IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

ANALYSIS ON APPLICATIONS OF BIG DATA ON HANDLING OF BIG DATA PROCESSING METHODS AND TECHNIQUES

Syed Faquaruddin Quadri,Student,Civil engineering, Deccan college of engineering and technology
Hyderabad

ABSTRACT

Large volumes of data are collected from numerous sources, including social networks like Facebook, transactional databases such as online shopping sites, and other data sources, such as weather. Distributed architecture is used to store large amounts of data. Many new social media sites have emerged, as has an increase in the number of digital computers and Internet connections. Emerging trends and warnings of impending disasters can be detected almost immediately if efficient techniques/algorithms are used to evaluate this enormous volume (such as the outbreak of a viral disease). Many relevant socioeconomic and political events can be discovered by diligent data mining, which can aid in the development of efficient public policy. Big data analytics are used in this study to examine how they are used to help people grow. Accordingly, data is a crucial factor in the Big Data situation because of its rapid growth over the last few years. New, high-performance processing is needed to handle the ever-increasing amounts of Big Data. In order to handle and analyse massive amounts of data, a large computational infrastructure is needed. This is a difficult and time-consuming task. It is discussed in this study how pretreatment approaches for data mining can be used to large amounts of data in big data.

1. INTRODUCTION

Big data analytics (ABDA) has captivated the interest of researchers and practitioners during the past decade. According to recent studies, the ABDA is a critical factor in the success of organisations in a wide range of industries. In addition, executives are quickly realising the benefits of the ABDA, which is a major step forward. The expenditure on big data analytics in both the public and private sectors around the world has increased dramatically in recent years. In the commercial world, the ABDA has the potential to be a game changer in terms of both strategy and operational efficiency The ABDA is a major differentiator and significant component for the success of high-performing enterprises.

Analysis of the competitive benefits acquired by firms was offered by experts in big data analytics. The "fourth paradigm" of science, according to some researchers and practitioners, is based on big data analytics. According to some, "big data analytics is the next frontier for innovation, competitiveness, and productivity" and a "new paradigm of knowledge assets". When it comes to business success, high-performing firms look to the ABDA as a major differentiator and growth driver.

"Big data," or the ever-increasing amount of data, presents both a burden and an opportunity to researchers. The ability to handle, store, and analyse large amounts of data has come a long way. Massive amounts of diverse unstructured data are now being generated every day, which can be processed and stored in a distributed fashion using emerging AI and ML technologies. This allows for valuable actionable knowledge to be extracted from the data, which can then be distributed across a cluster of computers. Researchers now have an excellent opportunity to make use of this data in order to get new insights and knowledge.

1.1 Understanding Big Data

Big Data processes a wide variety of data types. Despite the fact that machines produce the systematic outputs, solutions can be generated by humans or machines.

As a result of human engagement with technologies like internet services and digital gadgets, data that is generated by humans exists. Structured data, video, and textual data are all examples of data created by humans. In response to actual world occurrences, software programmes and hardware devices generate machine-generated data. To indicate what a consumer has purchased, the log document records each approved decision made by a security benefit, and a point-of-of-offer framework creates a trade against that stock.

Big Data is also harmed by the above listed drawbacks. As a result of the sheer magnitude of today's data sets and data streams, traditional ways of preparing Big Data are no longer viable. For the purpose of presenting an overview of the current state-of-the-art in data preparation for Big Data, we have gathered the most recent ideas. There are many hurdles to overcome when it comes to preprocessing large datasets, and we also address new technologies and learning paradigms that could be used to solve these problems.

2. LITERATURE REVIEW

There has been an increase in the volume, diversity, accuracy, and timeliness of data being produced in the digital age, leading to the emergence of the term "big data." In order to deal with the complexity and massiveness of diverse forms of data, the organisation was given permission to use a new strategy and new instruments in analytical aspects (structured, semistructured, and unstructured). In order to deal with the complexity of large amounts of data, a sophisticated technique known as "big data analytics" has been developed [1].

