Analysis of Big Data

Diya Deepak Makwana

Student, Dept. of Information Technology, Nagindas Khandwala College, Mumbai, India

---

Abstract - *Big Data is a term used to describe vast assemblages of data sets that are replete with knowledge. The important element in the market nowadays is extremely big data sets that can be computationally analyzed to uncover patterns, trends, and associations from unstructured data into structured ones to discover a solution for a firm. Despite the significant operational and strategic effects, little empirical study has been done to determine the business value of big data. With the goal of creating usable information from big data, big data analytics is fast becoming a popular method that many organizations are adopting. This paper offers a thorough examination of big data, including its characteristics, its applications, and the big data analytics techniques employed by various businesses to aid in decision-making. The paper also discusses various big data tools currently in use.*

Key Words: Big Data (BD), Big Data Analytics (BDA), Hadoop, MongoDB, Apache Spark, Apache Cassandra

1. INTRODUCTION

Every single thing in the modern, digitally connected world can be thought of as producing data. This data, which is being added to the ocean of Big Data already present and derived from a wide range of sources, including weblogs, cell phones, social networking sites, satellite images, human genome sequencing, consumer transaction data, astronomical and biological records, presents researchers with both enormous opportunities and challenges that must be met to yield fruitful results. The ability to manage enormous volumes of data is more important than merely the size of the data, which can be measured in petabytes or zeta bytes. The definition is given by the University of California, Berkeley —"Big data is when the normal application of current technology doesn't enable users to obtain timely, cost-effective, and quality answers to data-driven questions" [1]. Big Data is most frequently employed in the fields of marketing, sales, IT, healthcare, and finance; but, as big data's dependability increases, organizations are beginning to identify additional, long-term prospects in risk management and logistics. However, due to the difficulties associated with huge data, some caution is necessary. But we frequently limit the capabilities of this technology due to our ignorance of its possibilities and our worry about data security and privacy, especially in light of the Facebook data breach disaster. Large amounts of data, on the order of exabytes or zettabytes, with a wide variety of files including text, images, documents, videos, and log files that come in complex forms including structured, unstructured, and semi-structured formats, all with varying velocity requirements including batch processing and real-time or nearly real-time processing. Big Data dimensionality is expanded to include additional Vs, such as the value that can be extracted from the data and veracity, which describes how well understood the data is. The main problems with big data are the storage, manipulation, and conversion of enormous data into knowledge. Although it is frequently believed that important knowledge is buried beneath the enormous amount of Big Data and must be unearthed, analysts cannot just guess at the data's valuable content [2]. The variety and complexity given by big data, which is primarily unstructured or semi-structured and that comes in huge volume, are difficult for traditional data analytic approaches that work on structured data of low volume to handle.

1.1 Introduction to Big Data Analytics

Huge amounts of data that can't be processed efficiently by current standard applications are referred to as "big data." Big Data processing starts with the raw data, which is typically impossible to store in the memory of a single computer because it hasn't been aggregated. Big Data, a phrase for enormous amounts of data that are both structured and unstructured, inundates businesses daily. Big Data can be employed to analyze insights that can result in better business decisions and strategic company movements [3]. Big Data is described as "high-volume, high-velocity, and/or high-variety information assets that necessitate cost-effective, creative forms of information processing that enable greater insight, decision-making, and process automation," according to Gartner. Big data analytics is used to examine massive amounts of data in order to uncover previously unrecognized trends, connections, and other perspectives. With today's technology, you can quickly analyze your data and obtain insights from it, which is faster and more effective than with more conventional business intelligence solutions [4]. Businesses may benefit from using big data analytics to better understand the information included in their data and to pinpoint the data that will be most helpful for their current and forthcoming business choices. Big Data analysts frequently seek the information that results from data analysis [5]. Big data volumes are dramatically increasing and can be anywhere from a few terabytes (TB) to several hundred petabytes (PB) of enormous data in one location. Recording, preserving, scanning, sharing, reporting, and analyzing are some of the big data-related issues, though. Businesses today examine enormous volumes of highly structured data to discover previously unknown facts. Therefore, big data is where sophisticated analytical analysis on large data sets is done. But the more data collected, the more difficult it is to manage. We'll start by outlining the characteristics and significance of big data in this section. Business advantages will, of course, typically be attained through the processing of highly vast and complicated data, which calls for real-time technologies, but doing so contributes to a variety of requirements meant for contemporary data structures, computational tools, and procedures.

