ISSN 2278-3091

Volume 9, No.4, July – August 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse174942020.pdf

https://doi.org/10.30534/ijatcse/2020/174942020

Big Data Analytics: Importance, Challenges, Categories, Techniques, and Tools

Sarah Alswedani, Mostafa Saleh
Faculty of Computing & Information Technology
King Abdulaziz University, Jeddah 21589, Saudi Arabia
sabdullahalswedani@stu.kau.edu.sa
msherbini@kau.edu.sa

ABSTRACT

With the gigantic explosion of the volume of data generated every single day, big data analytics has born as a powerful technology for various organizations. This paper explores the importance of big data analytics for different domains, the challenges of utilizing big data analytics due to the complex nature of big data, and the technical approaches that are used for handling this complexity such as Hadoop, Spark, and Storm. This paper also reviews the challenges associated with big data based on the literature. Furthermore, big data analytics categories are investigated including text, audio, and video.

Key words: Big Data, Big Data Analytics, Machine Learning, MapReduce, and HDFS

1. INTRODUCTION

The gigantic explosion of digital technologies such as smartphones, the Internet of Things, mobile and social networking applications, healthcare applications, multimedia, cloud computing, and autonomous mobile resulted in massive amounts of data [1] which is growing enormously in every minute. More data is being generated every year. In 2017, 462,00 posts were uploaded by Instagram every minute, however, it increased to 174, 000 posts in 2018 [2]. Recently reported, more than 2.5 PB of data are imported and processed every hour only by Walmart retail corporation. By 2020, it is expected that the size of data will dramatically grow by 50% [3]. Based on a recent study, 2.5 quintillion bytes of data are being generated in every single day [4]. According to [5], by 2030 the number of sensors is expected to increase to become about 1 trillion sensors. Accordingly, the current amounts of data generated by the Internet of things will be considered small as compared to 2030. This huge amount of data drives the society to big data world. It is considered “a new oil” as was stated by IBM’s Chief Executive officer [6]. It provides a lot of valuable knowledge and deeper insights for smarter decisions, better products, and perfect solutions [7]. Business can achieve higher performance when utilizing decisions obtained by big Data analytics. It greatly aids in handling the problems instantly, providing important and valuable insights that can eventually provide them with a competitive advantage [8]. Its enormous volume of data is considered as an asset when analyzed comprehensively. Big Data analytics refers to the methods and capabilities that examine, process, detect, and reveal the invisible patterns, interesting relations, and crucial insights under analysis and investigation [9].

The paper is organized as follows. Section □ presents an overview of big data including different definitions of big data, as well as, its characteristics. Then in section □, the concept behind big data analytics with its importance for various sectors is introduced. After that, section □ discusses some of the challenges related to big data along with data storage, analysis methods, security, and scalability. Section □ demonstrates some categories of big data analytics involving analytics of text, audio, and video. Section □ presents various analytics techniques using machine learning mechanisms such as fuzzy, evolutionary algorithms, neural networks, and deep learning. Section □ reviews some of big data analytics tools. Finally, the conclusion is provided in section □.

2. BACKGROUND

2.1 Big Data Definitions

Big Data is the big dataset with size and nature exceed the capabilities of the traditional database systems for management, storage, and analysis. Various definitions of big data have been introduced. According to [10], big data involves vast collections of data with massive volumes that are updated rapidly on a frequent basis and it includes a vast range of various data formats. However, the authors in [11] define big data as it is a composition of various features including data variety, velocity, volume, and veracity that provides institutions with a chance of gaining competitive benefits in nowadays market. With respect to [12]. The support for big data includes merging these technologies to provide novel solutions that can produce considerable advantages to the business field. In [13], big data is described as a flood of data that affects every existing domain such as science, business, art, and government.

2.2 Big Data Characteristics

The big data holds three to four main characteristics (3 or 4 Vs) including Volume, Variety, and Velocity [14]. The volume refers to the massive volume of data that is being produced on a daily basis. Some of the existing technologies such as social media, smartphones, computers, sensors, websites, etc. generate data with enormous amounts that could reach tens of zettabytes. In fact, 40 Zetta-bytes of data are expected to be generated by 2020 [15]. Variety means that data exist with various formats which can be highly structured, unstructured, or semi-structured. Relating to velocity, it refers to the pace at which data is being produced, combined, ingested, and processed. With big data, data is generated very rapidly [16]. For example, with Snapchat application, 2.4 million snaps are created in a minute in 2018 [3]. Velocity also involves the speed of the data availability and delivery. Value reflects the usefulness of the data and whether it can be turned into value and produce insights or not. Another feature is known as Veracity and it refers to the trustworthiness (the quality) of the data including accuracy, certainty, and precision [16] [17]. It has two features: the reliability of the source, and the convenience of data for its intended users. Before data experience processing, it goes through quality testing [14]. Data complexity is considered other characteristics of big data [18]. Complexity refers to the degree of interdependence as well as the interconnectedness in the structures of big data.

3. BIG DATA ANALYTICS

The concept of big data analytics mainly refers to the mechanism of extracting insights and useful information from the huge amount of data. It involves multiple processes including the data collection, organization, and analysis of huge data sets to identify various useful and meaningful information and patterns. It is a collection of different technologies and techniques that needs new shapes of integration to reveal the hidden values from massive amounts of data having more complexity. The main emphasis of big data analytics is solving new or old issues in a more efficient way [19].