Analyzing large amounts of data can help companies become more innovative, productive, and competitive [2]. Big data analytics refers to approaches used to find patterns and uncover surprising relationships in complex contexts by analysing [3,4]. In a knowledge-based society, big data analytics can benefit from reduced complexity and the ability to control cognitive load. In addition, the most important element of big data analytics led to its success was the identification of features. This means that the most critical factors that have a significant impact on the outcome should be identified. Next, correlations between input and a dynamic point that changes over time are identified [3].

Large data sets may be handled and analysed by enterprises using big data analytics. Using big data analytics, social media-generated big data can be handled efficiently and effectively. Analysis of customer behaviour and five key characteristics of big data management (volume; velocity; value; variety; and veracity) are all possible. For businesses, big data analytics not only provides a comprehensive perspective of consumer behaviour, but it also enables firms to be more creative and effective in implementing initiatives. [4]

There is a strong case for the use of big data analytics in the decision-making process, according to Ref. [5]. This means that decisions will be based on strong evidence. The two primary methods used to derive insights from vast amounts of big data are data management and data analytics. The former involves the use of technology to collect, store, and prepare data for analysis, while the latter involves the use of techniques to analyse the data and derive knowledge from it. As a result, big data analytics has been referred to as a component of insight extraction.

It was necessary to broaden the scope of ML and DL to address BDA. Health, the Internet of Things (IOT), and search engines are all examples of industries where predictive analytics (ML) is applied. As a result of this, new data can be predicted based on the patterns learned in previous data. ML relies heavily on the development of features and the representation of data. DL, a subset of ML that is inspired by the human brain and uses neural signals to analyse data, is also being used to extract usable information from large amounts of data [6].

3. BIG DATA ANALYSIS TECHNIQUES

There are three V's that make up big data: the volume of data, how quickly it's handled, and the wide range of data. This expansion of data analytics into machine learning and artificial intelligence may be traced back to the second characteristic, velocity. Traditional statistical methods are also used in conjunction with computer-based data processing techniques. As a result, an organisation uses both real-time streaming data analysis and batch data analysis – to seek for patterns and trends – while conducting big data analysis to better understand its employees' behaviour. As the amount of data grows, so does the variety of methods for managing it. More and more innovation isfueled by data that is more insightful in speed, scale and depth.

In today's data-driven world, everything you do online, from what you watch on Netflix to what you buy, is being tracked and analysed every second.

According to McKinsey's examination of big data approaches and technologies, everything from statistics to computer science and applied mathematics is covered.

They can be applied to both large and small datasets because they draw on so many different areas of study.

1. A/B testing

This data analysis technique compares a control group with a variety of test groups in order to find which treatments or changes are most beneficial. If you'd want to see how to improve your website's conversion rates, McKinsey provides an example of how to do it. To be effective, large-scale studies using big data must include a sufficient number of participants to produce meaningful results.

2. Data fusion and data integration

Combining and analysing data from various sources and solutions can lead to more accurate and efficient insights, as well as potentially more accurate results.

3. Data mining

Statistical and machine learning approaches can be used to extract patterns from large data sets utilising data mining techniques in the database management environment. If you want to find out which demographics are most likely to buy a product, one way to do this is through mining consumer data.

4. Machine learning

Data can be analysed with the help of machine learning, which is widely utilised in the field of intelligent automation. Data-driven hypotheses are generated through the application of algorithms with roots in computer science. It would be impossible for a human to make the forecasts that it does.

5. Natural language processing (NLP).

Computer science, artificial intelligence, and languages all play a role in developing this data analysis tool.

6. Statistics.

To obtain, organise and analyse data, this strategy is commonly employed in surveys and experiments

Many more methods of data analysis exist, such as spatial analysis and a plethora of predictive modelling and association rule learning techniques. An ever-evolving set of technologies is required to effectively process, manage, and analyse this data. Is data beneficial in any form? Regardless of the format or size. In the right hands, it may be a powerful source of information about a company's products, services, and the market as a whole. The future of data mining and analysis is bright, but what can we look forward to? Even if data-driven innovation is unquestionably changing the face of business and society as a whole, the rate at which analytics and technology are evolving is difficult to gauge.

