International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 12– No.29, May 2020 – www.ijais.org

Big Data Analytics: Significance, Challenges and Techniques

Chatura Chinthana Gamage

ABSTRACT
Big data analytics have been embraced as a novel technology that will reshape domains such as business intelligence, cyber security and economic development that relies on data analytics to gain insights for better decision-making. In recent years, the rapid development of Internet, Information Systems, and Cloud Computing have led to the explosive growth of data in almost every industry and business area. Due to the rapid growth of such data, big data analytics techniques need to be explored and provided in order to process and derive value and knowledge from large datasets. Analysis of these data requires a lot of efforts at multiple levels of knowledge extraction for effective decision making. This paper aims to briefly introduce the concept of big data, analyze some of the different analytics methods and tools which can be applied to big data, as well as critically evaluate the significance of the big data analytics and challenges associated with the application of big data analytics in various decision domains.

General Terms
Data analytics, decision support, knowledge discovery, analytics techniques, pattern recognition

Keywords
Big data, big data analytics, data mining, hadoop, analytical complexity, data visualisation

1. INTRODUCTION
Big data has rapidly developed into an alluring subject that attracts great attention from academia, industry, and even governments around the world in the recent years [1], [2]. Using and mining big data has resulted in a new trend of predication, productivity growth, management information and consumer momentum.

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" [3]. From a macro point of view, big data can be seen as a bond that subtly links and integrates the physical world, the human society, and cyberspace [4]. In this context, big data can primarily be classified into three categories. The first is data from the physical world, which is typically obtained through sensing devices, experiments and observations. The next is data from the human society, which is often acquired from sources or domains such as social networks, health, finance, and transportation and the third is data from the cyberspace which delivers through the Internet.

In the era of big data, large amounts of data affect our work, life and study, even national economic development. It provides a new way of thinking and approaches to analyze and solve problems, which has increasingly emerged as a new ground for research. 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 [5]. These are available in structured, semi-structured, and unstructured format in gigabytes and beyond. 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.

Previous scholars have described big data characteristics using 3Vs to 4Vs where the "Volume" refers to the huge amount of data that are being generated everyday whereas "Velocity" is the rate of growth and how fast the data are gathered for being analysis. The third characteristic "Variety" provides information about the types of data such as structured, unstructured, semi-structured etc. Here, the fourth V refers to the term "Veracity" that includes availability and accountability [6]. Using these, the prime objective of big data analysis is identified as to process data of high volume, velocity, variety, and veracity using various traditional and novel intelligent computational techniques [7] to discover valuable knowledge. Such value can be provided using big data analytics, which is the application of advanced analytics techniques on big data. Big data analytics enforces industries to describe, diagnose, predict, prescribe, and cognate the hidden growth opportunities and leads them toward gaining business value [8]. Big data analytics deploys advanced analytical techniques to create knowledge from exponentially increasing amount of data, which will affect the decision-making process in decreasing complexity of the process [9].

Generally, Data warehouses have been used to manage the large datasets. However, extracting the precise knowledge from the available big data is the foremost issue, such that, most of the presented approaches in data mining are not usually able to handle big data successfully. One of the key problems in big data analytics is the lack of harmonization between database systems with analysis tools for data mining and statistical analysis [6]. These challenges generally arise when performing knowledge discovery and representation, for its practical applications [6]. Even though there is much research on the topic of big data analytics, it is still evolving as a research discipline and not yet established, thus, a clear understanding of its definition, significance and practices is yet to be fully established. At large, there has been a lack of research studies that comprehensively addresses the key challenges of big data, or which investigates opportunities for new applications in various decision domains or emerging and established analytical tools and techniques [10]. Accordingly, the research topic 'Big Data Analytics: Significance, Challenges and Techniques' was considered timely.

This paper aims to introduce the concept of big data, analyze some of the different analytics techniques and tools which can be applied to big data, as well as critically evaluate the significance of big data analytics and challenges perceived for the application of big data analytics in various decision domains. It is expected that this study contributes to the body of knowledge in the subject of big data analytics and benefit the organizations and practitioners.

2. LITERATURE REVIEW

In the information age, enormous amounts of data generated by multiple sources which can be used for informed decision making have become available to the organizations. Big data refers to datasets that are not only big, but also high in variety and velocity, which makes these datasets difficult to process using conventional tools and techniques [11]. Due to the rapid growth of such data, specialised solutions need to be developed and implemented in order to handle and extract valuable knowledge from these huge datasets [11], [12]. Here, the researchers, policy makers and organizations should have the ability to gain valuable insights from such huge, varied and rapidly growing data such as business transactions, customer behaviors and social networks. Studies have shown that such value can be gained using big data analytics, which is the use of advanced analytics techniques to process big data [4], [13], [11].

This Literature Review will first focus on locating the research topic within the discipline of data analytics. Next, it will explore the literature on concept of big data and big data analytics with a special emphasis on analytical methods and tools for big data. Next, after highlighting the findings of the previous studies on major trends big data analytics, the review concludes with emphasising the importance of the research question. This structure is adapted for the review so that it best fits with the author's research question; 'What are the opportunities, challenges and techniques of Big Data Analytics?'.

2.1 Concept of Big Data

In general, big data can be described as a huge and growing set of structured and unstructured data that cannot be handled by using conventional databases, analysis tools or techniques [14], [15], [16], [17]. It is evident that big data carries the potential for delivering business advantages by generating valuable insights into discovering opportunities and risks [18], [19] and is becoming a new technological trend in science, business and policy making [20], [21].