1.2 Characteristics of Big Data

One of the most important characteristics of large data is value. Saving a large volume of data in databases is important for IT infrastructure systems. The term "velocity" refers to the rapid rate of data creation. The potential of data is influenced by the speed at which it is produced. There is an enormous and constant influx of data. Variety refers to a variety of sources, and both structured and unstructured kinds of data are offered. Data in the form of movies, emails, audio files, word processing files, etc. is now also being taken into account. The concept of "Big Data," which is concerned with extraordinarily vast data, is called volume. Volume determines whether a set of data is considered big data or not. Therefore, "Volume" is the most crucial metric among others that should be taken into account when working with "Big Data." The term "variability" refers to the inconsistent nature of data. When working with a huge volume of data, data accuracy is not always guaranteed. Data accuracy and correctness are topics covered by validity. Concern with data security involves vulnerability. Big data breaches are still major breaches, after all. Volatility is a Big Data parameter that deals with the statistical measurement of the dispersion for a certain collection of returns. The current big data trait that deals with data visualisation is visualisation. The inconsistent rate at which the data is saved in our system is referred to as variability in the area of big data. Many different tools and methods are used to analyse large data.

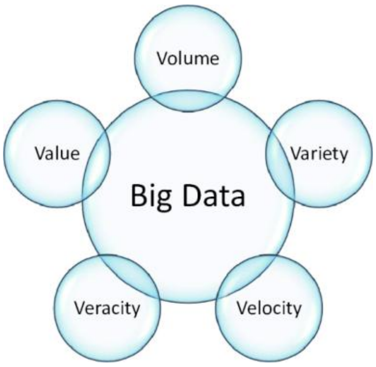

Fig –1: Characteristics of Big Data

2. TOOLS OF BIG DATA

Big Data architecture and the organization's support infrastructure need to be coordinated. All of the data that corporations currently use are static. Data comes from a variety of untidy and unstructured sources, including information from machines or sensors and vast collections of both public and private data. Prior to now, the majority of businesses were unable to either gather or store these data, and the technologies at hand could not handle the data in a timely manner. The new Big Data technology, however, enhances performance, encourages innovation in business model products and services, and supports decision-making [9][10]. According to Statistics Market Research Consulting [11], the global market for data science platforms reached $19.76 billion in 2016 and will increase at a CAGR of more than 36 % to $128.21 billion by 2022. Your analytics depend heavily on big data. If you have access to cutting-edge big data tools and methodologies, you can manage unstructured and imperfect data with ease and get valuable insights from it. 87% of businesses feel that Big Data Analytics will assist them in restructuring their operations within the next three years, and 89% believe that they will fall behind the competition if they do not use Big Data Analytics [12]. Today, practically all businesses use big data extensively to gain a competitive edge in their respective markets. In light of this, enterprises should choose open-source big data solutions for big data processing and analysis because they are less expensive and allow for better time management of data analytical jobs. Focusing on the opensource big data technologies that are advancing the big data sector is beneficial since businesses are creating new solutions quickly to gain a competitive edge in the market.

2.1 Apache Hadoop

One of the technologies used to process Big Data, which is the massive volume of combined structured and unstructured data, is Apache Hadoop. The open-source platform and processing architecture Apache Hadoop only offers batch processing. Map Reduce from Google served as Hadoop's primary inspiration. The entire program is broken down into a number of little sections in the Map Reduce software architecture. These little pieces are also known as fragments. Any system in the cluster can run these fragments [14]. There are numerous parts utilized in the construction of Hadoop. To process batch data, these factors all worked in concert. Principal elements include:

HDFS: The core element of the Hadoop software system is the Hadoop Distributed File System (HDFS). It is the Hadoop file system. HDFS is set up to store a lot of data. It makes use of widely dispersed, inexpensive hardware [13]. This fault-tolerant storage system can hold huge files that range in size from TB to PB. In HDFS, there are two different kinds of nodes: name nodes and data nodes.