4. BIG DATA PROCESSING TECHNIQUES

4.1 Data preprocessing

In the well-known Knowledge Discovery from Data process, represented in Figure 1, the set of approaches employed before applying a data mining method is referred to as data pretreatment for data mining. Data with inconsistencies and redundancies cannot be used to begin a data mining process since the data is likely to be imperfect. Data generating rates and sizes are likewise rapidly increasing in business, industrial, academic, and scientific applications. The increased volume of data collected necessitates the use of more advanced analysis methods. It is possible to process data that would otherwise be infeasible without preprocessing the data, as each algorithm has its own set of criteria.

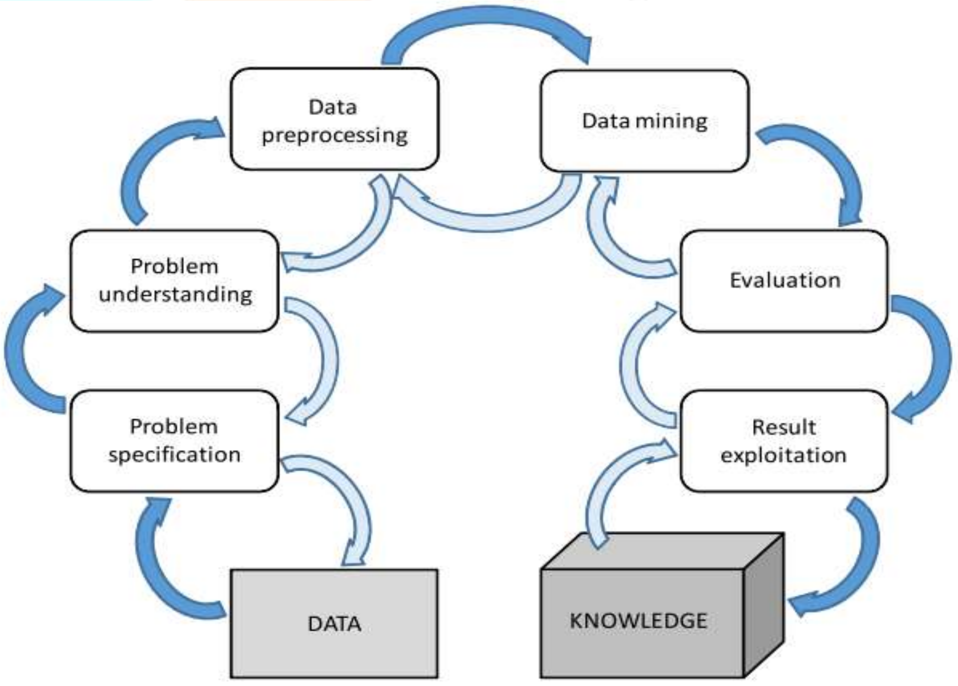

Fig. 1 KDD process

Although data preparation is a useful tool for treating and processing complex data, it can take a long time to run. Fig. 2 shows some of the many disciplines involved in data preparation and data reduction.

