where and how this data will be stored once it is acquired. The traditional methods of structured data storage and retrieval include relational databases, data marts, and data warehouses. The data is uploaded to the storage from operational data stores using Extract, Transform, Load (ETL), or Extract, Load, Transform (ELT), tools which extract the data from outside sources, transform the data to fit operational needs, and finally load the data into the database or data warehouse. Thus, the data is cleaned, transformed, and catalogued before being made available for data mining and online analytical functions. However, the big data environment calls for Magnetic, Agile, Deep (MAD) analysis skills, which differ from the aspects of a traditional Enterprise Data Warehouse (EDW) environment. First of all, traditional EDW approaches discourage the incorporation of new data sources until they are cleansed and integrated. Due to the ubiquity of data nowadays, big data environments need to be magnetic, thus attracting all the data sources, regardless of the data quality. Furthermore, given the growing numbers of data sources, as well as the sophistication of the data analyses, big data storage should allow analysts to easily produce and adapt data rapidly. This requires an agile database, whose logical and physical contents can adapt in sync with rapid data evolution. Finally, since current data analyses use complex statistical methods, and analysts need to be able to study enormous datasets by drilling up and down, a big data repository also needs to be deep, and serve as a sophisticated algorithmic runtime engine. Accordingly, several solutions, ranging from distributed systems and Massive Parallel Processing (MPP) databases for providing high query performance and platform scalability, to non-relational or in-memory databases, have been used for big data. Non-relational databases, such as Not Only SQL (NoSQL), were developed for storing and managing unstructured, or non-relational, data. NoSQL databases aim for massive scaling, data model flexibility, and simplified application development and deployment. Contrary to relational databases, NoSQL databases separate data management and data storage. Such databases rather focus on the high-performance scalable data storage, and allow data management tasks to be written in the application layer instead of having it written in databases specific languages. On the other hand, in-memory databases manage the data in server memory, thus eliminating disk input/output (I/O) and enabling real-time responses from the database. Instead of using mechanical disk drives, it is possible to store the primary database in silicon-based main memory. This results in orders of magnitude of improvement in the performance, and allows entirely new applications to be developed. Furthermore, in-memory databases are now being used for advanced analytics on big data, especially to speed the access to and scoring of analytic models for analysis. This provides scalability for big data, and speed for discovery analytics. Alternatively, Hadoop is a framework for performing big data analytics which provides reliability, scalability, and manageability by providing an implementation for the Map Reduce paradigm, which is discussed in the following section, as well as gluing the storage and analytics together. Hadoop consists of two main components: the HDFS for the big data storage, and Map Reduce for big data analytics. The HDFS storage function provides a redundant and reliable distributed file system, which is optimized for large files, where a single file is split into blocks and distributed across cluster nodes. Additionally, the data is protected among the nodes by a replication mechanism, which ensures availability and reliability despite any node failures. There are two types of HDFS nodes: the Data Nodes and the Name Nodes. Data is stored in replicated file blocks across the multiple Data Nodes, and the Name Node acts as a regulator between the client and the Data Node, directing the client to the particular Data Node which contains the requested data.

Big Data Analytic Processing After the big data storage, comes the analytic processing. According to there are four critical requirements for big data processing. The first requirement is fast data loading. Since the disk and network traffic interferes with the query executions during data loading, it is necessary to reduce the data loading time. The second requirement is fast query processing. In order to satisfy the requirements of heavy workloads and real-time requests, many queries are response-time critical. Thus, the data placement structure must be capable of retaining high query processing speeds as the amounts of queries rapidly increase. Additionally, the third requirement for big data processing is the highly efficient utilization of storage space.

Since the rapid growth in user activities can demand scalable storage capacity and computing power, limited disk space necessitates that data storage be well managed during processing, and issues on how to store the data so that space utilization is maximized be addressed. Finally, the fourth requirement is the strong adaptively to highly dynamic workload patterns. As big data sets are analysed by different applications and users, for different purposes, and in various ways, the underlying system should be highly adaptive to unexpected dynamics in data processing, and not specific to certain workload patterns. Map Reduce is a parallel programming model, inspired by the "Map" and "Reduce" of functional languages, which is suitable for big data processing. It is the core of Hadoop, and performs the data processing and analytics functions. According to EMC, the Map Reduce paradigm is based on adding more computers or resources, rather than increasing the power or storage capacity of a single computer; in other words, scaling out rather than scaling up The fundamental idea of Map Reduce is breaking a task down into stages and executing the stages in parallel in order to reduce the time needed to complete the task The first phase of the Map Reduce job is to map input values to a set of key/value pairs as output. The "Map" function accordingly partitions large computational tasks into smaller tasks, and assigns them to the appropriate key/value pairs. Thus, unstructured data, such as text, can be mapped to a structured key/value pair, where, for example, the key could be the word in the text and the value is the number of occurrences of the word. This output is then the input to the "Reduce" function. Reduce then performs the collection and combination of this output, by combining all values which share the same key value, to provide the final result of the computational task The Map Reduce function within Hadoop depends on two different nodes: the Job Tracker and the Task Tracker nodes. The Job Tracker nodes are the ones which are responsible for distributing the mapper and reducer functions to the available Task Trackers, as well as monitoring the results. The Map Reduce job starts by the Job Tracker assigning a portion of an input file on the HDFS to a map task, running on a node. On the other hand, the Task Tracker nodes actually run the jobs and communicate results back to the Job Tracker. That communication between nodes is often through files and directories in HDFS, so inter-node communication is minimized. Hadoop is a MAD system data as files into the disputations on the data. Hadoop loaded into Hadoop simply reduce interprets the data is capable of attracting all that may occur in such After big data is stored, useful insights by performing data analyses will be methods, and follow data analyses. Big Data Analytics Nowadays, people don't juking and importance of the d lyrics is the process of apply useful and unknown patter analytics are used to extract and information from large among the stored variable

