Sarcouncil Journal of Engineering and Computer Sciences

ISSN(Online): 2945-3585

Volume- 04| Issue- 07| 2025

Review Article

Received: 15-06-2025| **Accepted:** 10-07-2025 | **Published:** 25-07-2025

A Review of Big Data Analytics: Tools, Techniques, and Scalability Challenges in Data Science

Venkata Teja Nagumotu[1], Harsha Vardhan Reddy Kavuluri[2], Akhil Kumar Pathani[3], Ajay Dasari[4], Venkata Kishore Chilakapati[5] and Srikanth Reddy Keshireddy[6]

[1]Sr Network Engineer, Techno-bytes Inc
[2]Lead database administrator, Wissen infotech Inc
[3]Network Engineer, Ebay
[4]Senior Support Engineer, Microsoft
[5]Technical Advisor, Microsoft
[6]Senior Software Engineer, Keen Info Tek Inc

Abstract: Big data is the era in which advances in digital technology have made it possible to collect a lot of difficult data. To find useful insights in their massive amounts of structured, semi-structured, and unstructured data, more and more companies are using big data analytics. Innovation, operational optimization, and decision-making are all aided by this. With an emphasis on Hadoop, Spark, and cloud providers like AWS EMR, this overview lays out the fundamentals of big data analytics and its tools and methodologies. The use of analytical tools such as descriptive, predictive, prescriptive, text, sentiment and graph analytics are addressed based on how it helps in identifying patterns, trends and relationships within mega data. The paper also addresses another problem with big data analytics, which is the following: processing speed, data quality, security, heterogeneity, scalability, and fault tolerance. This review highlights the relevance of quality data administration and analytical approaches in the utilization of big data by offering a detailed insight into the state of art tools and methodologies in the subject. It provides information on how organizations can convert complex data to practical knowledge and use it to promote innovation, operational effectiveness, and competitiveness in various areas.

Keywords: Big Data, Data Analytics, Data Science, Hadoop, Apache Spark, Machine Learning, Scalability, Data Processing.

INTRODUCTION

The digital world is experiencing a rapid increase in the volume of data. The proliferation of digital applications, social media, sensors, e-commerce, and the Internet of Things (IoT) is responsible. The proliferation of big data, caused by the quick uptake of new technologies, has sparked ground-breaking innovations across many different sectors. The common belief is that traditional data processing and database management systems are underequipped to handle the massive amounts of complex data that are colloquially referred to as "big data." [Nerella, V. M. L. G. 2018]. These data sets may be structured, semi-structured, as well as unstructured and may often be in petabytes and higher [Acharjya, D. P., & Ahmed, K. 2016]. The rapid expansion of information has posed enormous problems to organizations in terms of storage, handling, and analysis, which necessitates intelligent analytical plans to derive useful insights.

Big data analytics has been a pivotal approach to organizations in order to control, process, and actualize operational insights on such large volumes of data. Data mining is a complex set of techniques used in various fields to find patterns, correlations, and trends in large datasets for the purpose of making informed decisions. These fields include healthcare, business intelligence, finance, and scientific research [Gopinathan, S. 2018]. Big data analytics allows improving organizational performance, innovation, discovering new sources of revenue, and competitive advantage by converting raw data into valuable information. The process centralizes on data science, which is an interdisciplinary area of study, and includes methods of gathering, administering, controlling, visualizing as well as conserving enormous volumes of data. Complementary processes include data mining which aims at analyzing massive data and drawing meaningful patterns and insights that can be used to inform strategy and decision-making.

Organizations use a set of tools and techniques to process large-scale, complex data to carry out effective big data analytics [Hu, H. *et al*., 2014]. Hadoop and Spark are popular in distributed computing, NoSQL databases are popular in flexible data storage, cloud services in scalable infrastructure and enhanced visualization software to graphically reflect insights [Kushwaha, A. *et al*., 2016; Garg, S. 2019]. These tools are used in conjunction with analytical tools to process and

***Corresponding Author:** Venkata Teja Nagumotu

DOI- https://doi.org/10.5281/zenodo.19488398

interpret data including ML, statistical modelling, data mining, and real time streaming analytics. Nevertheless, big data analytics has many challenges such as scalability both in storage, time and resources, data quality, integration, security, privacy and data heterogeneity [Gupta, P. *et al*., 2016]. A resolution to these issues is necessary for data-driven organizations to maximize big data in the modern day.