The major objective of big data analytics is to provide assistance for organizations to perform better prediction and data analysis, as well as, to make better and smarter decisions. Flipkart and snapdeal are two examples of online websites for big data analytics that are using Facebook or Gmail data to provide a view for the business information. Big data analytics is utilizing previous data that are usually unusable. The analysis of big data enables researchers, data analysts, business owners, and users to take better as well as faster decisions. Employing sophisticated analysis methods such as machine learning, text analytics, natural language processing (NLP), predictive analytics, as well as statistics can provide an analysis for unused data to obtain new insights that can help providing better and faster decisions. It provides supports in revealing hidden patterns, correlations, customer preferences, as well as market trends. It leads to more successful marketing, better services, etc. [19].

3.1 The Importance of Big Data Analytics

Big Data “is becoming a significant corporate asset, a vital economic input, and the foundation of new business models" as stated in [6]. It is a wealth of knowledge and it will make a significant contribution on the future life in smart cities. From a medical perspective, it can provide evidence in the detection of patterns and symptoms of diseases, recent health issues, or pandemics. Moving to economy and business in a smart city, big data analytics can provide significant help in which generated data from social networks or smart devices can be employed in finding users’ preferences, and realizing the relationships between business corporations (both competitive and collaborating). Accordingly, this will result in producing better products and services, or enhancing existing ones [5]. Obtaining knowledge for the preferences as well as the needs of customers will support businesses and organizations with a vital benefit over competitors. Finally, from an authority perspective, big data analytics can assist governments in providing citizens with better services by utilizing generated data. It also helps governments supporting modern societies by enhancing various sectors including education, public transport, healthcare, and other domains of life. For example, obtained data from the traffic domain can be used for the enhancement of the services of public transportation which is supported by the state and delivered to the population. [9]

4. BIG DATA ANALYTICS CHALLENGES

Utilizing big data analytics face numerous challenges due to the various forms of data, in addition to the massive data size in big data [20]. The challenges are classified into multiple groups including data storage, analysis methods, data security, and scalability.

4.1 Data Storage

Recently the data size is growing on an exponential scale. The Internet of Things (IoT) devices in addition to the social media applications are all considered major contributors behind the enormous data size. With Big data analytics, data accessibility has the highest priority since data needs to be accessed easily and quickly for processing. In the past, hard drives were used for storage but due to its poor performance SSD technology was introduced. Nowadays, all the previous technologies cannot provide the required performance for big data processing. However, storing this huge amount of data in a medium that provides high input/out speed for data processing is a major challenge [17].

4.2 Analysis Methods

Big data analysis has two main focuses. The first is to develop efficient approaches that can provide accurate predictions. The second is to obtain insights from the relations between the data features and attributes [21]. Suitable analysis tools are required for handling the inconsistency, uncertainty, and complexity of big data. However, finding the appropriate tool is considered a challenging issue for data scientists and analysts. [20]

4.3 Data Security

With big data analytics, enormous amounts of data are analyzed, correlated, and mined for obtaining the needed knowledge. Different organizations are imposing several polices for saving sensitive information. Keeping sensitive information safe is a significant matter in big data analytics [22]. Inappropriate use of individual data could destroy a company's trademark. For instance, Facebook has lost about one hundred billion dollars for breaching data [23]. The privacy of Big data can be maintained by multiple techniques including encryption, authorization, as well as authentication. Due to the fact that big data applications face a lack of intrusion system and network scale, privacy-based data models with multi-level security are required which are considered a major challenge.

4.4 Scalability

With the lack of scalability, big data analytics becomes a major bottleneck. The scalability of big data is the most critical challenge with big data analytics. As the data is expanding at a very fast pace even faster than the speed of CPU, processors are facing spectacular move being fabricated with a various number of cores [23]. This jump in processors technology has led to the emergence of the applications of parallel computing such as social networks and Internet search. Scalability challenges require scientists to relate more mathematical models to the field of computer science [17].

5. BIG DATA ANALYTICS CATEGORIES

In this section, some categories of big data analytics will be discussed including the analytics of data in the form of text, audio, and video. Several examples and case studies are provided. A summary of the documented categories is presented in Figure 1.

5.1 Text Analytics

The analytics of textual data refers to text mining which is the process of analyzing the text which is in an unstructured format in order to extract important information and then transform it into a structured format that can be utilized in various manners [24]. Examples of textual data are emails, news, social network feeds/posts, and documents. The process of text analysis is utilizing numerous techniques such as Natural Language Processing (NLP), machine learning, statistical analysis, and computational intelligence [25]. Text analytics is used in all types of analyses such as the analytics of social media, fraud detection, and churn prediction [24].