The concept of big data is characterized in the literature by different counts of "V´s" which vary from 3 V´s implied for volume, velocity and variety [22], [14], [23], [24], to 4 V´s which adds the characteristic "value" to the list [22], [25]. Further, extending this model some studies have identified 5V's by including the characteristic of veracity [26], [27], [28]. In literature these characteristics are explained as follows: The first characteristic, the Volume refers to the size or the vast amount of data that are being generated, and the second characteristic Velocity refers to the speed at which data is generated or the rate of growth [22], [24]. Next the third characteristic Variety refers to the different formats in which data is generated such as structured, unstructured, semi-structured [22], [24]. The fourth characteristic value refers to the ability of data to lead insights and better decisions [22], [25]. The fifth characteristic veracity refers to the accuracy of data or the data in doubt which describes the level of uncertainty in data due to inconsistent, latent, ambiguous data and their trustworthiness [22], [28].

2.2 Big Data Analytics

With growing rate of data production in digital era, big data analytics makes use of the data by extracting, transforming, loading, analysing, and preparing the data for decision making [22]. In this context, the large size, wide variety, and rapid change of big data adds analytical complexity and therefore, require a new type of data analytics, as well as different storage and analysis techniques to be properly analyzed for discovering valuable knowledge [11]. Here, the sophisticated techniques that aims to cope with complexity of the big data to extract valuable insights is known as big data analytics [29]. Big data analytics tools support analysis of a wide variety of digital information such as text analytics, audio analytics, video analytics, social media analytics, and predictive analytics. Thus, big data analytics can be considered as the main tool for analysing and interpreting all kinds of digital information [30] including but not limited to transactions, text, audio, video and social media.

Big data analytics needs novel and sophisticated algorithms that process and analyze real-time data which produce results with a high-accuracy and therefore, the technologies such as machine and deep learning allocate their complex algorithms in this process [31]. Big data analytics supports organizations in innovation, productivity, and competition [13] and has been defined as the techniques that are implemented to discover hidden patterns in data and bring insight into interesting relations in various domains by examining, processing, discovering, and visualising the results [32].

Big data analytics can provide advantages for adapting organizations for reducing complexity and managing cognitive burden for knowledge discovery in the data-driven society [33]. Today, business organizations globally have flourished due to the fast evolution of big data analytics and similarly, the governments are increasingly using big data analytics to deliver better services to their citizens [32]. Prior research on big data analytics have demonstrated that, when having right capabilities and approaches, businesses organizations can gain significant value, increase competitive advantage and boost firm performance by making informed decisions based on the results produced by such analysis , [34] [35], [36]. More, it has been found that big data analytics enables business organizations to create a complete view of the behavior of their consumers and also helps them to be more innovative and effective in deploying business strategies [37]. Evidence from big data research shows that, for big data analytics to realise its objectives and to deliver expected outcomes, it requires the right tools and techniques to be analyzed and implemented effectively and proficiently [38]. Thus, for gaining such advantages via big data analytics technologies, factors including skills, technology adaptation and infrastructure capabilities should be considered and developed [22], [35], [39].

2.3 Data Analytics Methods

The advantages of using big data for decision making are significant, however, quite constrained due to the availability of technologies, tools and skills for big data analysis [10]. For informed decision-making, the organizations need to implement efficient and appropriate methods in order to process huge volumes of various data into meaningful comprehensions [40].

In order to obtain value from data, previous studies highlight selective use of a set of analytical methods which shall be applied based on the type and nature of the data and the domain, such as:

1. descriptive analytics is the use of data analytics to examine past and present data and information to define the current state of a business situation in a way that developments, patterns and exceptions become evident, in the form of producing reports and alerts [10], [55], [42];
2. predictive analytics is the application of data analytics that use data and mathematical concepts to for forecasting and statistical modelling to determine the future possibilities based on the changes in the dataset [55], [43];
3. prescriptive analytics describes the application of data analytics using mathematical models to create a set of complex alternatives to find the most suitable solution. This is used for optimization and randomized testing to evaluate how businesses improve their service levels while reducing the expenses [41], [42];
4. diagnostic analytics is about the use of data analytics to investigate the effects and causes of situations and events [41].
5. pre-emptive analytics is the application of data analytics for gaining the ability to take protective actions on events that may have an undesirable effect on the organizational performance. For instance, identifying the possible future risks and recommending mitigating strategies [10], [44]; and
6. inquisitive analytics is about the application of data analytics for examining data to evaluate business proposals, for example, analytical drill downs into data, factor analysis, statistical analysis [10], [45].

Previous research suggests that, these types of analytical methods support for improved decision making and therefore, contribute for enhancing the organizational performance by making decision making process more transparent and accountable, while revealing inconsistencies, potential risks and opportunities [10].

3. RESEARCH METHODOLOGY

The purpose of this chapter is to review the research design appropriateness. This research evaluated the existing research published on the topics, big data and big data analytics, by using a well-established profiling method to investigate and analyze significance and opportunities, different challenges, tools, techniques, and approaches for big data analytics.

This study addresses the research question; 'What are the opportunities, challenges and techniques of Big Data Analytics?' and research objectives;

- Review the significance and opportunities of big data analytics.
- Review the challenges of big data analytics.
- Review the soundness and applications of tools and techniques for big data analytics.

It is vital to have this type of research to foster an in-depth understanding, not only of the big data and big data analytics subject area, but also of the state-of-the-art growth in the application opportunities in big data analytics within different sectors and disciplines, while discussing the prevalent challenges and analytical techniques.

3.1 Data Analytics Methods

This research involved of a comprehensive systematic literature review [46] which focused on identifying and generating detailed insights into the research topic and comprised of search, selection, analysis, and synthesis processes.