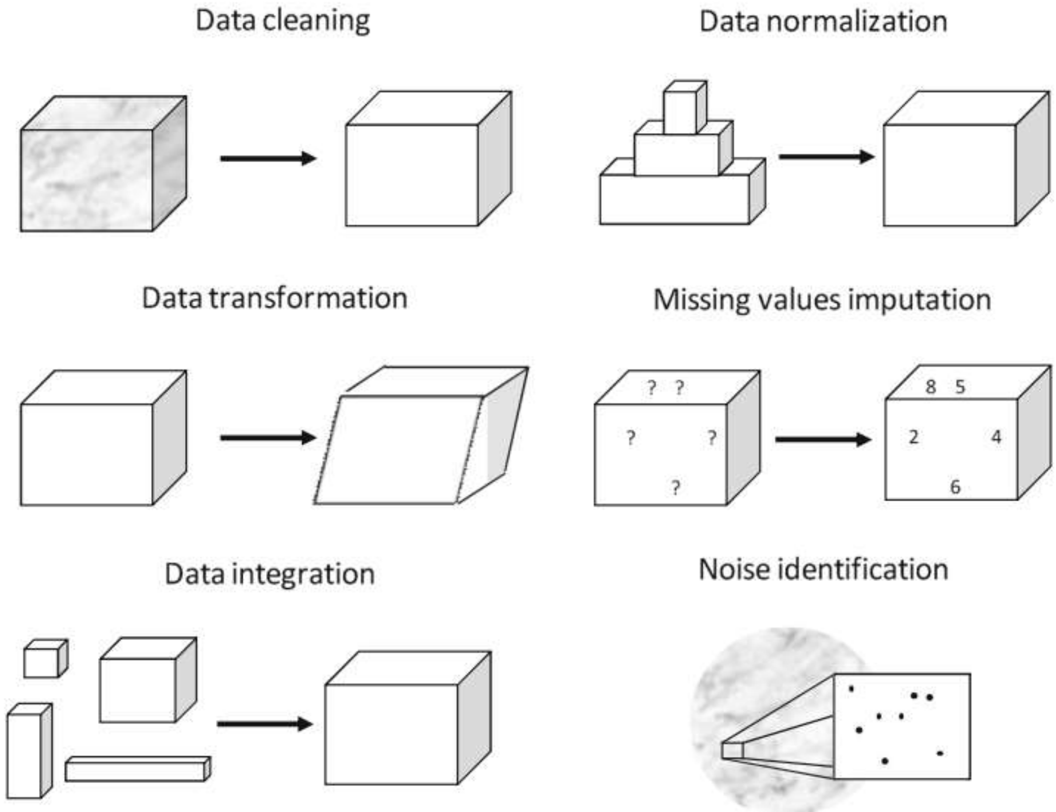

Fig. 2 Data preprocessing tasks

Feature selection and instance selection are all part of the first approach, whereas data normalisation and discretization are the goals of the latter. Any algorithm that uses a valid and appropriate source of information can be confident in its results after successful data preprocessing.

Hadoop

In order to build open-source software for trustworthy, adaptable and dispersed calculating, the Apache Hadoop project was established. MapReduce and Hadoop's distributed file system are the most well-known features of Apache Hadoop programming. HDFS and MapReduce are the two key components of the Hadoop system.

HDFS (Hadoop Distributed File System)

The Hadoop Distributed File System (HDFS) is a fault-tolerant distributed file system designed to run on commodity hardware. Because files are kept in a sequential, redundant fashion in HDFS, it provides an efficient way to access all of this data.

MapReduce Programming Frameworks

MapReduce is a Google-familiar product structure that creates structured data from unstructured data around 2004. Data access is controlled by the MapReduce namespace, which is a programming model namespace. The input key/esteem sets are mapped to a set of key/esteem sets that are in the middle of the road.

The Mapper step:

By dividing the problem into smaller sub-problems, the master hub can then distribute it to the worker hubs. This can be repeated by a worker hub, resulting in a tree structure with a staggered branching pattern. The minor problem is handled by the worker hub, and the correct response is promptly sent to the big master hub.

5. SYSTEM ARCHITECTURE

The term "big data" refers to a collection of methods that elicit a particular sort of integration and are employed to find enormous previously unknown values. It's hard to put into words how complicated these ideals are, and how enormous their scope. Scientists at NASA coined the term "Big Data" in a 1997 study paper. The term "big data" refers to the tremendous amount, velocity, and variety of data. An adage known as "big data" refers to a high volume of data. Structured and unstructured data can be found here. Practicing with conventional methods and equipment is also a challenge. In the IT industry, there is a large amount of big data that is shared by many departments, and no solution existed prior to the advent of big data to manage that data.

Big Data Tools and Techniques

Different stages of big data lifecycle can be categorised by the technologies used for the same purpose listed in Table I. These conclusions are based on their actual use and application.