Name Node: It performs the role of the master node. It includes every piece of data pertaining to every data node. It contains data about available space, node addresses, all of the data that they store, active nodes, and passive nodes. Additionally, it preserves the data from the task and job trackers. Data Node: A slave node is another name for a data node. The data is kept on a Hadoop data node. And TaskTracker's responsibility is to monitor ongoing jobs that are located on the data node and to take care of jobs coming from the name node.

MapReduce is a framework that enables programmers to write parallel operations for processing large amounts of unstructured data in a distributed architecture. Numerous components, including JobTracker, TaskTracker, JobHistorySever, etc., make up MapReduce. It is also known as the native Hadoop instruction execution engine. It was created in order to handle and store enormous amounts of data using common hardware. The huge volume of data is processed using clusters for record storage. The Map Reduce programming model is built on two functions: the Map function and the Reduce function. The Map function is functional in master node. It also acknowledges the input. Then, distribute the slave nodes with the submodules that were created from the accepted input.

The main Hadoop service, YARN (Yet Another Resource Negotiator), supports two key services: global resource management (ResourceManager) and per-application management (ApplicationMaster). It is the Hadoop stack's cluster coordination component. Execution is made possible through YARN [15] [16]. The MapReduce engine is in charge of making Hadoop usable. The MapReduce framework works with less expensive hardware. Nothing is thought to be saved in memory by it. Unimaginable measurability potential exists with MapReduce. Thousands of nodes have been created using it. The impact of this will be scaled back to varying degrees by other Hadoop enhancements, but it will always be a factor in the speedy execution of an idea on a Hadoop cluster.

2.2 Apache Spark

The AMP science lab at UC Berkeley is where Apache Spark, another open-source technology, was built [15]. It is a framework with support for stream processing. Spark was designed using many of the same principles as Hadoop's MapReduce engine and is completely devoted to accelerating process and instruction execution workloads by supplying full in-memory computation and processing enhancement. This allows us to perform in-memory analytics. This is frequently (100 times) faster than Hadoop. It works quite well with Hadoop's storage. According to information in Figure [17], Apache Spark implements associate degree in-memory batch processing using the RDDs (Resilient Distributed Datasets) model.

Flow process: Spark Streaming is packed with the ability to use Model Stream Process. The design of Spark itself is geared at batch-oriented workloads. Spark uses an idea known as micro-batches to handle the disparity between engine types and the features of streaming workloads. This approach is designed to handle information streams as a collection of incredibly small batches that will be handled using the batch engine's inherent linguistics. Spark is clearly superior to Hadoop MapReduce in terms of performance and advanced DAG programming. Spark's proficiency is one of its main advantages. It can be set up as an independent cluster or connected with an already-existing Hadoop cluster.

2.3 Apache Cassandra

An open source, distributed, and decentralized/distributed storage system (database) called Apache Cassandra is used to manage extremely huge amounts of structured data dispersed over the globe. It offers services that are highly accessible and have no single point of failure. Cassandra is a distributed storage system that offers highly available service with no single point of failure while storing extremely large amounts of structured data dispersed over many commodity machines. Product catalogues and playlists, sensor data and the Internet of Things, messaging and social networking, recommendation, personalization, fraud detection, and several more applications that deal with time series data are just a few examples of the many uses for Cassandra databases. Because of its scalable and fault-tolerant peer-to-peer architecture, flexible and adaptable data model that developed from the BigTable data model, declarative and user-friendly Cassandra Query Language (CQL), and extremely efficient write and read access paths, Cassandra Database has been adopted in big data applications. These features allow critical big data applications to be always on, scale to millions of transactions per second, and handle node and even entire data center faults [18].

Product catalogues and playlists, sensor data and the Internet of Things, messaging and social networking, recommendation, personalization, fraud detection, and several more applications that deal with time series data are just a few examples of the many uses for Cassandra databases. A distributed key-value store is Cassandra. Cassandra only let data to be queried by its key, unlike SQL queries, which allow the client to express arbitrary complicated constraints and joining criteria. Additionally, Cassandra does not contain a join engine, therefore the application must perform any necessary joining of related rows of data. Indexes on non-key columns are also not permitted. In order to maintain referential integrity, the Cassandra data modeller must select easily derivable or discoverable keys. Abstractions that Cassandra has adopted closely resemble the Bigtable design.