It was aimed to select papers that have explicitly included the term ''big data" in the title, abstract, keywords, or body of the paper and have an emphasis on the adoption, implementation, or use of data analytics technologies by various organizations, and optionally which have conversed the opportunities, challenges and techniques for big data analytics.

The main selection process involved two rounds. In the first round, the researcher inspected the papers in order to check whether the term ''big data" had been mentioned in the title, abstract, keywords and the body of the text. In the second round, the papers were primarily judged based on the use of keywords such as challenges, opportunities, and techniques as focal concepts in an organizational context.

The analysis primarily focused on analysing the existing perceptions on big data analytics, highlighting prevailing findings related to this topic, and identifying supporting evidence in the literature. Further, the paper primarily aimed to highlight the new insights and techniques that can contribute to the future research and therefore, moved further than simply mapping or describing the current discourse.

4. DISCUSSION

4.1 Significance and Opportunities of Big Data Analysis

Recent times have greatly increased the ability to gather huge amounts of data, introducing an opportunity to produce transformative changes in the way data is analyzed and interpreted [13]. Due to its great value and potential, big data has been fundamentally changing and transforming the way we live, work, and think [47]. This section discusses in detail the significance and opportunities of big data in various perspectives.

4.1.1 Big Data in National Development and Policy Making

"Data is the lifeblood of decision making and the raw material for accountability" [48]. Advances in computer infrastructure and data-science engineering in recent years have made it possible to process and analyze big data in real-time [48], [49]. At the national level, the capacity of accumulating, processing, and utilizing vast amounts of data has become a new landmark of a country's strength [4]. Novel insights gathered from such data analytics can complement government's survey data and official statistics, adding depth and tone to the information on people's behaviors, experiences and expectations.

In the context of big data, the scope of this information is vast, and bigdata applications can facilitate policymaking for national development in the countries that would otherwise require dedicated intensive and continuous human and financial resources. Further, it is possible for countries to use

big data applications for various domains including national security and intelligence, government tax administration, tourism and healthcare development. Policy makers could leverage the improvements in the modern big data techniques to gain real-time insights into people's wellbeing and to target aid interventions to vulnerable groups. New sources of data such as satellite data together with new technologies, and new analytical approaches, if applied properly, can enable more agile, efficient and evidence-based decision-making and can better measure progress on the Sustainable Development Goals in a way that is both inclusive and fair [48].

4.1.2 Big data for business model improvement and innovation

Big data has become a new point of economic growth [50]. With big data, companies have started to upgrade and transform in order to introduce shared service models such as Analysis as a Service (AaaS) for providing a better approach for delivering business intelligence, thereby changing the ecology of the information and communication technology (ICT) and other industries. In this context, the many major players in the ICT industry have already introduced several big data analytics solutions spanning from on-premise solutions to cloud-based solutions, which eliminates the need of huge on-premise hardware setups for big data.

Further, it is evident that the startup organizations are able to more easily adapt new data-driven business models when compared to the already established organizations [51] and use this ability as a competitive advantage to enter into different industries. Therefore, incumbent organizations need to rethink and redefine their prevailing business models and how these may be affected and modified by big data [26]. The organizations can leverage the power of big data by tapping new data sources and techniques, where they can improve the existing business processes in terms of productivity, efficiency and effectiveness [50],[52],[53]. Further supporting this, Loebbecke and Picot (2015) suggest that ''incremental enhancements to established business models through increased digitization and big data analytics may replace less efficient business models (and thereby companies) in the long run" [51]. This approach suggest that organizations can take advantage of big data analytics while generally continuing to function in the same manner, only more effectively and efficiently [50].

Organizations can use insights from big data analytics also for innovating new business models in developing novel value propositions, targeting different customer segments, or in changing the ways of interacting with customers. Business organizations can improve customer experience for their products and services using customer relationship management applications which can provide insights into products and services which customers are likely to buy. Such customer relationship management applications which have integrated with big data analytics solutions are often capable of providing a single view which can improve marketing campaign efficiency, drive better customer retention, create new cross-sell and up-sell opportunities, and provide more light on things such as customer lifetime value [50]. These abilities are possible because with big data, organizations have the opportunity to gain a more complete and accurate view of customers and their behaviors by incorporating wide variety of data sources such as social media interactions, call recordings, email conversations, feedbacks, purchase history records, search history records and forum interactions.

4.1.3 Big data for better predict the future

Advances with big data analytics together with machine learning technologies help researchers and organizations to look into the future in a more accurate and reliable manner and make predictions about future trends by analysing patterns. For such analysis, the organizations use predictive analytics to navigate through current and historical big datasets to detect trends and forecast events and conditions that may probably occur at a defined time in future.

With predictive big data analytics, organizations can discover and exploit patterns which exist within large sets of data in order to identify future opportunities and detect risks in advance which helps adopters in a variety of domains such as finance, retailing, airlines, hospitality, healthcare, automotive, pharmaceuticals and manufacturing. For instance, by using big data analytics business organizations can forecast their inventory requirements and reorder levels, manage delivery schedules and design store layouts in order to maximize profits. Also, the organizations in the airlines industry widely use predictive bigdata analytics for pricing tickets and travel packages using past travel patterns, seasonal and global trends. Further, the business organizations in the hospitality industry such as hotels, restaurants and tour operators have the ability to use big data analytics applications to forecast reservation levels on different time periods in order to maximize their occupancy and revenue. Further, the organizations have the ability to use insights from predictive bigdata analytics to prevent and reduce destructive activities by identifying unusual patterns in user activities including fraud, corporate spying and information security breaches such as cyber-attacks. Also, wide adaptation of big data analytics by diverse organizations is evident to predict future economic conditions, risk management, climate trends, infrastructure maintenance and investment needs, and is becoming more and more prevalent with the improvements in the big data analytics technology.