Table I: Types of Big Data Analytics Tools

	Data Collectiontools	**DataStoragetoolsandframeworks**	**Data filtering andextractiontools**	**Datacleaningandvalidationtools**
1	Semantria	ApacheHBase(Hadoopdatabase)	Scraper	DataCleaner
2	OpinionCrawl	CouchDB	OctoParse	MapReduce
3	OpenText	MangoDB	ParseHub	Rapidminer
4	Trackur	Apachespark	Mozenda	OpenRefine
5	SAS SentimentAnalysis	Oracle,NoSQLDatabase	ContentGrabber	Talend

A. Data collection tools:

There is a slew of Big Data tools available today, there is no denying that. A few of the more popular ones are Semantria, Opinion Crawl, OpenText, and Trackur, among others

B. Data Storage and frameworks tools:

Databases are required to store the collected data, whether it is structured or not. Some databases are needed to handle Big Data. Repositories like Apache and Oracle have built frameworks that can be used as analytics tools for retrieving and processing data stored there.

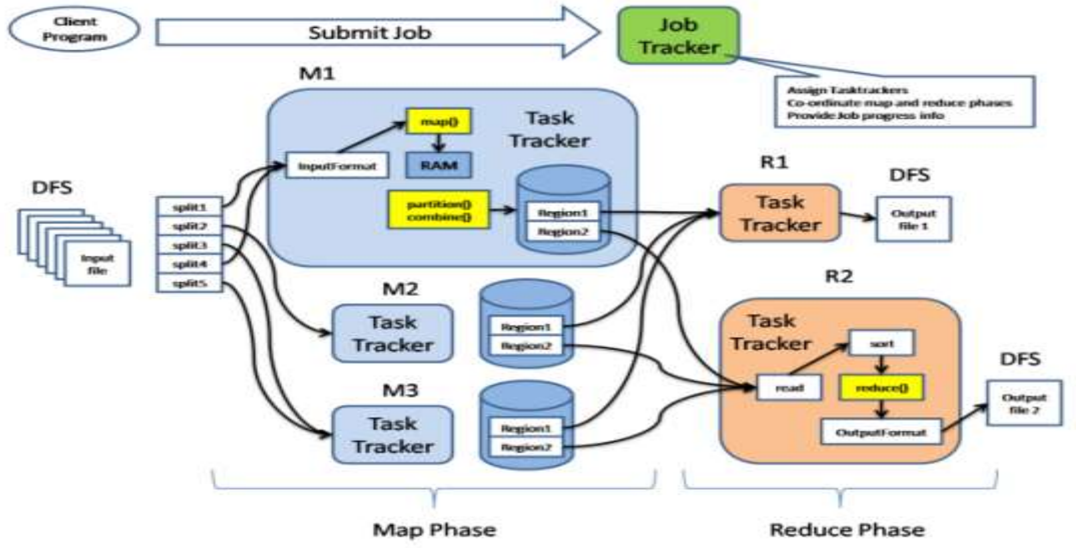

Fig 3. Architecture of Hadoop.

JobTracker is the only node in the Hadoop architecture that serves as the master server. TaskTracker's are slave nodes that run on a variety of machines..JobTracker's primary responsibility is to keep tabs on the slave nodes. It created a system of interfaces that can be used for a variety of functions. An application called JobTracker allows users to submit MR (MapReduce) jobs to be processed. It's first in, first out (FIFO). Figure 3 depicts the workings of Hadoop. JobTracke is in charge of coordinating the execution of the mapper and reducer [5]. To begin the JobTracker's functioning, the Map Task must be completed first. JobTracker now has the responsibility of providing TaskTracker with clear instructions. Afterwards, TaskTracker begins downloading files and combines them into a single entity in most cases (entity).

C. Data filtering and extraction tools:

Various tools are used to perform data extraction and filtering. These programmes are a lifesaver when it comes to sifting through the Internet for relevant data.

D. Data cleaning and validation tools:

An crucial part of the process is data cleansing and validation. When gathering data for analysis, a variety of validation rules are employed to make sure that they are both necessary and relevant. Because of the complexities of the data, it can be difficult to impose validation requirements. Data-cleansing solutions are quite beneficial since they reduce the amount of time it takes to process the data. They also slow down the speed of data analytics tools.