Structure of the Paper

The paper's structure is as follows: In Section II, we will go over how data scientists use big data analytics, Section III includes big data tools and platforms, Section IV discusses analytical methods of processing big data whereas Section V covers issues of big data, processing and management, Section VI contains a literature review of recent studies, and Section VII provides insights and future research directions.

BIG DATA ANALYTICS IN DATA SCIENCE

A new "digital universe" will be shaped by data as markets and corporations adapt to the ever-increasing amounts of data created, which is projected to reach 180 ZB in 2025. "Big Data" is arrived, ushering in a new era characterized by digital information harvested from various and intricate data sets that may originate from anywhere and at any moment. "Big data" describes datasets that are unmanageably enormous for conventional computer systems to store, process, or analyze efficiently [Pouchard, L. 2015]. There are an enormous amount and complexity of data here, including operational, transactional, sales, and marketing data among others. There are three different kinds of data: structured, semi-structured, and unstructured. There is a vast range of forms included in big data, which includes text, audio, video, images, and more. Data, process, and management issues are the three main categories into which data life cycle problems may be classified (Figure. 1).

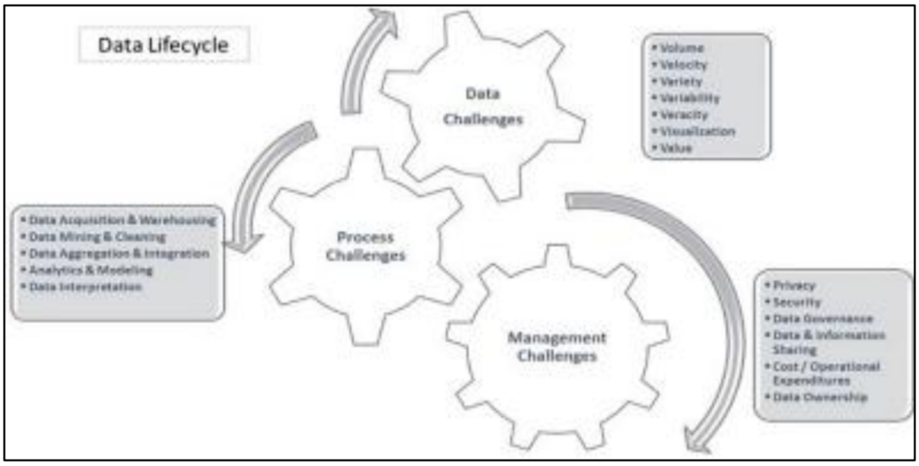

Fig 1: Challenges in Data Lifecycle [Vassakis, K. *et al*., 2017]

Classification of Big Data

The seven Vs—volume, variety, veracity, velocity, variability, visualization, and value—define big data.

- **Volume**: is a measure of how big the datasets are. The exponential growth of data generation can be attributed, in part, to the IoT and the proliferation of linked devices such as smartphones, sensors, and other electronic gadgets, as well as to the fast-paced evolution of ICTs, particularly AI. Data volume and velocity have grown exponentially, surpassing Moore's law, and new units of data storage capacity like exabytes, zettabytes, and yottabytes have evolved to accommodate this trend.
- **Variety**: stands for the ever-expanding variety of data formats and sources of data creation [Tong, H. 2014]. Emerging data formats are a result of Web 3.0's proliferation of internet and social media networks. Customer chats in retail, banking, and online purchasing; social media posts (e.g., Facebook and Twitter); SMS messages; mobile GPS signals; and a plethora of other sources gather this data.
- **Variability**: refers primarily to rapidly shifting meanings, but many people confuse it with diversity. Examples of this include the fact that

Publisher: SARC Publisher

words' meanings can change depending on the surrounding content; so, algorithms need to consider the entire context when determining a word's meaning (sentiment) in order to do correct sentiment analysis.

- **Velocity**: is defined by how quickly material is created. Information is flooding businesses in real time from all over the web and linked gadgets. Businesses rely on this quickness to take a variety of activities that make them more nimble, giving them an edge over competitors [Wamba, S. F. *et al*., 2015]. Big data analytics has made it possible for companies to understand and analyze data instantly.
- **Veracity of data**: means how accurate and reliable the info is. Unclean and inaccurate data is included in the dataset; consequently, data veracity describes the degree of uncertainty and dependability associated with a certain data type.
- **visualization Data**: is the field of data visualization. It provides graphical representations of quantitative and qualitative data that cannot be achieved by other means, such as text or tables, and which reveal trends, patterns, irregularities, consistency, and variety [Malallah, S. *et al*., 2018]. Enterprises, organizations, communities, and consumers may all profit from the data analysis process, which gives important insights that can be leveraged through big data.