4.1.4 Big data for regulatory compliance for preventing fraud and reporting risk

Financial institutions are frequently being pressured with variety of compliance and reporting requirements to tighten up their risk management and fraud prevention frameworks. Regulatory agencies do not impose these compliance requirements to suppress business operations of financial institutions. Instead, these regulators are enacted to ensure the stability, resiliency and sustainability of the financial system.

Here, the financial institutions can take advantage of big data analytics applications for real time identification and reporting of fraudulent behavioral patterns based on transactional data and historical data. Big data analytics enables financial institutions to deal with these highly volatile transaction data in real time with the possibility of taking immediate action to stop, hold or report suspicious transactions. Thus, big data analytics is the backbone of the technology that the institutions in the financial industry can use to comply with various regulatory requirements, including those that relate to financial crimes, money laundering and fraud.

Big data analytics can help financial institutions to build comprehensive compliance reports and perform regulatory stress tests, dramatically reducing compliance analytic cycle times. Further, it can help financial institutions to not only report aggregate client and financial reporting data, but also allows usage of a wider breadth of diverse data sources such

as email, voice to understand what is underpinning the credible differences as seen by the financial auditors. Thus, by adopting big data analytics for compliance, financial institutions can enhance efficiency and compliance across their business lines.

4.2 Challenges of Big Data Analytics

Many challenges exist in the process of leveraging the potential of big data, starting from the systems design of the big data processing systems at the lower layers to data analysis and presentation at the higher layers. Some of these challenges are caused by current analysis techniques, while some others are caused by the characteristics of big data and by the limitations of data processing systems currently in use. This section discusses the major issues and challenges of big data analysis.

4.2.1 Data complexity

Big data refers to datasets that are high in variety and velocity, which makes them complex to handle [5]. Diversified types and patterns, complicated inter-relationships, and greatly varied data quality are the typical characteristics of big data. The inherent complexity of big data resulting from complex types, complex patterns and structures makes observation, representation, understanding and computation of big data far more challenging and results in a significant increase in the computational complexity when compared to traditional analytical tools and techniques used for processing typical datasets. As a result, traditional data analysis and interpretation tasks, such as data retrieval, semantic analysis, topic-discovery and sentiment analysis, become particularly challenging when using big data. All these greatly limit the capacity to design efficient and effective computational models and methods for gaining valuable insights and solving problems using big data.

In this context, quantitatively describing the essential characteristics of the complexity of big data is challenging. The study on complexity theory of big data will help in understanding the characteristics, the formation of complex patterns in big data, simplify its representation, get better knowledge abstraction, and guide the design of computing models and algorithms for big data processing [4]. Additionally, data reduction, data selection, feature selection can be considered as essential tasks especially when dealing with large datasets and presents an unprecedented challenge for researchers and analysts. This is due to the fact that the existing algorithms for big data analytics may not always respond in an adequate time when dealing with these high dimensional data. Thus, automation of the processes for big data analytics, developing new data processing techniques and machine learning algorithms to ensure consistency has become a major challenge in recent years.

4.2.2 Data Storage complexity

The size of data available in different domains has grown exponentially during the recent years by various means such as enterprise resource planning systems, mobile devices, automated production systems, telco systems, social networks and demands huge storage needs. Therefore, one of the initial challenges for big data analysis is very high capacity storage mediums with higher operational efficiency and speed. Here, the accessibility of the data must be on the top priority for the knowledge discovery and representation and must be able to access promptly and easily for detailed analysis.

Here, the key requirements of big data storage are that, it should be able to handle extremely large amounts of data and should be able to keep scaling to keep up with growth, and should be able to provide higher amount of input/output operations per second (IOPS) necessary to deliver data to the analytics tools. In past decades, organizations and analysts used hard disk drives to store data which has a slower IOPS performance. However, in the recent years, storage solutions with high performance and high capacity which are based on the storage concepts such as solid-state drive (SSD) and phase change memory (PCM) have been introduced to overcome this limitation. Nevertheless, it is observed that the available storage technologies may be still not sufficient to achieve the required performance for processing big data [4].

4.2.3 Computational and Analytical complexity

Effectively analysing the big data sets for obtaining better knowledge is a significant challenge for researchers and big data analysts. The key features of big data, namely, huge volume, high velocity, high variety and veracity, make it difficult for traditional computing methods to effectively support the processing, computation, analysis and interpretation of big data [4]. In addition, knowledge discovery, representation and visualisation is also a major challenge in analysis of big data. Thus, it is required that the new approaches for big data analytics to focus on big data centric, novel and highly efficient computing models and provide innovative methods for processing and analysing big data, and support value driven applications in different domains [4], [5]. New features in big data processing, such as data wrangling, open and uncertain data relationships and uneven distribution of value density provide great opportunities but also introduce significant challenges, in the context of computability of big data and the development of new computing paradigms.

For addressing the computational complexity of big data applications it's required to study data centric computing models based on the characteristics of big data. Further, it is required to develop algorithms for distributed computing and introduce a computing framework based on big data where storage, communication and computing operations are smoothly integrated and optimized [4]. More, it is required to explore existing reduction-based computing methods which provides for reducing big data on demand, from huge datasets to being just enough, and to being valuable enough [4]. Furthermore, designing algorithms for machine learning to analyze data is necessary for enhancing efficiency and scalability of the analytical processes. However, it may be difficult to establish a comprehensive mathematical model that is broadly applicable to Big Data analytics. A standard process here is transforming the semi-structured or unstructured data into structured data, and then applying the data mining algorithms to extract knowledge and insights [54]. The analysis of domain specific big data can be performed by understanding the specific challenges and complexities applicable for those domains. Previous studies show that a significant amount of research and survey has been carried out in this direction with the objective of minimizing the computational cost of big data processing and minimizing the complexities [55], [56].