CONCLUSION

A wide range of industries have benefited from big data analytics. Using big data analytics, the healthcare industry has the ability to provide better treatment, save lives, and cut costs. Using customer log files to better understand what their needs are, for example, helps financial organisations better serve their customers. Once again, leveraging big data analytics in the retail industry can help managers better understand people's requirements, leading to the development of new and improved services for customers. Telecommunications companies utilise big data analytics to monitor machine records and handle quality issues. This paper provides an overview of large data analysis. Big data analysis is covered using a variety of tools and methodologies. The implementation costs of Hadoop are likely to be lower than those of competing systems for workloads that are done in batches only. This type of task is not time-critical. If time isn't a problem, Hadoop's instruction execution architecture is best suited for managing enormous datasets. With a wide range of processes, a spark could be an ideal solution. Unprecedented speed gains can be achieved by Spark instruction execution. Spark Streaming may be a smart stream processing option for applications that value throughput over latency. Library, storage, and integrations can all be tailored to meet your specific needs. MongoDB's data integrity and data splitting are two of its strongest features. Big data handling techniques and technologies that can be utilised to manage a large volume of data from numerous sources and increase overall system performance will be presented in this article.

REFERENCES

1. Ning J et al. A best-path-updating information-guided ant colony optimization algorithm. Information Sciences. 2018;433-434:142-162
2. Dong J, Yang C. Business value of big data analytics: A systems-theoretic approach and empirical test. In: Information & Management. 2018.
3. Wang H et al. Randomly attracted firefly algorithm with neighborhood search and dynamic parameter adjustment mechanism. Journal of Soft Computing. 2017;21(18):5325-5339
4. Delice Y et al. A modified particle swarm optimization algorithm to mixed-model two-sided assembly line balancing. Journal of Intelligent Manufacturing. 2017;28(1):23-36
5. Feng L et al. Rough extreme learning machine: A new classification method based on uncertainty measure. Neurocomputing. 2019;325:269-282
6. Harfouchi F et al. Modified multiple search cooperative foraging strategy for improved artificial bee colony optimization with robustness analysis. Soft Computing. 2017;22(19)
7. Patel, Aditya B., Manashvi Birla, and Ushma Nair. "Addressing big data problem using Hadoop and Map Reduce." Engineering (NUiCONE), 2012 Nirma University International Conference on. IEEE, 2012.
8. A, Katal, Wazid M, and Goudar R.H. "Big data: Issues, challenges, tools and Good practices." Noida: 2013, pp. 404 – 409, 8-10 Aug,2013.
9. Amrit pal, PinkiAggrawal, Kunal Jain, Sanjay Aggrawal "A Performance Analysis of MapReduce Task with Large Number of Files Dataset in Big Data using Hadoop" Forth International Conference on Communication Systems and Network Technologies, 2014.
10. Kaur, Anureet. "Big Data: A Review of Challenges, Tools and Techniques." International journal of scientific research in science, engineering and technology 2.2 pp. 1090-1093 ,2016.
11. Verma, Jai Prakash, et al. "Big data analytics: Challenges and applications for text, audio, video, and social media data." International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI) 5.1 2016.
12. Kaki, Gowtham, et al. "Safe Memory Regions for Big Data Processing." transfer (successorId, t, outList) pp. 17 -18 ,2016.
13. Big Data Black Book: Covers Hadoop 2, MapReduce, Hive, YARN, Pig, R and Data Visualization by DT Editorial Services Paperback ,2016.
14. Bhandari, Renu, Vaibhav Hans, and Neelu Jyothi Ahuja. "Big Data Security–Challenges and Recommendations." IJCSE pp. 93- 98 ,2016.
15. Tanuja, A and D. Swetha Ramana. "Processing and Analyzing Big data using Hadoop." International Journal of Computer Sciences and Engineering 4.4 : pp. 91-94 ,2016.