III. RELATED WORK

M. Viceconti described five major problems in healthcare data management systems. These are as follows; 1. Working with sensitive Data. 2. Analytics of complex and heterogeneous data spaces, including contextual information. 3. Distributed data management under security and performance constraints. 4. Specialized analytics to integrate bioinformatics and systems biology information with clinical observations at tissue, organ and organisms scales.5. Specialized analytics to define the "physiological envelope" during the daily life of each patient. J. Andreu-Perez provided an overview of recent developments in big data in the context of biomedical and health informatics. Yunchuan promoted the concept of ``smart and connected communities (SCC)", which is evolving from the concept of smart cities. SCC are envisioned to address synergistically the needs of remembering the past (preservation and revitalization), the needs of living in the present (liveability), and the needs of planning for the future (sustainability).X. W. Chen and X. Lin has given a brief overview of deep learning, and highlighted current research efforts and the challenges to big data, as well as the future trends. A. Fahad performed a survey on a comprehensive study of the clustering algorithms proposed in the literature. In order to reveal future directions for developing new algorithms and to guide the selection of algorithms for big data, they proposed a categorizing framework to classify a number of clustering algorithms. The categorizing framework is developed from a theoretical viewpoint that would automatically recommend the most suitable algorithm(s) to network experts while hiding all technical details irrelevant to an application. L. Xu reviewed the privacy issues related to data mining by using a user-role based methodology. They differentiated four different user roles that are commonly involved in data mining applications, i.e. data provider, data collector, data miner and decision maker. A. Belle reviewed that the Big Data focused on three areas of interest: medical image analysis, physiological signal processing, and genomic data processing. V. Sujatha analysed that the data sets from statistical models or complex pattern recognition models may be fused into predictive models that combines data set of patients' treatment information and prognostic outcome results. S. Vennila and J. Priyadarshini., promoted that the security in Big data is a challenging research issue. If Integration of Map Reduce, a machine for privacy preserving, is designed for the analysing of data would provide better privacy. Kovalchuk represented an early stage of the work aimed to the development of a general-purpose concept of the P4 CDSS rising from a treatment-level scope to a hospital-level scope. J. Cunha, C. Silvaa and M. Antunes proposed a generic functional architecture with Apache Hadoop framework and Mahout for handling, storing and analysing big data that can be used in different scenarios. Z. Liu et al., presented an agent-based model of emergency department that was implemented in Netlogo simulation environment. Case studies have been carried out for proving two of the possible uses of the simulator, one to meet the increasing patient arrival overcrowding problem, and the second a quantitative analysis of the influence of ambulance response time (for departure) over the ED behavior. M. Srivathsan and Y. Arjun proposed that Prognotive Computing recognize patterns and formulates its own structure to provide a solution or gives a predicted alert so as to find a solution by ourselves .The System provides a handle of Health care and life span of numerous life forms. A. Abbas stated that they propose a cloud based framework that effectively manages the health related Big-data and benefits from the ubiquity of the Internet and social media. The framework facilitates the mobile and desktop users by offering: (a) disease risk assessment service and (b) consultation service with the health experts on Twitter. F. Zhang proposed a task-level adaptive Map Reduce framework. This framework extends the generic Map Reduce architecture by designing each Map and Reduce task as a consistent running loop daemon. The beauty of this new framework is the scaling capability being designed at the Map and Task level, rather than being scaled from the compute-node level. Y. Wang, L. Kung and T. A. Byrd examined that health care industry has not fully grasped the potential benefits to be gained from big data analytics. K. Kambatla provided an overview of the state-of-the-art and focus on emerging trends to highlight the hardware, software, and application landscape of big-data analytics. J. Wang, M. Qiu and B. Guo developed a tele health system that covers both clinical and nonclinical uses, which not only provides store-and-forward data services to be offline studied by relevant specialists, but also monitors the real-time physiological data through ubiquitous sensors to support remote telemedicine. S. M. DeJong proposed that technology is likely to become increasingly important in healthcare. Any professionalism concerns must be weighed against the potential benefits of technology to patients. P. Nadkarni explained that the Institute of Medicine's idea of a learning health system, in which the boundaries between research and clinical practice are blurred. The historical roots of this idea are identified by exploring initiatives in the business world such as knowledge management, business process reengineering, and enterprise resource planning. M. Legg stated that the standardization required to achieve interoperability for pathology test requesting and reporting. Interoperability is the ability of two parties, either human or machine, to exchange data or information in a manner that preserves shared meaning. A. T. Janke et al., explained that clinical research often focuses on resource-intensive causal inference, whereas the potential of predictive analytics with constantly increasing big data sources remains largely unexplored. Basic prediction, divorced from causal inference, is much easier with big data. L.A. Winters-Miner predicted the development of a healthcare-cantered democracy and seen an explosion in the volume and velocity of patient generated.