4.2.4 System complexity

Information technology systems capable of handling a variety of data types, volumes and applications are the key to supporting scientific research of big data analytics. If big data systems cannot be used to provide or forecast critical business

decisions or provide the insights into business values, which are buried under huge amount of data at the right time when needed, then these systems lose their relevance. These big data analytical systems are complex by nature as they should support to handle huge volumes of data, complex structures, high computational complexity, long duty cycles, and real-time processing requirements. These requirements make it challenging to design the computing frameworks, system architectures and processing systems and also introduce severe constraints on their usability, operational efficiency and energy consumption. Answering these problems should lay the standards for designing, implementing, testing, and optimizing systems for big data processing and should form a solid foundation for designing and developing hardware and software system architectures with energy optimized and efficient distributed storage and processing capabilities. It's possible to develop systems for big data analysis with a high throughput in data acquisition, low energy consumption, and highly efficient computing through an agile process of design, implementation, and validation which focus on applying required mathematical and analysis models that can address the particular complexities.

4.2.5 Security complexity

In general, it is observed that the security mechanisms in big data technology is weak [57]. If not properly handled, big data architecture can lead to certain critical security implications, while undergoing storage, partition, replication and distribution among thousands of data processing nodes for distributed computation. Among the security concerns of big data, insecure computation can not only cause information leakages but also may cause corrupt data, leading to incorrect results in analysis and decision making.

An additional challenge to big data solutions is that users may retrieve sensitive information out of the data using ad-hoc queries. This is due to the fact that, as data is stored at thousands of nodes authentication, authorization and encryption of data at these nodes becomes challenging. Further, producing analytical results involves multiple challenges such as privacy exposures, intrusive marketing and unintentional disclosure of confidential information. Typically, big data solutions collect input from variety of data sources and therefore, it is important and essential to validate the input. This involves detecting whether the data is trusted and identifying the trusted and untrusted data sources in order to filter inaccurate or malicious data from the good data.

It is possible to enhance the security of big data by designing and using security centric big data architectures where techniques of authentication, authorization, and encryption is rationally implemented for specific big data problems. However, developing a multi-level security, privacy preserved data model for big data remains a major challenge.

4.3 Big Data Analysis Tools and Techniques

Today, there are a large number of tools available for big data analytics. This section discusses some current tools and techniques that are widely used for big data analytics with an emphasis on five important analytical tools namely, Apache Hadoop, Apache Spark, SAS, R Programming and Apache Storm. It is observable that, majority of the tools available for big data analytics focus on batch processing, steam processing and visual and interactive analysis [6]. Here, stream processing is mostly used when real time analytics is required. Currently, most batch processing tools used in the industry are based on the Apache Hadoop framework while, Apache Strom and Apache Splunk are widely used for large scale streaming platforms. Further, the interactive analysis processes offered by some of the tools allow users to directly interact in real-time for their own analysis. For instance, Microsoft Power BI, Qlik, Tableau, SAS Visual Analytics, Dremel and Apache Drill are some of the key platforms that support visual and interactive analysis. These tools make it possible to design and implement effective and efficient big data projects.

4.3.1 Apache Hadoop

The Apache Hadoop software library is a big data processing framework which allows for distributed processing of large datasets across groups of computers [58]. Hadoop has been designed to be highly scalable and it can be scaled up from a single server to thousands of servers, where each node offers storage and local processing. The Hadoop framework has been designed to detect and handle failures at the application layer instead of depending on the hardware to deliver high availability. This has enabled delivering a service with high availability on top of a cluster of computers which may be vulnerable to hardware failures. Apache Hadoop framework consists of modules namely, Hadoop Common, Hadoop Distributed File System (HDFS), YARN, Ozone, and MapReduce [58]. Here, MapReduce is an important module for parallel processing of large data sets based on divide and conquer method, where the divide and conquer method is applied in two steps namely, Map step and Reduce Step. Today, a large number of companies and organizations in various domains use Apache Hadoop for both research and production as it is the most established software platform for big data analysis [6].

4.3.2 Apache Spark

Apache spark is an open source big data processing framework by the Apache Software Foundation built for boosting performance of sophisticated big data analytics applications [59]. However, this can also be used for disk-based conventional data processing when datasets are too large to fit into the existing system memory. Due to its in-memory data processing engine it can provide fast access to data in SQL workloads enabling it to very quickly process data [59]. It's in-memory, distributed and iterative computation is principally valuable when dealing with machine learning algorithms and provides for data cleansing and transformation, model building, feature engineering. Apache Spark can process data from a range of data sources, including the Hadoop Distributed File System (HDFS), relational databases and NoSQL databases [59]. It is notable that the functions in this tool are mainly broken down into data transformations and data aggregations. It also provides an interface for programming entire clusters where users can run the application programs in different languages such as Java, R and Python. Spark has been widely adopted by organizations that work with big data analytics due to its processing speed as well as its ability to support multiple types of databases and ability to run various analytical application programs.

4.3.3 SAS (Statistical Analysis Systems)

SAS (Statistical Analysis Systems) is a product suite created by SAS Institute that performs multivariate investigations, advanced analytics, data management, business intelligence, and various different duties [60] and is not an open source

software. Data management on the SAS Platform includes several components that are fully enabled for the Hadoop ecosystem. SAS has a variety of tools for in-memory and in-database computing, including SAS Visual Analytics and SAS Visual Statistics, SAS In-Memory Statistics for Hadoop, SAS High-Performance Analytics, SAS Scoring Accelerator and SAS In-Database technologies, and SAS Data Loader for Hadoop [60]. More, SAS Federation Server simplifies data access, administration, security and performance by creating a virtual data layer without physically moving data. SAS platform supports every phase of the big data analytics life cycle starting from data, to discovery, to deployment [61].