2.4 MongoDB

A database built on JSON files is called MongoDB. It was introduced in 2009, is still growing, and is written in C++. MongoDB contains information that can be used by each in small packages with hundreds of users. In essence, the MongoDB database stores a set of data that lacks a clear schema. Data can be stored in the form of BSON documents and has no preset format like tables. Binary-encoded JSON-like items are known as BSON. If the demand is knowledge-intensive, the user should consider MongoDB rather than MySQL since it saves data and supports queries [19,20]. Based on C++, MongoDB is a member of the NoSQL database family. MongoDB was created specifically for the storage and retrieval of information. It is capable of processing and measurement. It was C++ compatible and belonged to the NoSQL group. Relative tables that aren't time-sensitive aren't trusted. It stores its records in document format.

3. DATA STORAGE AND MANAGEMENT

One of the first decisions that organizations must make when dealing with enormous amounts of data is where or how the data will be handled after it has been acquired. Data marts, relational databases, and data warehouses have historically been used in structured data collecting and extraction methods. Using software that retrieves information from outside sources and adapts it to satisfy technological needs, this data is transported to storage from operational data stores, stacked, and eventually stored in databases. Before the data collection and advanced analytics activities are allowed access, the data is processed, altered, and documented.

However, the big data ecosystem calls for expertise in analyzing models like Agile, Magnetic, Deep (MAD), which deviates even more from the components of an (EDW) environment. First off, traditional EDW techniques forbid the use of new data sources until they have been cleansed and integrated. Due to the current inequitable nature of data, which draws all data kinds regardless of quality performance, big data systems are forced to adopt dynamic responses. Additionally, having a huge amount of data storage would allow the analysts to quickly generate the results and alter the data due to the increasing and exponential number of sources for the different types of data as well as the complexity of these data analyses. This invariably comprises a flexible infrastructure, the structure of which will coordinate with quick system evolution. Finally, a large data database frequently needs to be deep and end up acting as an advanced runtime algorithm because current data analyses use complex statistical approaches and professionals must be able to evaluate enormous data sets by digging up or down [21].

Big data has been approached in so many different ways, from in-memory databases to Massive Parallel Processing (MPP) to distributed network databases that offer excellent query efficiency and application scalability. Databases like Not Only SQL (NoSQL) were developed for the storage and management of unstructured or nonrelational data. In contrast to relational databases, NoSQL databases disperse data processing and storage. These databases provide a greater emphasis on high-performance, scalable data storage, and they mandate that information administration tasks be carried out at the application layer rather than in the database itself [22]. On the other hand, memory repositories manage the priceless data in storage memory, eliminating disc input and output and enabling real-time database response. The complete database can be stored in silicon-based central memory instead of mechanical hard discs. Additionally, in-memory databases are now being used for cutting-edge research on massive amounts of data, notably to speed up the entry into and scoring of expository models for analysis. This allows for rapid revelation evaluation and adaption to vast amounts of information. As an alternative, by using the MapReduce model, Hadoop gives a basis for the success of Big Data Analytics that offers accuracy and consistency. A number of data nodes, including the name node, store the data in distributed file blocks. The name node serves as the supervisor between the data node and the client, directing the client to the correct data node that has the necessary data.

4. BIG DATA ANALYSIS METHOD

4.1 Predictive Analytics

Software and/or hardware solutions that let businesses identify, assess, optimize, and implement predictive models by examining big data sources in order to boost operational efficiency or reduce the risk [23]. Forecasting and statistical modeling to ascertain potential future outcomes are central to predictive analytics [24]. The need for new large data statistical methods is prompted by a number of causes. In order to analyze the importance of a given relationship, a small sample from the population is taken, and the results are compared with chance. This is the foundation of traditional statistical approaches. The generalization of the conclusion to the entire population follows. Big data samples, on the other hand, are enormous and comprise the majority, if not the whole population. As a result, big data does not really require the concept of statistical significance. Second, many traditional methods for small samples do not scale up to massive data in terms of computational efficiency. The third factor relates to the unique characteristics of big data, including heterogeneity, noise accumulation, misleading correlations, and incidental endogeneity [25].