Text analytics involve the following methods: Information extraction (IE) and Relation Extraction (RE). The techniques of Information extraction are used to extract data in structure format from the unstructured data, for instance, extracting the details from the medical prescription such as the name of the medical drug, the dosage, as well as how often the medicine should be taken. IE includes two sub-tasks: Recognition of Entity and Extraction of Relation. With respect to Recognition of Entity (ER), it locates the names and then classifies them into various categories which are defined in advance such as locations, persons, and companies. For Entity Recognition, supervised, semi-supervised, or unsupervised learning techniques can be used. Some examples of supervised learning methods are Support Vector Machine (SVM), decision trees, and Hidden Markov models (HMM). For semi-supervised, bootstrapping technique is an example while for unsupervised learning clustering algorithms can be used [25]. For Relation Extraction (RE), it uses different methods such as supervised and semi supervised approaches to extracts the semantic relationships between the entities [26]. Kernel and features based algorithms are examples of supervised approaches while bootstrapping and Snowball are examples of and semi-supervised approaches. [25]

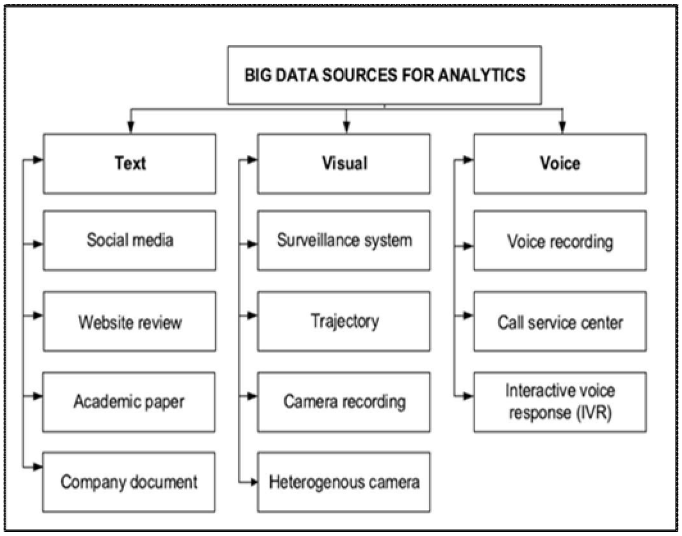

Figure 1: Big Data Analytics Categories [20]

Product defect discovery is a case study related to text analytics. The researchers in [27] proposed a framework for social media analytics using text. The system can detect a product defect from the postings on social media. Other case studies for text analytics can be found in [28] and [29]. The summary of these cases is depicted in Table1.

5.2 Audio Analytics

Audio analytic which is known as speech analytics refers to the process of extracting information from audio signals which is considered unstructured data. Audio Analytics is involved in various application areas such as call service centers, healthcare organizations, surveillance applications, and threats detection. For example, in the call service center and in order to measure the performance of the customer service, hundreds of thousands of hours of the recorded calls need to be analyzed in order to obtain insights, identify the issues related to the service, and improve the customer service [25]. The researchers in [29] designed a method for evaluating the performance of the call center by utilizing the big data. The recorded calls have been collected and then converted to textual format with Google Speech API. Then, call conversations are analyzed using Hadoop and MapReduce. The results were given with the required quality attributes such as accuracy, completeness, and reliability. Furthermore,
audio analytics can be used by health care providers for diagnosing and/or treating some medical cases as depressive disorders, schizophrenia, and cancer as they obviously affect the patterns of communication for the patient [10]. For depressed patients, their muffed speech cannot be easily understood by doctors. In [30], the scholars studied the relationship between the Mel-frequency cepstrum coefficient (MFCC) and depression using OpenSMILE. Audio analytic techniques can also be used for monitoring the infant's health and emotional status by analyzing infant's cries [11]. Voice analytics use two approaches: Transcript-based approach (LVCSR) and Phonetic-based approach:

Table 1: Summary of Big Data Analytics Categories with Case Studies [20]

Analytic Category	Case Study		
	Data Source	**Framework**	**Reference**
Text	Social Media	Social Media Analytics Framework using text (SMART)	[27]
	Website Review	WEKA	[28]
	Academic Paper	R	[29]
Audio	Voice recording	OpenSMILE and Mel-frquency ceptrum cofficients (MFCG)	[30]
	Call service center	Hadoop	[31]
	Voice recording	Speech Activity Detection	[32]
Video	Heterogenous Camera	Kestrel	[33]
	Surviellance System	OpenCV	[34]
	Video Recording	I-AVER	[35]

With respect to Transcript-based approach (LVCSR), it is an abbreviation for large vocabulary continuous speech recognition. The system based on LVCSR passes through two-steps: first indexing and then searching. In the indexing stage, the content, in terms of speech, of the audio is transcribed where the sounds are matched to words using algorithms for automatic recognition of speech. The words are recognized using a predefined dictionary. In case the exact word is not found in the dictionary, the most similar word should be returned. The output of the first phase is an indexed file that contains information about the order of the words that have been mentioned in the speech. In the second phase, text-based mechanisms are utilized for finding the target term in the index file [25].

For Phonetic technique, the phonemes are considered which are sound units that differentiate between the words. For example, the phonemes/k/and/m/ distinguishes between the word “cat” and the word “mat”. This approach also involves two phases: first is phonetic indexing and then searching. During the first phase (indexing stage), the input is mapped to a series of phonemes. However, in the search phase, the system searches the output of the indexing phase. [25]. Additional case studies for audio analytics are found in [31] and [32]. The summary of these cases is depicted in Table1.