4.3.4 R Programming

R programming language is an open source free software environment designed for statistical computing and graphics. The R language is widely used for developing statistical software and data analysis by data analysts and statisticians for data visualisation, data cleansing and exploratory data analysis (EDA) to obtain insights from data. It is also possible to use R programming for parallel or cluster computation using Apache Spark (SparkR). In the application for big data analytics, the main constraint with R is that R can run only on in-memory data by default. However, to overcome this limitation, it is possible to use R with big data by using techniques such as down-sampling data, compressing data before moving into R and chunking of data which enables each data portion to be pulled separately and processed by R serially or in parallel.

4.3.5 Apache Storm

Apache Storm is a free and open source distributed computation system with a real-time stream processing model, which can process the unbounded streams of data very fast. It has been specially designed for real-time processing of data, in contrasts with Hadoop, which is designed for batch processing [49]. Apache Strom is easy to use and has a high fault tolerance and reliability level. More, this tool is highly scalable where it can process the data in parallel through a cluster of machines. Further, Apache Storm provides a clean architecture to build applications called Topologies where it enables developers to build their logic virtually in any programming language, which supports communication over a JSON-based protocol [62]. Further, with the graph analytics offered by Storm, companies can easily determine the linkages between many different data points simplifying the task of linking people, places, processers, services, times, and products, and can speed times and improve clarity for gaining business insights.

4.3.6 Advanced Data Visualisation Tools

Advanced Data Visualisation is one of the strongest potential developments among big data analytics tools and has surfaced as a compelling technique to discover knowledge from data [11]. It helps organizations and analysts to group many data points, to understand various relationships, concentrate on questions in real time and quickly determine research focus [63]. Further, it enables data engineers to discover hidden data patterns and possible methods for processing. [63].

Today, organizations are able to use, Advanced Data Visualisation tools which blends with data analysis methods with interactive visualisation, to enable comprehensive and dynamic data discovery. These tools for Advanced Data Visualisation are helpful in many business analysis situations and also fits well in situations where analysts have little knowledge about the data as it is a data driven exploratory approach [64]. It is evident that with huge volume and velocity of data generated and due to the increased complexity of data structures, a growing demand has developed for Advanced Data Visualisation solutions from many organizational domains [11], [65]. Such visualisation analyzes enables decision makers to comprehensively analyze data at both the summarised and the detailed levels by taking advantage of human cognitive and reasoning abilities. Due to the adaptation of interactive statistical graphics and simplified user-friendly interfaces, Advanced Data Visualisation tools can enable speedier analysis, better decision making, and more effective presentation and comprehension of results [11].

5. CONCLUSION

In recent years, with the explosive growth of data, big data has made a strong impact in almost every industry and business area. This paper evaluates the significance of big data, challenges and some key tools and techniques for big data analytics. Accordingly, the literature was examined in order to provide an insight into the concept of big data, its characteristics and the big data analytics methods which are being researched, as well as their importance to decision making. In addition, some of the key big data analytics tools and techniques were examined. Thus, this research has provided the data analysts, organizations and big data solution developers, with an insight on significance and challenges of the big data analytics and various big data tools and techniques which can be adopted to successfully implement big data solutions. By adopting and implementing such solutions for big data analytics, valuable knowledge can be discovered and scrutinized to enhance informed and cognitive decision making which can bring significant improvements in terms of efficiency, productivity, profitability and compliance.

6. REFERENCES

[1] Brunswicker, S., Bertino, E. & Matei, S. 2015, ‘Big Data for Open Digital Innovation - A Research Roadmap’, Big Data Research, 2 (2), pp. 53-58.

[2] Cuzzocrea, A. 2014, ‘Privacy and Security of Big Data: Current Challenges and Future Research Perspectives’, Proceedings of the Conference on Information and Knowledge Management, pp. 45-47.

[3] Wikipedia. 2019, Big data. [Online]. Available at: https://en.wikipedia.org/wiki/Big_data. [Accessed 23 October 2019]

[4] Jin, X., Wah, B. W., Cheng, X. & Wang, Y. 2015, ‘Significance and challenges of big data research’. Big Data Research, 2 (2), pp. 59-64.

[5] Das, S., Bhuyun, U. C., Panda, B. S. & Patro, S. 2016, ‘Big Data Analysis and Challenges’. International Journal of Engineering and Management Research, 6 (5), pp. 203-207.

[6] Acharjya, D. P. & Kauser, A. P. 2016, ‘A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools’. International Journal of Advanced Computer Science and Applications, 7(2), pp. 511-518.

[7] Kakhani, M. K., Kakhani, S. & Biradar, S. R. 2015, ‘Research issues in big data analytics’. International Journal of Application or Innovation in Engineering & Management, 2(8), pp.228-232.

[8] Wang, H., Wang, W., Zhou, X., Sun, H., Zhao, J., Yu, X. & Cui, Z. 2017, 'Firefly algorithm with neighborhood attraction'. Information Sciences, 382 (383), pp. 374-387.

[9] Maulik, U. & Bandyopadhyay, S. 2000, 'Genetic Algorithm-Based Clustering Technique', Pattern Recognition, 33(9), pp. 1455-1465.

