Big Data Analytics: Characteristics, Tools and Methods, Data Analysis techniques, Challenges, Security and its Applications - An Overview

Vijayabhaskar .V

Assistant Professor, Dept. of Science and Humanities,
R.M.K. Engineering College, Kavaraipettai – 601 206, India.
vvbhaskar75@gmail.com

Dr. Pavai Madheswari .S

Professor and Head , Dept. of Mathematics,
R.M.K. Engineering College, Kavaraipettai – 601 206, India.
pavai.arunachalam@gmail.com

Abstract - Big data is currently a buzzword in both academia and industry, with the term being used to describe a broad domain of concepts, ranging from extracting data from outside sources, storing and managing it, to processing such data with analytical techniques and tools. This paper work thus aims to provide a review of current big data analytics concepts in an attempt to highlight big data analytics' importance to decision making.

Due to the rapid increase in interest in big data and its importance to academia, industry, and society, solutions to handling data and extracting knowledge from datasets need to be developed and provided with some urgency to allow decision makers to gain valuable insights from the varied and rapidly changing data they now have access to. Many companies are using big data analytics to analyze the massive quantities of data they have, with the results influencing their decision making. Many studies have shown the benefits of using big data in various sectors, and in this paper, various big data analytical techniques and tools are discussed to allow analysis ofthe application of big data analytics in several different domains.

Keywords: Big Data, Characteristics, Tools and Methods, Challenges, Security and Big Data Applications.

I. INTRODUCTION - BIG DATA

Big data generally refers datasets that have grown too large for and become too difficult to work with by means of traditional tools and database management systems. It also implies datasets that have a great deal of variety and velocity, generating a need to develop possible solutions to extract value and knowledge from wide-ranging, fast-moving datasets [15].

According to the Oxford English Dictionary, "Big data" as a term is defined as "extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions". [4] argued that this definition does not give the whole picture of big data, however, as big data must be differentiated from data as being difficult to handle using traditional data analyses. Big data thus inherently requires more sophisticated techniques for handling complexity, as this is exponentially increased.

II. BIG DATA CHARACTERISTICS

Based on the various big data definitions, it is obvious that the size is the dominating characteristic despite other characteristics' importance. [25] proposed the three V's as the dimensions of challenge to data management, and the three V's constitute a common framework [25] [9]. These three dimensions are not independent of each other; if a one-dimension change, the probability of changing another dimension also increases [16].

A further two dimensions are often added to the big data characteristics, *veracity* and *variability* [16] as shown in Figure 1. The five V's reflect the growing popularity of big data. The first V is, as always, volume, which is related to the amount of generated data[18]. The second V is for the velocity (big data timeliness), as all data collection and analysis should be conducted in a timely manner [9]. The third V refers to variety, as big data comes in many different formats and structures such as ERP data, emails and tweets, or audio and video [15][38]. The fourth V refers to big data's "huge value but very low density", causing critical problems in terms of extracting value from datasets [15][9][32]. The fifth V refers to veracity, and questions big data credibility where the sources are external, in most cases [2] [3] [18].

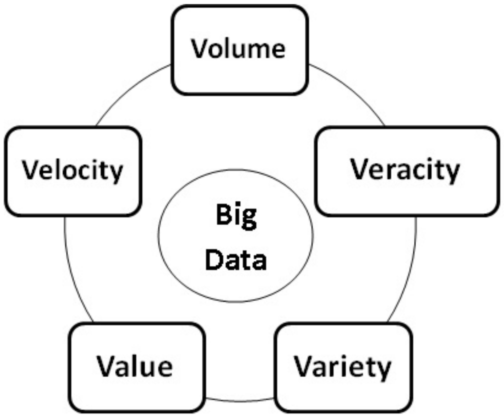

Figure 1: Big data in terms of the 5 V's.

Multi-dimensional data can be used to add historical context to big data. The variety of big data is as important as its volume, while velocity or speed can describe how difficult big data may be to handle. Velocity may refer to data generation frequency or data delivery frequency. Depending on data inconsistency, incompleteness, ambiguity, latency, deception, and approximations, big data quality can also be characterized as undefined, good, or bad [12].

Volume represents the sheer size of the dataset due to the aggregation of a large number of variables and an even larger set of observations for each variable. **Velocity** reflects the speed at which data are collected and analyzed, whether in real time or near real time from sensors, sales transactions, social media posts, and sentiment data for breaking news and social trends. **Variety** in big data comes from the plurality of structured and unstructured data sources such as text, videos, networks and graphics among others. **Veracity** ensures that the data used are trusted, authentic and protected from unauthorized access and modification. **Value** represents the extent to which big data generates economically worthy insights and/or benefits through extraction and transformation. **Variability** concerns how insight from media constantly changes as the same information is interpreted in a different way, or new feeds from other sources help to shape a different outcome. **Visualization** can be described as interpreting the patterns and trends that are present in the data [28].

3Vs: Volume, Velocity, Variety.

4Vs: Volume, Velocity, Variety, Veracity.

5Vs: Volume, Velocity, Variety, Veracity, Value.

7Vs: Volume, Velocity, Variety, Veracity, Value, Variability, Visualization.