5.3 Video Analytics

Video analytics refers to the process of analyzing the data in the video format which is known as video content analysis (VCA). It involves different techniques for extracting useful information by monitoring and analyzing video streams. The demand for video analysis is increasing, various approaches are adopted for video analysis [12], however, they are still not very efficient due to the issues arise from the size of the video data. In fact, a video of one second with high definition is about 2000 pages of a textual document; What about hundreds of hours for videos are uploaded to YouTube every minute! Video analytics has been utilized effectively for surveillance systems where surveillance functions are performed such as identification of unattended objects, detection of violations of restricted areas, tampering with monitoring devices like cameras. After the detection of threats, the system will provide real-time notification or activate automatic alerts such as sound alarm. From video data, the number of clients as well as how long they spend in various areas of the market can be obtained which can help to detect the customer movement patterns. Analyzing video data will help to provide business owners with valuable insights that can help with the decisions for the place, time, cost, promotion policies of the merchandise, and the layout of the store and staffing. Furthermore, video can be indexed to facilitate searching and retrieval based on metadata, the

visual content of the video, and the audio track.
For analyzing video data, three different approaches exist server-based, edge-based, and agent Vi's distributed architecture. With server-based approach, the captured videos are forwarded to a centralized server for analysis, while with edge-based the analysis process is performed locally. For the distributed architecture, video analysis is distributed between the server and the end node device. Each approach has its advantages and disadvantages. For example, a server-based approach provides easier maintenance but less accuracy because of the limited bandwidth. However, an edge-based approach gives better analysis results since the entire content is available on the client-side, but the maintenance is pricey and the computing power is smaller as compared to the server-based approach [25]. Here is an example of a video analytics case study. The research study in [33] has applied video analytics for tracking vehicles. Their analytics system can search for incidents or events quickly using CNN. Further case studies for video analytics are found in [34] and [35]. The summary of these cases is depicted in Table1.

6. BIG DATA ANALYTICS TECHNIQUES

With big data, the form of data is complex in which various data formats are included such as structured, unstructured, and semi-structured. Thus, unlike regular analysis for data, more sophisticated techniques are needed for analysis purposes [36]. For big data analytic, various techniques are used such as machine learning, text analytics, data mining, statistical analytics, visualization, predictive analytics, as well as deep learning. In the next section, examples of some of the most current trends for big data analytic are presented.

6.1 Machine Learning

Machine learning (ML) approaches provide different techniques to detect relationships and make predictions by modeling patterns as well as correlations from big data. The learning forms of ML algorithms are categorized into supervised learning, unsupervised learning, and reinforcement learning. The approaches of ML can be classified to clustering approaches, regression techniques, density estimation methods, and dimensionality reduction approaches. Some examples of ML approaches are artificial neural networks (ANN), decision trees, support vector machines (SVM), deep learning (DL), clustering, and classification [9].

Computational Intelligence (CI) is one of the approaches under ML. CI provides ability to process complex and uncertain data sources by imitating human processing and reasoning. It involves a number of methodologies for addressing complex problems that are unsolvable by traditional models due to its nature of being complex and uncertain to process. The basic approaches of CI are Fuzzy Logic (FL), Artificial Neural Networks (ANN), and Evolutionary Algorithms (EA). [9]

Fuzzy Logic (FL) is an approach that is designed to analyze data with uncertainty and inaccurate nature. Based on the use of fuzzy sets that are used for inference and decision making, FL provides reasoning and modeling for qualitative data. Previous research studies have proven the effectiveness of Fuzzy logic in handling uncertainties of data. For example, the systems based on Fuzzy logic have proven their capability to process human emotions which contain considerable amounts of uncertainty in data with satisfactory accuracy and performance. Fuzzy Logic has been utilized in different application areas of Big Data analytics such as social networks, and medicine. [9]

Evolutionary Algorithms (EA) are another CI approach that satisfies the requirements of big data. They have ability to deal with a high degree of sparseness and dimensionality of Big data with good performance as demonstrated by [37]. They are also considered proven methods for some machine learning problems such as feature selection, clustering, and others. [9]

Artificial neuron networks (ANNs) are a CI approach that is designed to imitate the structure as well as the functionality of the biological brain (neural network). They are statistical and nonlinear data modeling tools that can model the complicated connections between inputs and outputs as well as to find available patterns. They are essential algorithms for image analysis, pattern recognition, as others. However, with Big Data, ANNs suffer from poor performance due to the complexity of ANNs in the learning process where it is time and memory consuming. Accordingly, ANNs have been enhanced in a distributed and parallel setting such as Hadoop and Map/Reduce in order to reduce the consumption of both time and memory. [38]
The combination of CI methods is adopted in various application domains. This integration is applied to provide smart systems for data analysis and decision support. It is required by applications that are characterized by huge amounts of complicated information where analysis is necessary for getting practical decisions with a low cost. [9]