BIG DATA ANALYTICS TOOLS AND PLATFORMS

Big data analytics tools, as software components in data science, support the analysis, transformation, and derivation of patterns, correlations, and insights of large, high-dimensional datasets. Platforms are holistic ecosystems, which combine storage, computing systems, and analytics processes, enabling scalable, efficient, and systematic process of big data, to support advanced model and decision-making.

Big Data Tools

Big data tools are computer programs and libraries that compute, study and depict huge volumes of data effectively. Some are given:

Hadoop

One example of an open-source framework is Hadoop. This Java-based project made its debut in October 2003. The ability to manage and handle enormous datasets by means of computer clusters employing basic software models is a significant benefit that Hadoop has provided. Hadoop is an application framework that aims to increase the processing and storage capacity of massive data sets by using thousands of computers instead of just one node [Hannan, S. A. 2016]. Hadoop relies on its data storage component, HDFS, which is one of its two primary features. Data processing based on the MapReduce paradigm is the responsibility of the second. The core of the work is built on using MapReduce to partition huge data files into smaller ones, which are then distributed across the machines in the cluster for parallel processing. As seen in Figure 2, the following four modules make up the Hadoop paradigm's basis.

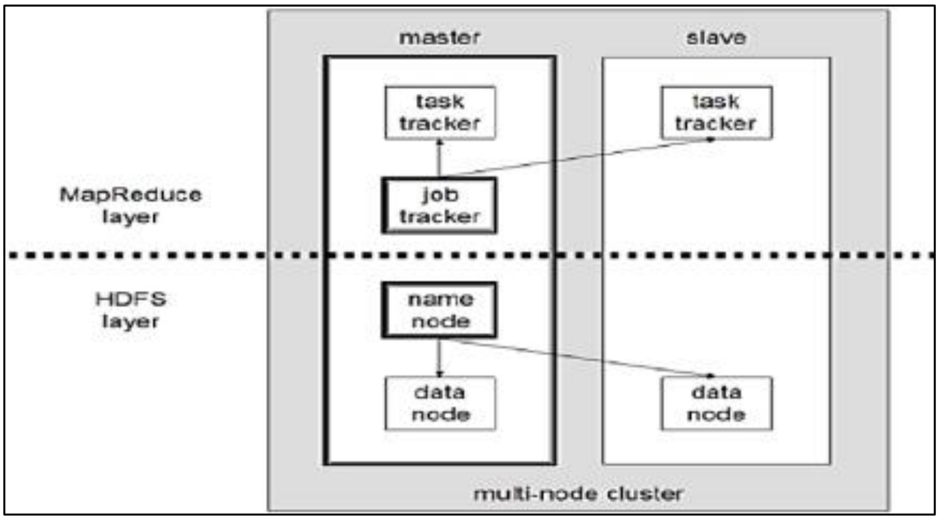

Fig 2: Hadoop Architecture [Mukhopadhyay, D. *et al*., 2014]

Publisher: SARC Publisher

Apache Spark

Apache Spark is an efficient and versatile engine for processing massive amounts of data, according to its official website. Compared to Hadoop MapReduce, its performance is superior, with memory speeds that are up to 100 times faster and storage speeds that are 10,000 times faster. Making it easy to construct parallel programs, it may be used interactively from the Python or Scala shells [V. K. S. V. 2019]. It comes with an impressive set of advanced capabilities for streaming, SQL, and advanced analytics. It can access multiple data sources, such as S3, HDFS, Cassandra, etc., and runs anywhere from Hadoop and Mesos to the cloud and standalone. Very high processing speed is required for the enormous data sets to be meaningful. Figure 3 shows the Apache Spark stack's general design.