4.2 Prescriptive Analytics

Prescriptive analytics focuses on optimization and randomized testing to evaluate how companies might raise their service standards while cutting costs [26]. This kind of analytics is used to establish the causal link between analytical findings and business process optimization guidelines. Predictive analytic models' input is used by enterprises to improve their business process models in prescriptive analytics [27]. Prescriptive analytics help manage the information transition and the ongoing growth of business process models, despite being challenging to implement [28]. There aren't many real-world instances of effective prescriptive analytics. The fact that most databases are limited in the number of dimensions they can store is one of the causes of this limitation [29]. As a result, the analysis of such data offers, at most, fragmentary insights into a challenging business issue. The simulation optimization approaches have only been used in a few preliminary investigations with the BDA. For instance, the framework known as multi-fidelity optimization with ordinal transformation and optimal sampling was proposed by Xu, Zhang, Huang, Chen, and Celik in 2014. (MO2TOS). Under the BD context, the framework offers a platform for descriptive and prescriptive analytics. Two sets of high- and low-resolution models were created for the MO2TOS framework. The authors emphasized that the vast amount of data can make the construction of high-resolution models quite slow. The low-resolution models, on the other hand, grew far more quickly and used only a little amount of data. To easily combine both resolution models and optimize targeted systems in the BD environment, the MO2TOS framework has been presented [30]. Prescriptive solutions often aid business analysts in making decisions by identifying activities and evaluating their influence on company objectives, needs, and limitations. What if, for instance, simulation tools had been able to shed light on the likely options that a company would decide to apply in order to keep or improve its current market position?

4.3 Descriptive Analytics

In the form of providing standard reports, ad hoc reports, and alerts, descriptive analytics examines data and information to define the current status of a business scenario in a way that developments, patterns, and exceptions become apparent [31]. Banerjee, Bandyopadhyay, and Acharya (2013) mentioned the use of dashboard-type applications as another type of descriptive analysis when a business regularly generates various metrics, including data to monitor a process or many processes over time. For instance, this type of application might be helpful to comprehend the financial health of a business at a specific point in time or to compare it to other businesses or its own over time. In descriptive analytics, analysts must have the ability to read facts from figures, relate them to the pertinent decision-making process, and ultimately make a data-driven conclusion from a business standpoint. The majority of business decision analysis (BDA) is often descriptive (exploratory) in nature, and using descriptive statistical approaches (data mining tools) enables firms to identify unidentified correlations or beneficial patterns that may be used for corporate decision-making [32]. Root cause analysis and diagnostics are also types of descriptive analysis, according to Spiess, T'Joens, Dragnea, Spencer, and Philippart (2014). These analyses entail reading and interpreting data passively as well as taking specific actions on the system being tested and reporting the findings. According to the author, root cause analysis is a complex procedure that involves continuously analyzing data and linking different insights in order to identify the one or more underlying causes of an incident [33]. Using analytical drill-downs into data, statistical analysis, and factor analysis, for example, are all examples of inquisitive analytics that include exploring data to confirm or reject business hypotheses [34]. Preemptive analytics refers to the ability to take preventative measures in response to events that could unfavorably affect an organization's performance, such as identifying potential threats and advising mitigation measures well in advance [35].

4.4 Text Analytics

Techniques that extract information from textual data are known as text analytics (also known as text mining). Examples of textual data held by corporations include social network feeds, emails, blogs, online forums, survey replies, corporate documents, news, and call center logs. Machine learning, computational linguistics, and statistical analysis are all used in text analytics. Businesses can transform massive amounts of text produced by humans into insightful summaries that enhance evidence-based decision-making by using text analytics. Text analytics, for instance, can be used to forecast stock market movements using data gleaned from financial news [36]. Let's take a quick look at text analytics techniques below. Techniques for information extraction (IE) take unstructured text and extract structured data from it. For instance, IE algorithms may extract structured data from medical prescriptions, such as medicine name, dosage, and frequency. Entity Recognition (ER) and Relation Extraction (RE) are two IE sub-tasks [37]. ER identifies names in text and groups them into pre-established groups like person, date, location, and organization. In the text, RE identifies and extracts semantic links between entities (such as people, organizations, medications, genes, etc.).