III. BIG DATA ANALYTICS (BDA): TOOLS AND METHODS

The most difficult problem that needs to be solved to handle big data effectively is storage; it is not necessarily easy to deal with large quantities and varieties of data[15][41].

There are many big data storage and analysis models. Where the large amount of data is caused by the sheer variety of users and devices, a data centre may be necessary for storing and processing the data. Establishing network infrastructure is necessary to help gather this rapidly generated data, which is then sent to the data centre before being accessed by users .

Research by [40] identifies the components of the network that must be established, such as an original data network, the bridges used for connecting and transmitting to data centres, and at least one data centre.

Structured data storage and retrieval methods include "relational databases, data marts, and data warehouses" [15]. Data is extracted from outside sources, then transformed to fit operational needs, and finally loaded into the database. The data is then uploaded from the operational data store to longer-term storage using Extract, Transform, Load (ETL) or Extract, Load, Transform (ELT) tools. The data is then cleaned, transformed, and catalogued before use [5][15]. A big data environment requires analysis skills, unlike the Enterprise Data Warehouse (EDW) traditional environment [19].

The big data environment accepts and demands all possible data sources. On the other hand, EDW approaches data sources with caution, as it is more streamlined towards supporting structured data [15][19]. Due to increasing number of data sources and data analyses possible, big data storage requires agile databases to give analysts the opportunity to produce and adapt to data easily and quickly [15][19].A big data repository must be deep, allowing analysts to analyze the datasets deeply by using complex statistical methods [15][19].

Hadoop is a popular big data analytics framework. Hadoop "provides reliability, scalability, and manageability by providing an implementation for the MapReduce paradigm as well as gluing the storage and analytics together"[15]. Hadoop includes HDFS which is for the big data storage and MapReduce for big data analytics, and it can process extremely large amount of data by dividing the data into smaller blocks, then specifying datasets to be distributed across cluster nodes [32][15]. Hadoop incorporates several technologies: "Hive is a data warehouse implementation for Hadoop, MapReduce is a programming model in Hadoop, and Pig is a querying language for Hadoop which has similarities to the SQL language for relational databases" [42].

First-generation technology generated the Apache Spark project in software terms [38], but Hadoop has a great deal more power, which offers advantages to analytics in terms of memory. It can work with both batch and real-time workloads, is easy to program with Java code, and can connect to Apache projects and other software within a closed ecosystem. Hadoop's components are [38]: Spark SQL runs SQL-like queries on structured data, Spark streaming provides real-time data processing, MLib provides a machine learning library of algorithms and utilities and Graph X provides application algorithms.

IV. BIG DATA ANALYTICS PROCESSING

Analytics processing is the next issue after big data storage. According to [20], big data analytics processing has four critical requirements:

a) *Fast data loading*: limited interference between disk and network, to speed up query execution.

b) *Fast query processing*: workloads are heavy, therefore real-time requests should be processed as quickly as possible to satisfy user requirements. The data placement structure should also have the ability process multiple queries as query volumes increase.

c) *Highly efficient utilization of storage space*: as user activities grow rapidly, they need scalable storage capacity and computing power. As disk space is limited, it is necessary to manage data storage during processing and address the space issues adaptively.

d) *Strong adaptivity to highly dynamic workload patterns*: the underlying system should be highly adaptive, as data processes have different workload patterns and the analyzing of big datasets has many different applications and users, with different purposes and methods [15].

The work presented by [17] shows that using big data frameworks for storing, processing, and analyzing data has changed the context of knowledge discovery from data, mainly in terms of data mining processes and pre-processing, with a particular focus on the rise of data pre-processing in cloud computing. The presented solution covered various data pre-processing technique families with factors such as maximum size supported examined in terms of big data and data pre-processing throughout all of the families of methods. Moreover, various big data framework such as Hadoop, Spark, and Flink were discussed.

2.1Big data analytics

Big data growth continues apace, and many organizations are now interested in managing and analyzing data. Organizations trying to benefit from big data are adopting big data analytics to facilitate faster and better decisions, as it is not easy to analyze datasets with analysis techniques and infrastructure based on traditional data management [10]. The need for new tools and methods specialized for big data analytics is thus also growing. The emergence of big data is affecting everything from data itself to its collection and processing, and, finally, the extracted decisions. Providing big data tools and technologies can help in managing the growth of network-produced data, which is otherwise exponential, as well as in increasing the capability of organizations to scale and capture the required data to reduce database performance problems [15].

Big data analytics: a means to analyze and interpret any kind of digital information. The technical and analytical advancements in BDA, which in large part determine the functional scope of today's digital products and services, are crucial for the development of sophisticated artificial intelligence, cognitive computing capabilities and business intelligence. Big data analytics: technologies (e.g. database and data mining tools) and techniques (e.g. analytical methods) that a company can employ to analyze large-size, complex data for various applications intended to augment firm performance in various dimensions. Big data analytics, defined as tools and processes often applied to large and disperse datasets for obtaining meaningful insights, has received much attention in IS research given its capacity to improve organizational performance. Big data analytics is defined as the application of multiple analytic methods that address the diversity of big data to provide actionable descriptive, predictive and prescriptive results. Big data analytics is the statistical modeling of large, diverse and dynamic datasets of user-generated content and digital traces [28].