Deep-Learning (DL) is another popular ML approach that is used to discover correlations and extract relevant information from complex datasets [9]. It performs deep analysis to discover invisible hierarchies, unobservable correlations, unnoticed parameters, as well as to find the complicated distribution of variables [39]. It supports various types of data such as text, images, and sound [9]. It is based on the foundation of utilizing ANN with many hidden layers with various neurons (processing units) as the sigmoid function in every layer which makes it a very powerful tool for modeling complex data [39]. It is using unsupervised learning for generating representations that can then be used for training a classifier using supervised training. It has been utilized in different Big Data applications such as pattern recognition, image analysis, genomic medicine, and text mining. Most of the analytics techniques for big data are developed on the basis of deep-learning approach as it uses classification optimization, and statistical estimations. Exploiting the different mechanisms of deep learning in big data analytics have been demonstrated in various new review studies. [9]

7. TOOLS

Selecting the perfect tools for processing big data is a significant and challenging task. Choosing an inappropriate tool could result in some issues. For example, as stated in [39], 22% of the tools suffer from low processing, 23% have issues with scalability, 21% experience slow data loading and 32% of tools have insufficient database analytics. Thus, for selecting the suitable Big data analytics tool, several factors should be taken into consideration including the processing speed, data size, and development model [40]. In this section, various platforms from recent studies are presented briefly.

7.1 Hadoop

It is an open-source distributed framework for storage as well as processing large data sets. It is fault-tolerant and it is designed to provide reliability and to scale up to thousands of nodes. Hadoop stack consists of various components [41]:

- Hadoop Common: it contains libraries need by other Hadoop components.
- Hadoop Distributed File System (HDFS): it is a distributed file storage system to store huge amounts of data on a cluster of computers [42]. It exists at the bottom of Hadoop software stack.
- YARN – it is a resource management layer in which it is responsible for managing available resources and scheduling tasks over the clusters.
- MapReduce: it is a programming model for processing data with a very huge scale in various clusters. It was introduced by Dean and Ghemawat at Google. It is considered the basic data processing scheme in Hadoop.

Here is a brief summary of the most important components:

A. Hadoop Distributed File System (HDFS)

It is a distributed and scalable storage system. It is designed with capabilities to handle data with very large data sizes. Every file is split into blocks that are distributed across the cluster. The architecture of HDFS has two types of nodes working in a master-slave architecture: the master node is referred to as Name Node and the slave nodes are known as Data Nodes, as shown in Figure 2. The small cluster includes one master node and multiple worker nodes and it could reach thousands of worker nodes in a large cluster. The master node is responsible for managing the file system and it controls the access by clients as well the processing on the cluster. It acts as a regulator who forwards the client to a specific Data Node that contains the needed data. For the Data Node, it stores the data in its storage space. Each worker node has its local file system. Data is stored in a replicated file block, by default three copies, on multiple Data Nodes [43] [44] [14]. HDFS provides the cluster nodes with fast data transmission [45] and high availability of data by using a replication mechanism [43]. Accordingly, the design of HDFS ensures reliability and tolerance.

B. MapReduce

It is a parallel programming model and framework for speeding up processing massive amounts of data in various clusters. It is used in Hadoop for processing the data sets with parallel computing. It is considered the core unit in Hadoop as it performs data processing and analytics. It supports horizontal scalability in which it is based on adding more computers rather than increasing the computing power of a single computer. The basic idea of MapReduce is to reduce the time needed to complete the tasks by dividing the tasks into smaller tasks and executes tasks in parallel [14] [43].

1) Components of MapReduce

Job Tracker: It is software located on the Name Node. It schedules the Map or Reduce tasks to the Task Trackers based on its awareness about data location on Data Nodes.

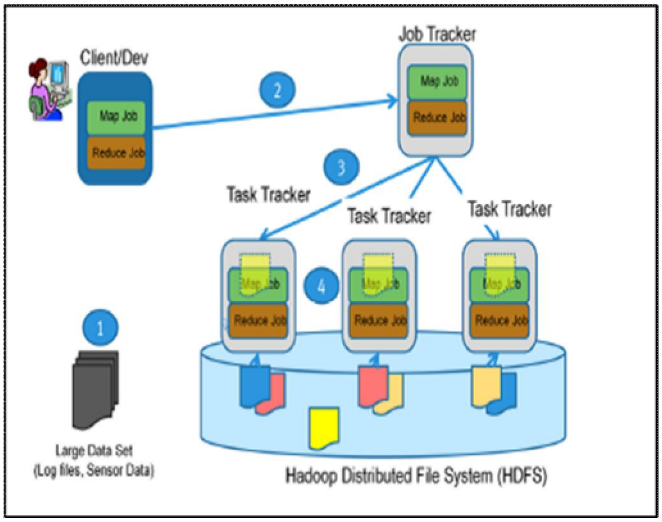

Figure 2: MapReduce and HDFS [44]

Additionally, it monitors the execution of assigned tasks to Task Trackers. [14]

Task Tracker: It is responsible of the execution of the assigned task to the Data Node. Upon the completion of task execution, it informs the Job Tracker on the Name Node. [14]

2) Working Mechanism of MapReduce

The MapReduce starts with Job Tracker assigning an input file on HDFS. In the first phase of MapReduce, Map function divides large tasks to smaller tasks and then assign them to the suitable key/value pairs. In the second phase, reduce function receives the input which is the output of Map function, and it combines all output values that have the same key value to provide the final result. Figure 2 displays how MapReduce and HDFS work together.