For instance, a RE system can extract relations like Founder Of [Steve Jobs, Apple Inc.] or Founded In [AppleInc., 1976] from the line "Steve Jobs co-founded Apple Inc. in 1976." Examples of industry-standard QA systems include Apple's Siri and IBM's Watson. These programs have been used in marketing, finance, healthcare, and education. Complex Natural Language Processing (NLP) methods are used by QA systems. Sentiment analysis (opinion mining) techniques examine opinionated material, which includes views on things like goods, services, people, and events. Sentiment analysis has become increasingly popular as a result of businesses collecting more data about the attitudes of their customers [37].

4.5 Data Visualization

A generic word used to describe any effort to explain the importance of data by placing it in a visual context is data visualization. With the aid of data visualization software, patterns, trends, and correlations that could go unnoced in text-based data can be exposed and identified more easily. It enables applications to get data devoid of technical limitations imposed by data formats, data locations, etc. Data virtualization is one of the most popular big data technologies, employed by Apache Hadoop and other distributed data stores providing real-time or near real-time access to data stored on many platforms [38].

5. PROCESSING OF BIG DATA

The big data storage is followed by analytics processing [39]. The first requirement is for the data to load quickly. However, as file and network/internet traffic mostly correlates with query performance, during data preparation, the loading time for data must be decreased. Quickly responding to inquiries is the second requirement. To meet the requirements of increasing workloads and real-time requests, the majority of queries become critical in terms of response time. If a result, the data structure should be able to maintain excellent query speeds even as query quantities grow rapidly.

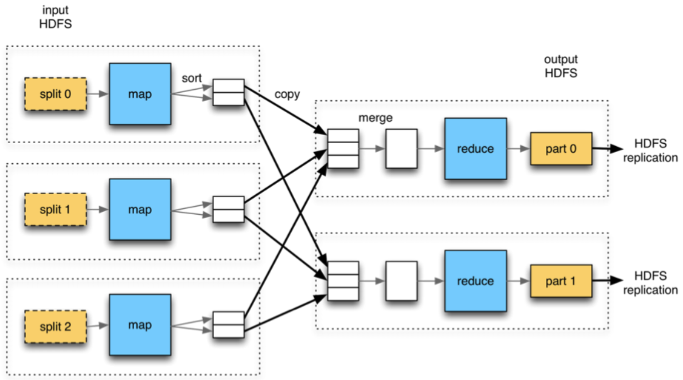

Fig -2: Representation of Map Reduce [40]

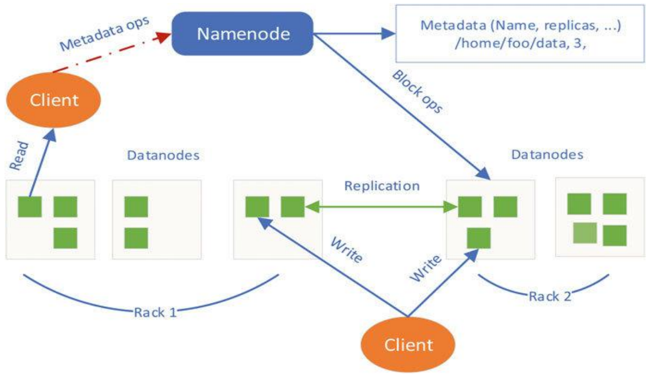

Fig -3: HDFS Architecture [41]

MapReduce is a distributed computing-based programming approach and software configuration for Java-based systems. Map and Reduce are the two primary responsibilities of the MapReduce algorithm. In order to create another data set in which distinctive features are divided into tuples, Map first prepares a set of data. In the second task, "reduce," features are extracted using the map as an input, and these other distinct lists of data are then transformed into a smaller subset of lists. The reduction task is always completed immediately following the map task, as suggested by the name MapReduce. The MapReduce feature in Hadoop depends on two different nodes: the Work or job Tracker and the Task Tracker nodes. The Job Tracker node is in charge of providing two functions, such as the function that maps, which is the mapper function, and the reducer functions to the corresponding Task Trackers that are available, as well as for tracking the results.