In this section, various big data analyses will be discussed, beginning with the data analysis techniques available and some of the common big data analytics suites, finally discussing several big data platforms and tools. Data analysis techniques can be characterized into four types, as shown in Figure 2

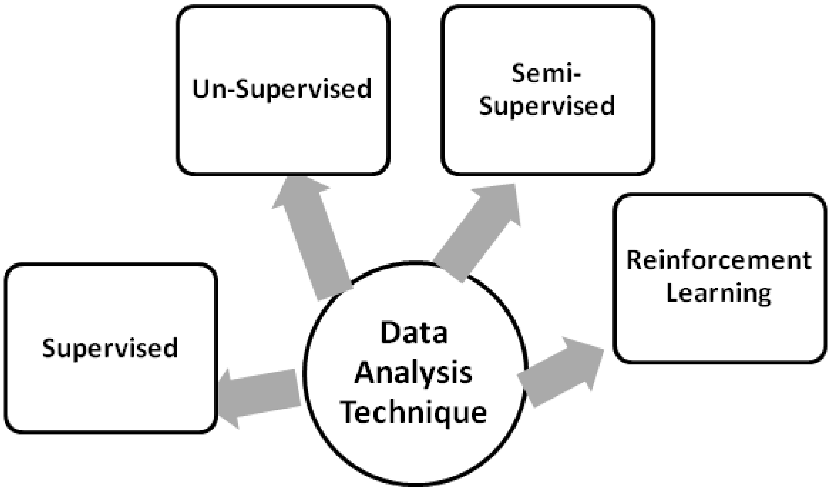

Figure 2: Data analysis techniques

2.1.1. Supervised techniques

A supervised technique refers to where data are trained and tested, and the training data is labeled. Labeled means that the full history of what has happened to the data is known, and thus the history for the data variables is known.

Supervised learning involves training a system based on labeled data and this requires a supervisor with the ability to expect the output from each input that can train the system according to its expectations. When the system is trained, it can give predictions within “many applications of classification and fault detection and channel coding and decoding” [23][11][44]. This technique is used for approximating a function between the input and output.

The idea is for the system to learn the training dataset's classifiers (the labeled documents) then to automatically apply this classification to an unknown dataset's un-labeled documents. This learning technology thus involves learning from example [6][7][30].

Regression is an example of the supervised learning algorithm, as are Linear Regression, Decision Trees (DT), Support Vector Machine (SVM), K Nearest Neighbour (K-NN), Naive Bayes Classifier (NBC), Random Forest, and neural networks (NN). However, many of these supervised techniques cannot be used with wireless networks, and as the learning techniques are dependent on the data training, the results are also restricted [11].

Regression Analysis: is mathematical tool used to discover correlations between several variables based on experimental or observed data. Where analysis defines the relationships between variables as non-random, such analysis may make the correlations between variables appear simpler and more regular [26], as shown in Figure 3.

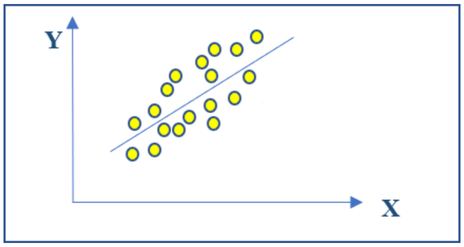

Figure 3: Regression analysis

Structured data mostly utilizes predictive analytics, and this overshadows other analytics forms for 95% of big data [16]. However, new statistical techniques for big data have emerged which clarify the differentiation of big data from smaller data sets. In practice, however, most statistical methods were designed for smaller datasets, in particular, samples.

Usually, scientists make predictions based on theories in the prediction domain. However, big data analytics can deliver predictions that depend on the sequence of data processing and execution. According to [22][30].

- big data brings new challenge as it is generated from different system sources. The data retrieved from each source system should thus be sent to a central repository;
- the relationship between operations should be defined to allow reconstruction of datasets from multiple sources;
- the knowledge discovery process should be automated from data or datasets to make predictions;
- generating new theories is required to create and improve models. Predicted target theory generates a set of predictors; however, some theories explain the relationships between independent and dependent predictors more effectively;
- there is a shift from theory-driven to process-driven prediction based on analyzing the BDA steps and identifying the challenges, theoretically informing future BDA needs throughout data acquisition, pre-processing analysis, and interpretation.

2.1.2. Un-supervised techniques

Here, the training data is unlabelled. Unlabelled means that the history of the data is missing, there is no history available for data variables, and the data have not been trained and tested. Thus, unsupervised techniques require separate training data [6][7][30].

Unsupervised learning requires deducing functions for presenting unknown structures from unlabelled data. This technique does not require a supervisor, which means that the system must have the ability to proceed independently with training based on unlabelled data input [44] [11].

Examples of unsupervised learning algorithms include clustering algorithms, combinatorial algorithms, A priori algorithms, Self-Organizing Maps (SOM), and applications of game theory. These techniques are used for classifying the input data into different clusters or classes based on the data distribution [11][21].