1) Data Processing

Hadoop offers a batch processing mode. The processing of data occurs at MapReduce and it is applying the algorithm of map, then shuffle, and finally reduce phase. The process involves 1) fetching datasets from the HDFS system and then 2) breaking the data into smaller pieces to distribute them on the nodes that are available to perform the computations on every node where the intermediate data are written to the HDFS, 3) the intermediate results are redistributed by the key, and finally, 4) the results on the individual nodes are combined together and written to HDFS. [45] [14]

Data processing methodology of MapReduce requires multiple reading and writing per task which is time consuming. However, it can handle massive amounts of data since it relays on disk space which is one of the resources which is considered plentiful. Furthermore, it is considered cost effective since it is designed to run on low cost hardware components. Moreover, it provides good scalability in which thousands of nodes can be added. With respect to analytical operations, the ecosystem of Hadoop is very large, and it can be used as core elements for building other software tools (e.g. HBase, Hive, and advanced algorithms) that exploit YARN manger and HDFS [42]. In summary, Hadoop is suitable applications with enormous amounts of data where the time is not very critical [45]. A summary of the advantages and limitations of Hadoop is presented in Table 2.

7.2 Spark

It is an open-source framework for big data processing with batch and stream analytics capabilities. It supports real time (interactive) processing for large data sets. With Spark, data processing is 10-100 faster than Hadoop [41] since it is adopting an in-memory approach for processing and computation. However, the main focus with Spark is accelerating the processing mechanism by using full in-memory processing and computation which minimizes the number of read/write operations to hard disk by storing the intermediate data in memory. Spark also provides another interesting feature which is its flexibility in which it can be deployed as a standalone framework or it can be integrated with a capable and appropriate storage layer. It is considered a standalone platform if it is integrated with capable and appropriate storage layer, however, it is considered an alternative to MapReduce if it is paired with Hadoop [46] [20]. It has an ecosystem of various libraries that can be utilized for various tasks such as Spark MLlib which provides widely used machine learning algorithms including clustering mechanisms, regression techniques, classification approaches, and reduction of dimensionalities [45]. Spark also supports both processing mechanisms (bath and stream). With batch processing, great speed is offered since all the data is processed in memory and dealing with the disk only occurs at the beginning to load data to the memory and at the end to store the latest results. This in-memory approach is achieved with DAG (Directed Acyclic Graphs) scheduling in which it enables the processor to coordinate the work in a smarter way by providing all the operations to be performed in addition to the data that will be operated on together with the relationships between the operations. Furthermore, (RDDs) Resilient Distributed Datasets is used for in-memory processing for maintaining fault tolerance, avoiding replication [45], and escaping accessing disk with each write operation.

For stream processing, Spark is designed with a strategy known as micro-batches. It deals with the stream of data as multiple small batches that can be managed with batch engine by buffering data streams. This approach results in increasing the throughput but increases the latency as well since the stream needs to wait to be flushed. In summary, Spark is considered a good option for applications with diverse processing workloads [46] [20]. A summary of the advantages and limitations of Hadoop is presented in Table 2.

7.3 Storm

It is a big data framework (an open and free source) developed by Storm for near real-time big data processing. Initially, it started as a BackType platform which is designated for social media analytics [20]. It has the capability to handle very huge amounts of data with very low latency as compared to other solutions. It is considered a suitable choice when the processing time is very critical in which it affects the

Table 2: Advantages and limitations of Hadoop, Spark, and Storm

Tool	Advantages	Limitations
Hadoop	1- Scalability in which it can be expanded with thousands of nodes. 2- Compatibility and integration with other frameworks in which it be used as a core element for other software tools. 3- Cost effective (Economical) since it run on low-cost hardware components.	1- Slow which makes it unsuitable for real time applications.
Spark	1. Fast processing (faster than Hadoop) since it is adopting a strategy for in-memory processing which minimizes the number of read/write operations to hard disk. 2. Flexibility in which it can be used as a separate framework or with an existing Hadoop framework. 3. Support both Stream and Batch processing.	1- Spark can cost more since RAM is more expensive. 2- Not suitable to work on shared clusters since it consumes more resources.
Storm	1- Flexibility to be integrated with Hadoop's YARN 2- Strict latency requirements	1- Does not support batch processing.

experience of the user since it has strict latency requirements. It supports stream processing; however, batch processing is not supported. Currently, it is used in various applications including online machine learning, distributed RPC, and real-time analytics. With regard to stream processing, it is achieved using DAG in addition to topologies which describes the steps that are taken on each data input. The topologies consist of data streams, spouts (data streams sources), and bolts that perform operations on the input data [47]. A summary of the advantages and limitations of Storm is presented in Table 2.

8. CONCLUSION

The tremendous blast of the amounts of data generated every day has led to the appearance of big data analytics as a powerful technology for many organizations. With big data analytics, valuable knowledge, deeper insights, better products, perfect solutions, and smarter and wiser decisions based on scientific conclusions can be obtained. Big data analytics has a promising future in various sectors mainly in health care, business, education, politics, and banking.