[10] Sivarajah, U., Mustafa, M., ZahirIrani, K. & Weerakkody, V. 2017, 'Critical analysis of Big Data challenges and analytical methods', Journal of Business Research, 70, pp. 263-286.

[11] El-Gendy, N. & Elragal, A. 2014, 'Big Data Analytics: A Literature Review Paper', ICDM.

[12] Singh, A. 2015, 'Data Mining Techniques, Applications and Scope', International Journal of Engineering and Management Research, 5 (2), pp. 358-365.

[13] Zhou, Z., Chawla, N. V., Jin, Y. & Williams, G. J. 2014, 'Big data opportunities and challenges: discussions from data analytics perspectives [discussion forum]', IEEEComput. Intell. Mag., 9 (4), pp. 62-74.

[14] Batra, S. 2014, 'Big Data Analytics and its Reflections on DIKW Hierarchy', Review of Management, 4 (1/2), pp. 5–17.

[15] Chaudhary, R., Pandey, J. R., & Pandey, P. 2015, 'Business model innovation through big data', Proceedings of the 2015 International Conference on Green Computing and Internet of Things IEEE, pp. 259-263.

[16] Chen, H., Chiang, R., & Storey, V. 2012, 'Business Intelligence and Analytics: From Big Data to Big Impact', Management Information Systems Quarterly 36 (4), pp. 1165-1188.

[17] Ebner, K., Buhnen, T., and Urbach, N. 2014, "Think Big with Big Data: Identifying Suitable Big Data Strategies in Corporate Environments," in Proceedings of the 47th Annual Hawaii International Conference on System Sciences, pp. 3748–3757.

[18] McAfee, A., & Brynjolfsson, E. 2012. 'Big Data: The Management Revolution', Harvard Business Review 90 (10), pp. 60–68.

[19] Shim, J., French, A., Guo, C., & Jablonski, J. 2015, 'Big Data and Analytics: Issues, Solutions, and ROI', Communications of the Association for Information Systems, 37 (1), pp. 797-810.

[20] Bedi, P., Jindal, V., & Gautam, A. 2014, 'Beginning with big data simplified', 2014 International Conference on Data Mining and Intelligent Computing, pp. 1-7.

[21] Wood, J., Andersson, T., Bachem, A., Best, C., Genova, F., Lopez, D. R. & Vigen, J. 2010, 'Riding the wave: How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data A submission to the European Commission', European Commission.

[22] Fernando, F., Engel, T. 2018, 'Big Data and Business Analytic Concepts: A Literature Review', Twenty-fourth Americas Conference on Information Systems, New Orleans

[23] Russom, P. 2011, 'Big Data Analytics'. TDWI Best Practices Report (Fourth Quarter). TDWI.

[24] Singh, D. S., Singh, G., 2017, 'Big data - A Review', International Research Journal of Engineering and Technology 4 (4), pp 822-824.

[25] Liao, Z., Yin, Q., Huang, Y., & Sheng, L. 2014, 'Management and application of mobile big data', International Journal of Embedded Systems, 7(1), pp. 63-70.

[26] Gillon, K., Aral, S., Lin, C., Mithas, S. & Zozulia, M., 2014, 'Business analytics: radical shift or incremental change?' Communications of the Association for Information Systems, 34 (13), pp. 287-296.

[27] Vajjhala, N. R. & Ramollari, E. 2016, 'Big Data using Cloud Computing - Opportunities for Small and Medium-sized Enterprises', European Journal of Economics and Business Studies, 2 (1), pp. 130-138.

[28] Baesens, B., Bapna, R., Marsden, J., & Vanthienen, J. 2016, 'Transformational Issues of Big Data and Analytics in Networked Business', Management Information Systems Quarterly, 40 (4), pp. 807-818.

[29] Ning, J., Zhang, Q., Zhang, C., Zhang, B. 2017, 'A best-path-updating information-guided ant colony optimization algorithm', Information Sciences. 433 (434), pp. 142-162.

[30] Koonce, D. & Tsaib, S. 2000 'Using data mining to find patterns in genetic algorithm solutions to a job shop schedule', Computers & Industrial Engineering, 38(3) pp. 361-374.

[31] Harfouchi F., Habbi, H., Ozturk, C. & Karaboga, D. 2017, 'Modified multiple search cooperative foraging strategy for improved artificial bee colony optimization with robustness analysis', Soft Computing, 22(08), pp. 6371-6394.

[32] Wang, H. et al. 2017, 'Randomly attracted firefly algorithm with neighborhood search and dynamic parameter adjustment mechanism', Journal of Soft Computing, 21(18), pp. 5325-5339

[33] Mishra, N., Lin C. & Chang, H. 2015, 'A cognitive adopted framework for IoT big data management and knowledge discovery prospective', International Journal of Distributed Sensor Networks, 11(10), pp. 1-13

[34] Davenport, T. H., & Harris, J. G. 2007, 'Competing on analytics: The new science of winning', Harvard Business Press.

[35] Akter, S., Wamba, S. F., Gunasekaran, A., Dubey, R., & Childe, S. J. 2016. 'How to improve firm performance using big data analytics capability and business strategy alignment?', International Journal of Production Economics, 182 (December), pp. 113-131.

[36] Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J. f., Dubey, R., & Childe, S. J. 2017, 'Big data analytics and firm performance: Effects of dynamic capabilities', Journal of Business Research, 70, pp. 356-365.

[37] Delice, Y., Aydogan, E., Özcan, U. & İlkay, M. S. 2017, 'A modified particle swarm optimization algorithm to mixed-model two-sided assembly line balancing',

Journal of Intelligent Manufacturing, 28(1) pp. 23-36.