Cluster Analysis: This method is based on grouping objects and classifying them depending on shared features. It is used for differentiation between objects to allow division into clusters. Thus, data which are related to each other or have the same features will be placed in a cluster or a group and unrelated data will be in other groups [11] [39], as shown in Figure 4.

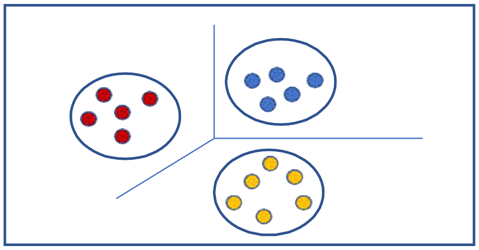

Figure 4: Cluster analysis.

2.1.3. Semi-supervised techniques

Where some of the data is labeled and some is unlabelled, supervised and unsupervised techniques can also be mixed. Algorithms are applied for both labeled and unlabelled data, and even with incomplete information or missing training sets, some of the dataset's classifiers can be learned. Both supervised and unsupervised techniques focus on one aspect (target separation or independent variable distribution, respectively), and using them together may thus give better results [7] [44].

2.1.4. Reinforcement learning (RL)

Reinforcement learning involves setting and classifying real-time data changes in a way that allows the learning framework to adapt based on those changes [11][39].

The components of an RL algorithm are the agent; the environment; and the actions. The actions are taken by the algorithm based on the environment, and depending on the feedback from the environment, it determines whether the action is positive, thus using it again in future, or negative, thus discarding it. An example of reinforcement learning is Markov Chains (Markov Decision Process) [30]. The difference between RL and supervised or unsupervised learning is that RL works based on the feedback which is either good or not depending on the situation and is hence dynamic, while supervised and unsupervised learning give static solutions[11] [44].

The RL process includes an actor which acts in the environment with its own copy of the data; the data can thus be stored in a separate replay memory and sampled by the learner to be computed within the policy parameters. The actor learners then receive the updated policy parameters [29].

The Map-Reduce framework was utilized by for parallelizing batch reinforcement learning methods with linear function approximation [29]. Applying parallelism helped speed up large matrix operations but did not assist the collection of experience or stabilize learning.

V. ANALYTICS TECHNIQUES

5.1. Correlation Analysis

This is an analytical method used to determine the relationships such as "correlation, correlative dependence, and mutual restriction, among observed phenomena and accordingly conducting forecast and control" [9], as shown in Figure 5. Positive correlation on the left means, while one variable increases so does the other. No linear correlation on the middle means there is no visible relationship between the variables. Negative correlation on the right means as one variable increases, the other decreases [9].

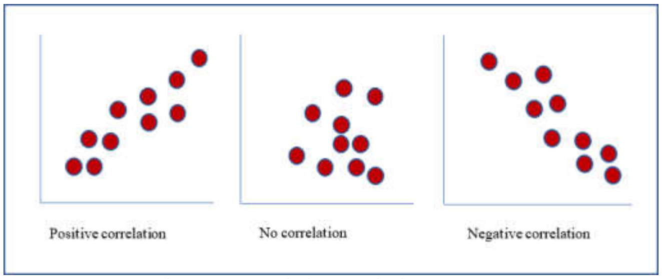

Figure 5: Correlation Analysis.

5.2. Text Mining

This converts the content from unstructured text to structured text in order to help uncover the meaning and the information contained.

5.3. Factor Analysis

This groups several related variables into a single factor, which means that fewer factors are used in analysis, which is thus simpler.

The research presented in [34] examines the state-of-the-art in big data at that time and discusses research agendas. In addition, it defines the basic technology and toolsets used. It is not easy to analyze datasets with traditional data management techniques [10]; therefore, new methods and tools have been developed for big data analytics, as well as for storing and managing such data. These solutions thus need to be studied in terms of handling datasets and extracting knowledge and value. In addition, the rapid changes in data volume, variety, velocity, and value require decision makers to know how to obtain valuable insights.

Finally, advanced data visualization is becoming an important analysis tool, as this enables faster and better decision making [15]. Some of the more common models and analyses are explained further below, and shown in Figure 6.

- **Text analytics:**
  - ➢ **Sentiment Analysis:** This is based on understanding the subjects' emotions from their text patterns to help in organizing viewpoints into good or bad, positive or negative. This analysis helps firms by alerting themwhere customers are dissatisfied or seeking to shift to other products, allowing preventative actions to be taken [15].
- **Audio analytics or speech analytics using technical approaches:**
  - ➢ **LVCSR**: large-vocabulary continuous speech recognition, indexing and searching.
  - ➢ **Phonetic-based systems:** work with sounds or phonemes [16].
- **Social media and social network analysis (SNA):** Social media depends on multiple tools and frameworks for collecting, monitoring, summarizing, analyzing, and visualizing social media data, and SNA depends on social entities' relationships with each other to measure the knowledge linking parties, including who shares information, what information, and with whom. SNA tries to get develop network patterns, while social media tries to uncover useful patterns and user information using text mining or sentiment analysis [15][16].
- **Data Visualization**: This can be used even by decision makers with little knowledge about the data, as it presents the information visually prior to deep analysis. Advanced Data visualization (ADV) offers strong potential growth to big data analytics as it allows analysis of data at several levels by taking advantage of human perceptual and reasoning abilities [15].
- **Predictive analytics**: This is based on statistical methods such as associative rules, clustering, classification and decision trees, regression, and factor analysis [7].