6. BIG DATA ANALYTICS AND DECISION MAKING

Big data is significant from a decision-perspective since it can provide important knowledge and facts on which to base policies. The managerial and decision-making process has been a crucial and meticulously covered topic in research for years. Big data is now a far more valuable resource for companies. Scanners, mobile phones, rewards programs, the internet, and web technologies are just a few of the sources that produce enormous amounts of incredibly precise information that present chances for businesses to generate large gains. That's only feasible if the data are properly analyzed to unearth insightful knowledge that enables firms to take advantage of the wealth of historical and current data produced by distribution networks, industrial processes, consumer preferences, etc. Additionally, businesses typically evaluate essential papers including revenue, imports, and stockpiles. The usage of big data will have gathered knowledge and insight, but there has been a need to analyze datasets, such as customer needs and suppliers. With the amount of complicated text available, both in size and format, reasonable decisions must be made frequently based on certain presumptions about the data.

7. FUTURE WORK

In order to facilitate meaningful results from these concepts, machine learning concepts and technologies are also becoming more and more popular among researchers. The processing of data, the use of algorithms, and optimization have been the main areas of research in the field of machine learning for big data. Many of the recently developed machine learning technologies for big data require significant adjustment to be adopted. We contend that even if each tool has benefits and drawbacks of its own, more effective tools can be created to address big data's inherent issues. The effective tools that need to be created must be able to deal with noisy and unbalanced data, uncertainty and inconsistent results, and missing numbers.

8. CONCLUSION

Throughout this study, I focused on the novel topic of big data, which has drawn a lot of attention due to its perception of opportunities and advantages that are unmatched by anything else. We live in an information age when significant variations are produced globally with high-speed data, yet within them there are underlying specifics and trends of hidden patterns that can be deduced and utilized. Therefore, by attempting to apply various statistical methods to big data and exposing both powerful and valuable knowledge, big data analytics can be used to exploit change in business as well as optimize decision-making. Big Data Analytics, in my opinion, is extremely pertinent in this era of data overspill and also offers unexpected additional insight, which is used by decision-makers in various fields. If properly used and implemented, big data analytics has the immense potential to lay the groundwork for advancements in the fields of science, mathematics, and humanities. Big data analysis methods and tools are covered in detail. Hadoop might be a good option for batch-only applications and is probably less expensive to adopt than alternative solutions. Time-sensitive workloads are not batch-only workloads. Hadoop offers a thoroughly established methodology for the execution of instructions that is best suited to handle extremely large datasets where time isn't a major concern. For individuals with a variety of process workloads, a spark can be a nice option. The benefits of speed when executing Spark instructions are remarkable.

REFERENCES

[1] T. Kraska, "Finding the Needle in the Big Data Systems Haystack," IEEE Internet Computing, vol. 17, no. 1, pp. 84-86, 2013.

[2] F. Shull, "Getting an Intuition for Big Data," IEEE Software, vol. 30, no. 4, pp. 3-6, 2013.

[3] Data Science vs. Big Data vs. Data Analytics 2018, simplilearn, accessed 14 May 2018, https://www.simplilearn.com/data-science-vs-bigdata-vs-data-analytics-article

[4] Big Data Analytics What it is and why it matters, SAS, accessed 14 May 2018

[5] Vangie Beal. (2018), Big Data Analytics, accessed 10 May 2018, https://www.webopedia.com/TERM/B/big_data_analytics.html

[6] Big Data Tools and Techniques: A Roadmap for Predictive Analytics Ritu Ratra, Preeti Gulia

[7] Rick. Smolan, Jennifer. Erwitt, “The Human face of Big Data, Ed. Against all odds production”, Sausalito, CA 2012.

[8] https://www.researchgate.net/figure/big-data-5-Vs-Volume-According-to-a-Fortune-magazine-11-we-created-5-Exabytes-of_fig1_268157581

[9] J. Manyika, C. Michael, B. Brown et al., “Big data: The next frontier for innovation, competition, and productivity,” Tech. Rep., Mc Kinsey, May 2011.

[10] J. Manyika, M. Chui, B. Brown et al., “Big data: the next frontier for innovation, competition, and productivity,” McKinsey Global Institute, 2011.

[11] Rick Whiting. (2018), 2018 Big Data 100: The 10 Coolest Data Science And Machine Learning Tools, accessed 10 May 2018

[12] A COMPLETE LIST OF BIG DATA ANALYTICS TOOLS TO MASTER IN 2018, accessed 10 May 2018

[13] Kaisler S, Armour F, Espinosa JA, Money W. “Big data: issues and challenges moving forward” In: System sciences (HICSS), 2013 46th Hawaii international conference on, IEEE. 2013. pp. 995–1004.