This paper provides an overview of big data from various perspectives including the definition, characteristics, and importance. It also discusses the significance of big data analytics for different domains, as well as, the challenges of utilizing big data with the complex and varying characters. Moreover, some of the most current trends for big data analytic techniques are presented. Furthermore, the paper provides a review on big data analytics categories including text, audio, and video. The technical approaches that are used for handling big data complexity such as Hadoop, Spark, and Storm are also discussed.

REFERENCES

1. R. H. Hariri, E. M. Fredericks, and K. M. Bowers, **Uncertainty in big data analytics: survey, opportunities, and challenges**, *Journal of Big Data*, vol. 6, no. 1, p. 44, 2019. https://doi.org/10.1186/s40537-019-0206-3
2. Bluesyemre. **What Happens in an Internet Minute in 2018? (2018 vs. 2017)**. Available: https://bluesyemre.com/2018/05/30/what-happens-in-an-internet-minute-in-2018-2018-vs-2017-infographic-lorilewis-officiallychadd/
3. R. A. A. Habeeb, F. Nasaruddin, A. Gani, I. A. T. Hashem, E. Ahmed, and M. Imran, **Real-time big data processing for anomaly detection: A Survey**, *Int J Inf Manage*, vol. 45, pp. 289-307, 2019.
4. X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, **Data mining with big data**, *IEEE Trans Knowl Data Eng*, vol. 26, no. 1, pp. 97-107, 2013.
5. I. A. T. Hashem et al., **The role of big data in smart city**, *Int J Inf Manage*, vol. 36, no. 5, pp. 748-758, 2016.
6. D. D. Hirsch. **The glass house effect: Big Data, the new oil, and the power of analogy**, *Me. L. Rev*., vol. 66, p. 373, 2013.
7. N. Khan, A. Naim, M. R. Hussain, Q. N. Naveed, N. Ahmad, and S. Qamar. **The 51 V's Of Big Data: Survey, Technologies, Characteristics, Opportunities, Issues and Challenges**, presented at *the Proc. of the Int. Conf. on Omni-Layer Intell. Syst.*, Crete, Greece, 2019. https://doi.org/10.1145/3312614.3312623
8. I. D. Constantiou and J. Kallinikos. **New games, new rules: big data and the changing context of strategy**, *J. Inf. Technol*., vol. 30, no. 1, pp. 44-57, 2015.
9. R. Iqbal, F. Doctor, B. More, S. Mahmud, and U. Yousuf. **Big data analytics: computational intelligence techniques and application areas**, *Technol Forecast Soc Change*, pp. 119253, 2018.
10. C. K. Davis**. Beyond data and analysis**, *Commun. ACM*, vol. 57, no. 6, pp. 39-41, 2014. https://doi.org/10.1145/2602326
11. M. Schroeck, R. Shockley, J. Smart, D. Romero-Morales, and P. Tufano. **Analytics: The real-world use of big data**, *IBM Global Business Services*, vol. 12, no. 2012, pp. 1-20, 2012.
12. C. White. **Using big data for smarter decision making**, BI research, pp. 1-10, 2011.
13. K. Slavakis, G. B. Giannakis, and G. Mateos. **Modeling and optimization for big data analytics:(statistical) learning tools for our era of data deluge**, *IEEE Signal Process Mag*, vol. 31, no. 5, pp. 18-31, 2014.
14. R. Elmasri, Fundamentals of database systems. Pearson Education 2016.
15. IBM Big data and analytics hub. **The Four V's of Big Data**, 2011. [Online]. Available: https://www.ibmbigdatahub.com/infographic/four-vs-big-data
16. C. K. Emani, N. Cullot, and C. Nicolle. **Understandable big data: a survey**, *Comput Sci Rev*, vol. 17, pp. 70-81, 2015.
17. D. P. Acharjya and K. Ahmed. **A survey on big data analytics: challenges, open research issues and tools**, *Int J Adv Comput Sci Appl*, vol. 7, no. 2, pp. 511-518, 2016.
18. S. Kaisler, F. Armour, J. A. Espinosa, and W. Money. **Big data: Issues and challenges moving forward**, in *46th Hawaii International Conference on System Sciences*, 2013, pp. 995-1004: IEEE. Systems, vol. 44, no. 1, pp. 364-379, 2019.
19. J. P. Verma, S. Agrawal, B. Patel, and A. Patel, **Big data analytics: Challenges and applications for text, audio, video, and social media data**, *Int. J. on Soft Comput., Artif. Intell. and Appl. (IJSCAI)*, vol. 5, no. 1, pp. 41-51, 2016.
20. F. Amalina et al. **Blending big data analytics: Review on challenges and a recent study**, *IEEE Access*, vol. 8, pp. 3629-3645, 2019.
21. J. Fan, F. Han, and H. Liu. **Challenges of big data analysis**, *Natl. Sci. Rev*., vol. 1, no. 2, pp. 293-314, 2014.