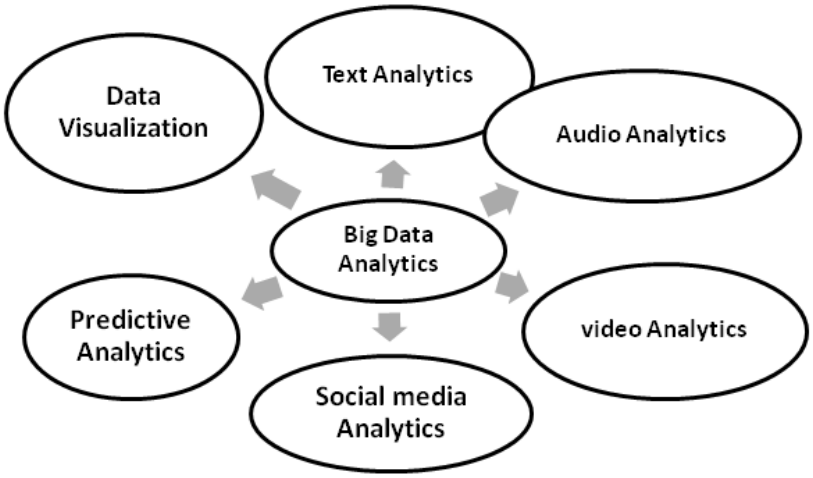

Figure 6: Common big data analytic methods.

The other types of big data analytics used for systematic review are presented by [18], and these include descriptive analytics, diagnostic analytics, predictive analytics, and prescriptive analytics.

VI. BIG DATA PLATFORMS AND TOOLS

There are now multiple big data analytics tools and the study done by [31] showed the importance of carefully choosing the right tool for the circumstances. The choice is dependent on the “nature of datasets (i.e., volumes, streams, distribution), the complexity of analytical problems, algorithms and analytical solutions used, systems capabilities, security and privacy issues, the required performance and scalability in addition to the available budget” (ibid). Some of big data platforms and tools are

6.1. Apache Mahout

This is an open source machine learning software library that can be used for executing algorithms via MapReduce, a framework for processing large datasets. Mahout encompasses several Java libraries, ensuring efficiency of processing large datasets by allowing application of large-scale machine learning applications and algorithms. It provides an optimized algorithm in which Mahout converts machine learning tasks presented in Java into MapReduce jobs [1].

6.2. R

This is a programming language often used for big data analysis, which offers relatively easy solutions to performing advanced analysis on large data sets via Hadoop. As compared to Mahout, in term of types and algorithms, R provides a more complete set of classification models; however, it is limited by its nature as an object-oriented programming language, which can cause problems with memory management compared to other solutions. In many cases, use in combination with Mahout is thus recommended [35], as R can be used to execute small data exploration while Hadoop/Jsql executes the larger operations.

6.3. Alteryx

This tool offers data blending and an advanced analytics platform where analysts can merge internal business processes, third-party tools, and cloud data centres. Also, it allows data analytics utilizing some tools in a single workflow [36].

6.4. Google Cloud Platform (GCP)

It is one of the leaders among cloud Application Programming Interfaces (APIs). Despite the fact that it was established a few years ago, GCP has realized a significant growth since it suits the public cloud services that are based on massive, solid infrastructures. It gives the developer the ability to build a range of programs starting from simple websites to complex world-wide distributed applications. GCP platform contains a set of physical assets (e.g., computers and hard disk drives) and virtual resources (e.g., virtual machines, a.k.a. VMs) hosted in Google’s data centres around the globe [8].

6.5. H2O

It is an open source framework offering parallel processing, analytics, math, and machine learning libraries beside data pre-processing and assessment tools. Furthermore, it offers a web-based user interface that eases its use by analysts and statisticians who have limited programming backgrounds. It also provides support for Java, R, Python, andScala [24].

6.6. MicroStrategy

It provides an integrated big data analytics platform where the data is stored in Hadoop clusters and the users are given permission to access the desktop computer and mobile devices. This tool offers real-time visualization and interactions to implement fast decisions [36].

6.7. RapidMiner

It is a programming-free data analysis platform. It provides the user with the ability to "design data analysis processes in a plug-and-play fashion by wiring operators". It allows importing operators for various data formats (e.g., Excel, CSV, XML). It prepares a set of operators for massive datasets with further attributes from open data sources which give an advantage for better predictive and descriptive models [33].

6.8. Datameer

Datameer Analytics Solution (DAS) is a business integration platform for Hadoop. It contains data source integration, “an analytics mechanism with a spreadsheet interface”, designed with analytic functions and visualization to help business users in reports, charts and dashboards. Datameer can bring data from both structured such as Oracle, IBM DB2, and unstructured sources such as Twitter, Facebook, LinkedIn or e- mails [14].

6.9. Microsoft

Microsoft platform provides predictive analytics capability called SSAS and integrated in the SQL Server. This platform offers "efficiency in Azure’s cloud data source’s integration and deployments as a web service" also, the simplicity of utilizingfor data scientists [36].