[14] Kubick, W.R. “Big Data, Information and Meaning”, In: Clinical Trial Insights, pp. 26–28 (2012)

[15] TechAmerica: “Demystifying Big Data: A Practical Guide to Transforming the Business of Government”, In: TechAmerica Reports, pp. 1–40 (2012)

[16] Ms. Komal , “A Review Paper on Big Data Analytics Tools” (IJTIMES), e-ISSN: 2455-2585 Volume 4, Issue 5, May-2018, pp 1012-1017.

[17] G. George and D. Lavie, “Big data and data science methods for management research”, Academy of Management Journal, vol 59, issue 5, pp. 1493 – 1507, 2016.

[18] Cassandra as a Big data Modeling Methodology for Distributed Database System https://www.ijedr.org/papers/IJEDR1703135.pdf

[19] https://www.google.com/imgres.

[20] MongoDB, Inc. (2015, Aprilie), MongoDB Ops Manager Manual Release 1.6,[Online].Available: https://docs.opsmanager.mongodb.com/current/opsmanager-manual.pdf

[21] Strohbach, M., Daubert, J., Ravkin, H., Lischka, M. (2016). Big Data Storage. In: Cavanillas, J., Curry, E., Wahlster, W. (eds) New Horizons for a Data-Driven Economy. Springer, Cham. https://doi.org/10.1007/978-3-319-21569-3_7

[22] Mazumdar, S., Seybold, D., Kritikos, K. *et al.* A survey on data storage and placement methodologies for Cloud-Big

Data ecosystem. *J Big Data* **6**, 15 (2019). https://doi.org/10.1186/s40537-019-0178-3

[23] Gil Press. (2016), Top 10 Hot Big Data Technologies,accessed 11 May 2018

[24] 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), 77–84.

[25] Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. National ScienceReview, 1(2), 293–314.

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

[27] Bihani, P., & Patil, S. T. (2014). A comparative study of data analysis techniques. International Journal of Emerging Trends & Technology in Computer Science, 3(2), 95–101.

[28] Rehman, M. H., Chang, V., Batool, A., & Teh, Y. W. (2016). Big data reduction framework for value creation in sustainable enterprises. International Journal of Information Management (Accepted).

[29] Banerjee, A., Bandyopadhyay, T., & Acharya, P. (2013). Data analytics: hyped up aspirations or true potential. Vikalpa. The Journal for Decision Makers, 38(4), 1–11.

[30] Xu, J. S., Zhang, E., Huang, C. -H., Chen, L. H. L., & Celik, N. (2014). Efficient multi-fidelity simulation optimization. Proceedings of 2014 winter simulation conference. GA: Savanna.

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

[32] Banerjee, A., Bandyopadhyay, T., & Acharya, P. (2013). Data analytics: hyped up aspirations or true potential. Vikalpa. The Journal for Decision Makers, 38(4), 1–11.

[33] Spiess, J., T'Joens, Y., Dragnea, R., Spencer, P., & Philippart, L. (2014). Using big data to improve customer experience and business performance. Bell Labs Technical Journal, 18(4), 3–17.

[34] Bihani, P., & Patil, S. T. (2014). A comparative study of data analysis techniques. International Journal of Emerging Trends & Technology in Computer Science, 3(2), 95–101.

[35] Szongott, C., Henne, B., & von Voigt, G. (2012). Big data privacy issues in public social media. 6th IEEE international conference on digital ecosystems technologies (DEST) (pp. 1–6).

[36] Chung, W. (2014). BizPro: Extracting and categorizing business intelligence factors from textual news articles. International Journal of Information Management,34(2), 272–284.

[37] Jiang, J. (2012). Information extraction from text. In C. C. Aggarwal, & C. Zhai (Eds.),Mining text data (pp. 11–41). United States: Springer.

[38] Margaret Rouse. (2017), Data visualization, accessed 11 April 2018, https://searchbusinessanalytics.techtarget.com/definition/data-visualization

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

[40] https://training-course material.com/index.php?title=Hadoop_Administration&action=slide

[41] B. Abu-Salih et al., Social Big Data Analytics