IV. Conclusion

In this research, we have examined the innovative topic of big data, which has recently gained lots of interest due to its perceived unprecedented opportunities and benefits. In the information era we are currently living in, voluminous varieties of high velocity data are being produced daily, and within them lay intrinsic details and patterns of hidden knowledge which should be extracted and utilized. Hence, big data analytics can be applied to leverage business change and enhance decision making, by applying advanced analytic techniques on big data, and revealing hidden insights and valuable knowledge. Accordingly, the literature was reviewed in order to provide an analysis of the big data analytics concepts which are being researched, as well as their importance to decision making. Consequently, big data was discussed, as well as its characteristics and importance. Moreover, some of the big data analytics tools and methods in particular were examined. Thus, big data storage and management, as well as big data analytics processing were detailed. In addition, some of the different advanced data analytics techniques were further discussed. By applying such analytics to big data, valuable information can be extracted and exploited to enhance decision making and support informed decisions. Consequently, some of the different areas where big data analytics can support and aid in decision making were examined. It was found that big data analytics can provide vast horizons of opportunities in various applications and areas, such as customer intelligence, fraud detection, and supply chain management. Additionally, its benefits can serve different sectors and industries, such as healthcare, retail, telecom, manufacturing, etc. Accordingly, this research has provided the people and the organizations with examples of the various big data tools, methods, and technologies which can be applied. This gives users an idea of the necessary technologies required, as well as developers an idea of what they can do to provide more enhanced solutions for big data analytics in support of decision making. Thus, the support of big data analytics to decision making was depicted. Finally, any new technology, if applied correctly can bring with it several potential benefits and innovations, let alone big data, which is a remarkable field with a bright future, if approached correctly. However, big data is very difficult to deal with. It requires proper storage, management, integration, federation, cleansing, processing, analysing, etc. With all the problems faced with traditional data management, big data exponentially increases these difficulties due to additional volumes, velocities, and varieties of data and sources which have to be dealt with. Therefore, future research can focus on providing a roadmap or framework for big data management which can encompass the previously stated difficulties. We believe that big data analytics is of great significance in this era of data overflow, and can provide unforeseen insights and benefits to decision makers in various areas. If properly exploited and applied, big data analytics has the potential to provide a basis for advancements, on the scientific, technological, and humanitarian levels.

References

[1] Adams, M.N.: Perspectives on Data Mining. International Journal of Market Research 52(1), 11–19 (2010)

[2] Asur, S., Huberman, B.A.: Predicting the Future with Social Media. In: ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 1, pp. 492–499 (2010)

[3] Bakshi, K.: Considerations for Big Data: Architecture and Approaches. In: Proceedings of the IEEE Aerospace Conference, pp. 1–7 (2012)

[4] Cebr: Data equity, Unlocking the value of big data. in: SAS Reports, pp. 1–44 (2012)

[5] Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J.M., Welton, C.: MAD Skills: New Analysis Practices for

Big Data. Proceedings of the ACM VLDB Endowment 2(2), 1481–1492 (2009)

[6] Cuzzocrea, A., Song, I., Davis, K.C.: Analytics over Large-Scale Multidimensional Data: The Big Data Revolution! In: Proceedings of the ACM International Workshop on Data Warehousing and OLAP, pp. 101–104 (2011

[7] Economist Intelligence Unit: The Deciding Factor: Big Data & Decision Making. In: Capgemini Reports, pp. 1–24 (2012)

[8] Elgendy, N.: Big Data Analytics in Support of the Decision Making Process. MSc Thesis, German University in Cairo, p. 164 (2013)

[9] EMC: Data Science and Big Data Analytics. In: EMC Education Services, pp. 1–508 (2012)

[10] He, Y., Lee, R., Huai, Y., Shao, Z., Jain, N., Zhang, X., Xu, Z.: RCFile: A Fast and Space efficient Data Placement Structure in Map Reduce-based Warehouse Systems. In: IEEE International Conference on Data Engineering (ICDE), pp. 1199–1208 (2011)