Figure 7 is adopted from [32] and shows 1) data sources; the big data states that need to be processed and transformed; 3) big data tools and platforms wherein these decisions are made depending on the inputs, tool selection, and analytical models chosen; and 4) the big data analytics applications.

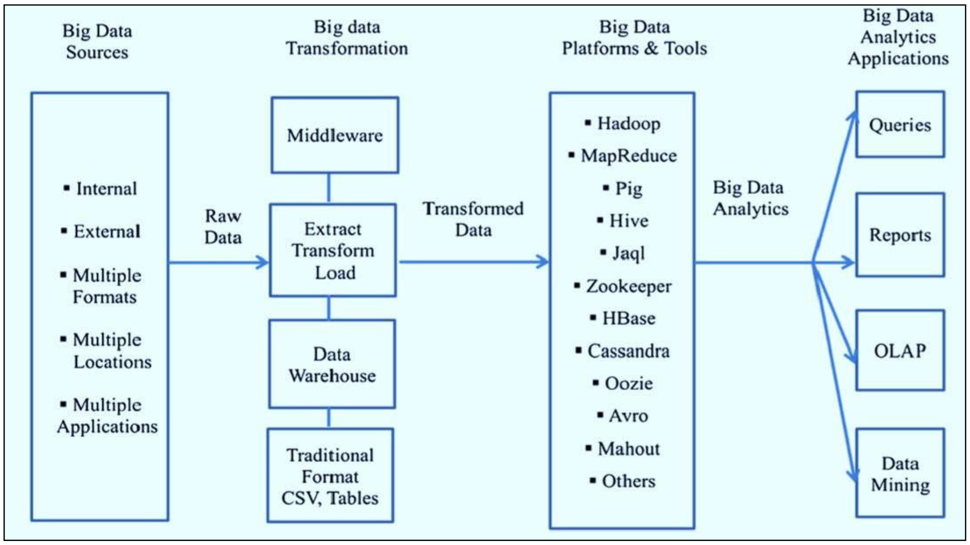

Figure 7: An applied conceptual architecture of data analytics, adopted from[32].

VII . BIG DATA ANALYTICS CHALLENGES

Many studies have focused on the use of analytics techniques such as data mining, visualization, statistical analysis, and machine learning; however, there is a need to develop new analytic approaches in order to handle big data challenges such as the time required for processing when the volume of the data is very large [31]. Thus the difficulties are presented in applying current analytical solutions, including machine learning, deep learning, incremental approaches, and granular computing.

Similarly addressed big data applications, opportunities, and challenges, and examined several techniques to handle big data challenges, such as cloud computing and quantum computing, to examine their efficacy. big data overview that included four categories: 1) concepts, big data characteristics, and processing paradigms; (2) state- of-the-art techniques for decision making in big data; (3) decision making applications of big data in social science; and (4) big data’s current challenges and future directions[9].

VIII. DATA SECURITY ISSUES

In public affairs, privacy, internet access disparities, and legal and security issues are key concerns, and managers and policymakers in these areas should work to overcome these limitations. Public managers and policymakers are also, however, generally working under the restrictions of a limited budget, multiple constituencies, and short time frames for extracting knowledge big data [27][18].

Some security issues with big data and gave some suggestions for avoiding big data security risks. The security concern inherent in big data include the fact that big data comes from many different sources, some of which may have weak security as well as a variety of formats and large volumes. Any security breaches may thus affect multiple companies and result in financial losses, and thus, appropriate actions should be taken to reduce such big data security risks [38].

IX. BIG DATA ANALYTICS APPLICATIONS

Big data analytics has the potential to be applied to demand forecasting, analyzing potential needs based on previous work used to classify analytics techniques. Big data analytics have also been applied in many areas, serving different sectors.

9.1. Healthcare

Big data might help in reducing waste and improving efficiency in clinical operations, research and development, and public health by means of statistical tools and algorithms; predictive modeling to produce new drugs and devices more quickly; analyzing records of diseases to improve epidemiology [15]; allowing faster development of vaccines; and identifying the data relevant to provide services and prevent crises.

[32] described big data analytics in healthcare and identified several remaining challenges; big data analytics has the power to develop care, save lives, and minimize the costs, using the recent data explosion to extract insights in order to allow healthcare providers to make better decisions. The potential benefits gained from using big data in healthcare include, but are not limited to, discovering diseases quickly, thus making treatment easier and more effective; identifying healthcare fraud quickly in order to manage specific individuals; and improving population health.

9.2. Banking

Handling massive volumes of data of many different types is not easy. Big data analytics offers potential benefits to industries such as banking by allowing analysis of customer log files and the handling of customer interactions. Combining structured and unstructured data types in this way can give companies a better view of both their customers and operations [13].

9.3. Retail

Big data analytics has a massive impact on retail industries, improving the customer experience and reducing fraud [37].Big data analytics provides these organizations with more information on market decisions and help in segmenting customer based on their characteristics. Social media analytics can also be used to inform companies about what their customers prefer. Applying sentiment analysis to such data provides the organization with early warnings when the customer turns to different products, allowing action to be taken by the organization [15].