22. H. Zhu, Z. Xu, and Y. Huang. **Research on the security technology of big data information**, in *4th Int. Conf. on Inf. Technol. and Manage. Innov.*, 2015: Atlantis Press.
23. H. J. Watson. **Update tutorial: Big Data analytics: Concepts, technology, and applications**, *Commun. Assoc. Inf. Syst*., vol. 44, no. 1, pp. 364-379, 2019.
24. J. S. Hurwitz, A. Nugent, F. Halper, and M. Kaufman. ***Big data for for dummies***. John Wiley & Sons, 2013.
25. P. Vashisht and V. Gupta. **Big data analytics techniques: A survey**, in *Int. Conf. on Green Comput. and Internet of Things (ICGCIoT)*, 2015, pp. 264-269: IEEE.
26. W. Chung. **BizPro: Extracting and categorizing business intelligence factors from textual news articles**, *Int. J. of Inf. Manage*, vol. 34, no. 2, pp. 272-284, 2014. https://doi.org/10.1016/j.ijinfomgt.2014.01.001
27. A. S. Abrahams, W. Fan, G. A. Wang, Z. Zhang, and J. Jiao. **An integrated text analytic framework for product defect discovery**, *Prod Oper Manag*, vol. 24, no. 6, pp. 975-990, 2015.
28. R. Chen, Y. Zheng, W. Xu, M. Liu, and J. Wang. **Secondhand seller reputation in online markets: A text analytics framework**, *Decis Support Syst*, vol. 108, pp. 96-106, 2018.
29. C. S. Ishikiriyama, D. Miro, and C. F. S. Gomes. **Text Mining Business Intelligence: a small sample of what words can say**, *Procedia Comput Sci*, vol. 55, pp. 261-267, 2015.
30. T. Taguchi et al. **Major depressive disorder discrimination using vocal acoustic features**, *Journal of affective disorders*, vol. 225, pp. 214-220, 2018.
31. B. Karakus and G. Aydin. **Call center performance evaluation using big data analytics**, in *Int. Symp. on Netw., Comput. and Commun (ISNCC)*, 2016, pp. 1-6: IEEE.
32. N. Bassiou et al. **Privacy-preserving speech analytics for automatic assessment of student collaboration**, in *INTERSPEECH*, 2016, pp. 888-892.
33. H . Qiu et al. **Kestrel: Video analytics for augmented multi-camera vehicle tracking**, in *IEEE/ACM Third Int. Conf. on Internet-of-Things Design and Implementation (IoTDI)*, 2018, pp. 48-59: IEEE
34. D. Singh, C. Vishnu, and C. K. Mohan. **Visual big data analytics for traffic monitoring in smart city**, in 2016 *15th IEEE Int. Conf. on Mach. Learn. and Appl. (ICMLA)*, 2016, pp. 886-891: IEEE.
35. K. P. Seng and L.-M. Ang. **Video analytics for customer emotion and satisfaction at contact centers**, *IEEE Trans Hum Mach Syst*, vol. 48, no. 3, pp. 266-278, 2017.
36. M. Wook, Z. Abdul Jabar, M. Halim, N. Razali, S. Ramli, N. Hasbullah, and N. Zainuddin. **Big Data Analytics Application Model Based on Data Quality Dimensions and Big Data Traits in Public Sector**, Int. J. Adv. Trends Comput. Sci. Eng., Vol. 9, no. 2, pp. 1247-1256, March-April 2020. https://doi.org/10.30534/ijatcse/2020/53922020
37. M. Bhattacharya, R. Islam, and J. Abawajy. **Evolutionary optimization: a big data perspective**, *J. Netw. Comput. Appl.*, vol. 59, pp. 416-426, 2016.
38. A. Mohamed, M. K. Najafabadi, Y. B. Wah, E. A. K. Zaman, and R. Maskat. **The state of the art and taxonomy of big data analytics: view from new big data framework**, *Artif Intell Rev*, vol. 53, no. 2, pp. 989-1037, 2020.
39. H. Shu. **Big data analytics: six techniques**, *Geo Spat Inf Sci*, vol. 19, no. 2, pp. 119-128, 2016.
40. P. Russom, **Big data analytics**, *TDWI best practices report*, fourth quarter, vol. 19, no. 4, pp. 1-34, 2011.
41. D. Singh and C. K. Reddy. **A survey on platforms for big data analytics**, *J. Big Data*, vol. 2, no. 1, p. 8, 2015.
42. S. Šuman, P. Poščić, and M. G. Marković. **Big Data Management Challenges**, *Int. J. Adv. Trends Comput. Sci. Eng*., vol. 9, no. 1, pp. 717-723, January-February 2020. https://doi.org/10.30534/ijatcse/2020/102912020
43. A. Maheshwari, ***Big Data***. McGraw-Hill Education, 2017.
44. N. Elgendy and A. Elragal. **Big data analytics: a literature review paper**, in *Industrial Conference on Data Mining*, 2014, pp. 214-227: Springer.
45. E. M. Saida, E. B. Younès, and H. Nabil. **Towards a reference big data architecture for sustainable smart cities**, Int. J. Adv. Trends Comput. Sci. Eng., vol. 9, no. 1, pp. 820–827, Feb. 2020 https://doi.org/10.30534/ijatcse/2020/118912020
46. H. Guy. ***Next generation databases: NoSQL, newSQL, and big data***, Apress, 2015.
47. V. Gurusamy, S. Kannan, and K. Nandhini. **The real time big data processing framework: Advantages and limitations**, *Int. J. Comput. Sci. Eng*., vol. 5, no. 12, pp. 305-312, 2017 https://doi.org/10.26438/ijcse/v5i12.305312