[38] Al Nuaimi, E., Al Neyadi, H., Mohamed, N., & Al-Jaroodi, J. 2015, 'Applications of big data to smart cities', Journal of Internet Services and Applications, 6(1), pp. 1-15.

[39] Garmaki, M., Boughzala, I., and Wamba, S. F. 2016, 'The Effect of Big Data Analytics Capability on Firms Performance', Proceedings of the 20th Pacific Asia Conference on Information Systems.

[40] Gandomi, A., & Haider, M. 2015, 'Beyond the hype: Big data concepts, methods, and analytics', International Journal of Information Management, 35(2), 137-144.

[41] Baum, J., Laroque, C., Oeser, B., Skoogh, A. & Subramaniyan, M. 2018, 'Applications of Big Data analytics and Related Technologies in Maintenance - Literature Based Research', Machines, 6 (54), pp. 1-12.

[42] Joseph, R. C., & Johnson, N. A. 2013, 'Big data and transformational government', IT Professional, 15(6), pp. 43-48.

[43] Waller, M. A., & Fawcett, S. E. 2013, 'Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management', Journal of Business Logistics, 34(2), pp. 77-84.

[44] Szongott, C., Henne, B., & von Voigt, G. 2012, 'Big data privacy issues in public social media', Sixth IEEE international conference on digital ecosystems technologies, pp. 1-6.

[45] Bihani, P., & Patil, S. T. 2014, 'A comparative study of data analysis techniques', International Journal of Emerging Trends & Technology in Computer Science, 3(2), pp. 95-101.

[46] Webster, J., and Watson, R. 2002, 'Analyzing the Past to Prepare for the Future: Writing a Literature Review', Management Information Systems Quarterly, 26 (2), pp. xiii–xxiii.

[47] Mayer-Schonberger V. & Cukier, K., 2013, 'Big Data: A Revolution That Will Transform How We Live, Work, and Think', Reprint, Eamon Dolan/Mariner Books, United States, 2014.

[48] The United Nations. 2019. Big Data for Sustainable Development. [Online] Available at: https://www.un.org/en/sections/issues-depth/big-data-sustainable-development/index.html. [Accessed 2 November 2019].

[49] Althaf, R. S., Sai, R. .K. & Girija, R. K., 2018, 'Challenging tools on Research Issues in Big Data Analytics', International Journal of Engineering Development and Research, 6 (1), pp. 637-644.

[50] Günther, W. A., Mohammad H.Rezazade Mehrizi, M. H. R., Huysman, M. & Feldberg, F. 2017, 'Debating big data: A literature review on realizing value from big data', The Journal of Strategic Information Systems, 26 (3), pp. 191-209.

[51] Loebbecke, C. & Picot, A., 2015, 'Reflections on societal and business model transformation arising from digitization and big data analytics: a research agenda', The Journal of Strategic Information Systems, 24 (3), pp. 149-157.

[52] Ghoshal, A., Larson, E. C., Subramanyam, R., Shaw, M. J., 2014, 'The impact of business analytics strategy on social, mobile, and cloud computing adoption', Proceedings of the Thirty Fifth International Conference on Information Systems, Auckland, New Zealand, December, pp. 14–17.

[53] Woerner, S. L. & Wixom, B. H. 2015, 'Big data: extending the business strategy toolbox', Journal of Information Technology, 30 (1), pp. 60-62.

[54] Ali, G. & Nithya, A. 2017, 'Challenges and Open Research Issues and Tools on Big Data Analytics', International Journal of Advanced Research in Computer Engineering & Technology, 6 (11), pp. 1690-1703.

[55] Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & K. Taha, K. 2015, 'Efficient machine learning for big data: A review', Big Data Research, 2 (3), pp. 87-93.

[56] Yoo, C., Ramirez, L. & Liuzzi, J. 2014, 'Big data analysis using modern statistical and machine learning methods in medicine', International Neurology Journal, 18, pp. 50-57.

[57] Bhandari, R., Hans, V. & Ahuja, N. J. 2016. 'Big Data Security - Challenges and Recommendations', International Journal of Computer Sciences and Engineering, 4 (1), pp. 93-98

[58] The Apache Software Foundation. 2019, Apache Hadoop. [Online] Available at: https://hadoop.apache.org/. [Accessed 15 November 2019].

[59] The Apache Software Foundation. 2019, Apache Spark. [Online] Available at: https://spark.apache.org/. [Accessed 22 November 2019].

[60] Wikipedia. 2019, SAS (software). [Online] Available at: https://en.wikipedia.org/wiki/SAS_(software). [Accessed 17 November 2019].

[61] SAS. 2019, Big Data Analytics: What it is and why it matters. [Online] Available at: https://www.sas.com/en_us/insights/analytics/big-data-analytics.html. [Accessed 25 November 2019].

[62] Iqbal, M. H. & Soomro, T. R. 2015, 'Big Data Analysis: Apache Storm Perspective', International Journal of Computer Trends and Technology, 19 (1), pp. 9-14.

[63] Chawla, G., Bamal, S. & Khatana, R. 2018, 'Big Data Analytics for Data Visualization: Review of Techniques', International Journal of Computer Applications, 182 (21), pp. 37-40.

[64] Shen, Z., Wei, J., Sundaresan, N., & Ma, K. L. 2012, 'Visual Analysis of Massive Web Session Data', Large Data Analysis and Visualization, pp. 65-72.

[65] Zhang, L., Stoffel, A., Behrisch, M., Mittelstadt, S., Schreck, T., Pompl, R., Weber, S., Last, H. & Keim, D. 2012, 'Visual Analytics for the Big Data Era - A Comparative Review of State-of-the-Art Commercial Systems', IEEE Conference on Visual Analytics Science and Technology, pp. 173-182.