9.4. Telecommunications

Big data analytics can improve the quality of management in telecommunications by making use of real-time data analyses and monitoring machine logs. Predictive analytics can also be used to minimize performance variability and to prevent quality issues by providing early warning alerts [15].

X. CONCLUSION AND FUTURE RESEARCH

This paper presented data analysis techniques characterized in four sections: supervised, unsupervised, semi-supervised, and reinforcement learning. Some analytics techniques were also presented, such as clustering, correlation, regression, and factor analytics, and some big data tools and platforms such as Hadoop, Apache Mahout, and R were explained in relation to these. Big data storage, management, and analytics processing were also discussed, and some emergent advanced data analytics techniques further examined.

Various big data tools, methods, and technologies have been discussed in this research, offering readers examples of the necessary technologies, and prompting developers to come up with ideas about how to provide additional big data analytics solutions to help in decision making.

Big data analytics has been applied in various areas, serving many different sectors. Big data analytics has the potential to improve care, save lives, and reduce costs in the healthcare sector. It also benefits industries such as financial institutions by allowing analysis of customer log files to help develop a better understanding of customer needs. The retail sector has a significant impact on society and usingbig data analytics in this sector can again help managers to better understand people’s needs, thus prompting the development of better services. Big data analytics are also used in the telecommunications sector, where they help in monitoring machine logs and addressing quality issues.

Some big data analytics challenges were discussed in this work, particularly with regard to security and privacy. Some examples of how big data analytics can be used to handle issues such as intrusion detection and big data characteristics such as size, velocity, variety, value and external sources were also given. Finally, some real-world big data analytics applications were introduced.

REFERENCES

[1]. Acharjya, D.P. and Ahmed, K., A survey on big data analytics: challenges, open research issues and tools. *International Journal of Advanced Computer Science and Applications,* pp.511- 518, 2016.

[2]. Addo-Tenkorang, R. and Helo, P.T., Big data applications in operations/supply-chain management: A literature review. *Computers & Industrial Engineering journal,* Volume 101, pp. 528-543, 2016.

[3]. Al-Barashdi, H. and Al-Karousi, R., Big Data in academic libraries: literature review and future research directions.. *Journal of Information Studies and Technology,* p. 13, 2019.

[4]. Arunachalam, D., Kumar, N. and Kawalek, J.P., Arunachalam, D., Kumar, N. anUnderstanding big data analytics capabilities in supply chain management: Unravelling the issues, challenges and implications for practice.. *Transportation Research Part E: Logistics and Transportation Review journal,* Volume 114, pp. 416—436, 2018.

[5]. Bakshi, K., 2012. *Considerations for big data: Architecture and approach conference.* s.l., IEEE, pp. (1-7).

[6]. Boyd-Graber, J., Mimno, D. and Newman, D.,. Care and feeding of topic models: Problems, diagnostics, and improvements.. *Handbook of mixed membership models and their applications Journal ,* Volume 225255, 2014

[7]. Breed, D.G. and Verster, T., An empirical investigation of alternative semi-supervised segmentation methodologies. *South African Journal of Science,* Volume 115, pp. pp.92-98, 2019.

[8]. Challita, S., Zalila, F., Gourdin, C. and Merle, P., *A Precise Model for Google Cloud Platform..* s.l., IEEE, pp. 177-183, 2018 .

[9]. Chen, H., Chiang, R.H. and Storey, V.C., Business intelligence and analytics: from big data to big impact.. MIS quarterly, pp. 1165-1188, 2012.

[10]. Constantiou, I.D. and Kallinikos, J.,. New games, new rules: big data and the changing context of strategy. *Journal of Information Technology,* Volume 30, pp. 44-57, 2015

[11]. Cui, Q., Gong, Z., Ni, W., Hou, Y., Chen, X., Tao, X. and Zhang, P.,. Stochastic Online Learning for Mobile Edge Computing: Learning from Changes.. *IEEE Communications Magazine journal,* Volume 57, pp. 63-69, 2019

[12]. Data, D.B., A Practical Guide to Transforming the Business of Government.. *TechAmerica Foundation" s Federal Big Data Commission Journal,* 2012.

[13]. Davenport, T.H. and Dyché, J., Big data in big companies. *International Institute for Analytics,* 2013.

[14]. Di Martino, B., Aversa, R., Cretella, G., Esposito, A. and Kołodziej, J., Big data (lost) in the cloud.. *International Journal of Big Data Intelligence,* Volume 1, pp. 3-17, 2014.

[15]. Elgendy, N. and Elragal, A., *Big data analytics: a literature review paper.* s.l., Springer, cham, pp. 214-227, 2014.

[16]. Gandomi, A. and Haider, M., Beyond the hype: Big data concepts, methods, and analytics.*International Journal of Information Management,* Volume 35, pp. 137-144, 2015.

[17]. García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J.M. and Herrera, F., Big data preprocessing: methods and prospects.. *Big Data Analytics Journal,* p. 9, 2016.

[18]. Grover, P. and Kar, A.K., Big data analytics: a review on theoretical contributions and tools used in literature.. *Global Journal of Flexible Systems Management,* 18(3), pp. 203-229, 2017.

[19]. Hartmann, T., Fouquet, F., Moawad, A., Rouvoy, R. and Le Traon, Y., GreyCat: Efficient what-if analytics for data in motion at scale.. *Information Systems journal,* 2019.

[20]. He, Y., Yu, F.R., Zhao, N., Yin, H., Yao, H. and Qiu, R.C., Big data analytics in mobile cellular networks. *IEEE access,* Volume 4, pp. 1985-1996, 2016.

[21]. Jiang, C., Zhang, H., Ren, Y., Han, Z., Chen, K.C. and Hanzo, L., Machine learning paradigms for next-generation wireless networks.. *IEEE Wireless Communications journal,* Volume 24, pp. 98-105, 2017.

[22]. Kitchin, R.,. Big Data, new epistemologies and paradigm shifts.. *Big Data & Society Journal,* p. 2053951714528481, 2014.

[23]. Kotsiantis, Sotiris B and Zaharakis, I and Pintelas, P,. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering journal,* 160(2), pp. 3—24, 2007.

[24]. Landset, S., Khoshgoftaar, T.M., Richter, A.N. and Hasanin, T., A survey of open source tools for machine learning with big data in the

Hadoop ecosystem.. *Journal of Big Data,* Volume 2, p. 24, 2015.

[25]. Laney, D., 3D data management: Controlling data volume, velocity and variety.. *META group research note,* Volume 6, p. 1, 2001.

[26]. Lei, Y., Jia, F., Lin, J., Xing, S. and Ding, S.X., An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data.. *IEEE Transactions on Industrial Electronics journal,* Volume 36, pp. 3137-3147, 2016.

[27]. Mergel, I., Rethemeyer, R. K., & Isett, K., Big data in public affairs. *Public Administration Review Journal,* pp. 928-937, 2016.

[28]. Mikalef, P., Pappas, I.O., Krogstie, J. and Giannakos, M., Big data analytics capabilities: a systematic literature review and research agenda.. *Journal of Information Systems and e-Business Management,* Volume 3, pp. 547-578, 2018.

[29]. Mnih, Volodymyr and Kavukcuoglu, Koray and Silver, David and Rusu, Andrei A and Veness, Joel and Bellemare, Marc G and Graves, Alex and Riedmiller, Martin and Fidjeland, Andreas K and Ostrovski, Georg and others, Human-level control through deep reinforcement learning. *Nature journal,* Volume 518, p. 529, 2015.

[30]. Mouthami, K., Devi, K.N. and Bhaskaran, V.M., *Sentiment analysis and classification based on textual reviews..* s.l., IEEE., 2013.

[31]. Müller, O., Junglas, I., Brocke, J.V. and Debortoli, S., Utilizing big data analytics for information systems research: challenges, promises and guidelines. *European Journal of Information Systems,* pp. 289-302, 2016.

[32]. Oussous, A., Benjelloun, F.Z., Lahcen, A.A. and Belfkih, S., Big Data technologies: A survey. *Journal of King Saud University-Computer and Information Sciences,* pp. 431-448, 2018.

[33]. Raghupathi, W. and Raghupathi, V.,. Big data analytics in healthcare: promise and potential. *Health information science and systems,* p. 3, 2014.

[34]. Ristoski, P., Bizer, C. and Paulheim, H., Mining the web of linked data with rapidminer., 2015.

[35]. Schelén, O., Elragal, A. and Haddara, M.,. *A roadmap for big-data research and education.,*s.l.: Luleå tekniska universitet., 2015

[36]. Team, R.C., R language definition.. *R foundation for statistical computing.,* 2000.

[37]. ur Rehman, M.H., Chang, V., Batool, A. and Wah, T.Y., Big data reduction framework for value creation in sustainable enterprises. International Journal of Information Management, 36(6), pp.917-928.. *International Journal of Information Management,* Volume 36, pp. 917-928, 2016.

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

[39]. Watson, H., Tutorial: Big data analytics: Concepts, technologies, and applications..*Communications of the Association for Information Systems CAIS,* Volume 34, p. p.65, 2014.

[40]. Wu, Celimuge and Yoshinaga, Tsutomu and Chen, Xianfu and Zhang, Lin and Ji, Yusheng, Cluster-based content distribution integrating LTE and IEEE 802.11 p with fuzzy logic and Q- learning. *ieee Computational intelligenCe magazine journal,* pp. 41-50., 2018.

[41]. Yi, X., Liu, F., Liu, J. and Jin, H., Building a network highway for big data: architecture and challenges. *IEEE Network journal,* Volume 28, pp. 5-13, 2014.

[42]. Zhong, R.Y., Newman, S.T., Huang, G.Q. and Lan, S., Big Data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives. *Computers & Industrial Engineering journal,* pp. 572-591, 2016.

[43]. Zuech, R., Khoshgoftaar, T.M. and Wald, R.,. Intrusion detection and big heterogeneous data: a survey.. *Journal of Big Data,* p. 3, 2015

[44]. Dr. K. Rajasekaran., and P. Saravanan, Conceptual Methodology on Machine Learning and Types of Learning Algorithms.. *JAC : A JOURNAL OF COMPOSITION THEORY, Volume XIII, Issue V, pp.* 234-235, 2020.