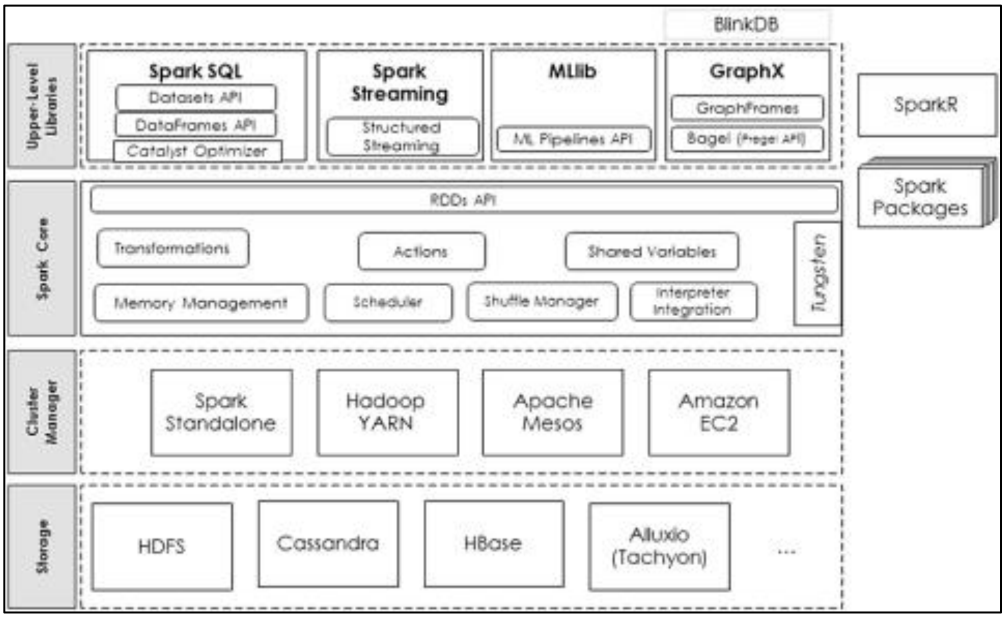

Fig 3: Architecture of Apache Spark stack [Salloum, S. *et al*., 2016]

Big Data Platforms

Big data systems offer an integrated platform that allows storage, processing, and analytics of large-scale data. They are in favor of distributed computing, real time processing, and scalable infrastructure.

Apache Hadoop Ecosystem

Big data may be easily processed, stored, and analyzed with Hadoop, an open-source platform. The core components of Hadoop are HDFS, which allows for fault-tolerant and scalable data storage, YARN, which coordinates cluster resources, and MapReduce, which allows for parallel processing of batch data [Landset, S. *et al*., 2015]. These central elements are been surrounded by ecosystem tools that expound the functionality of Hadoop. Hive supports SQL-like querying and data warehousing and Pig makes it easy to transform data with the help of the scripting language. HBase provides real time and column NoSQL storage on HDFS. Information ingestion using Sqoop and Flume, and orchestration of the work processes using Oozie and ZooKeeper. Figure 4 provide the whole ecosystem of the Hadoop.

Publisher: SARC Publisher

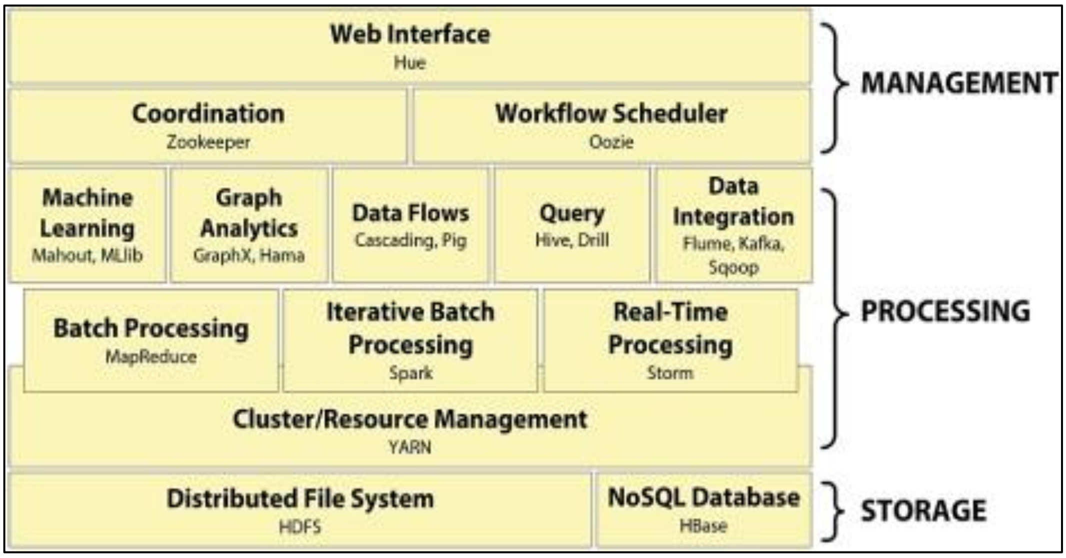

Fig 4: Hadoop Ecosystem [Landset, S. *et al*., 2015]

AWS EMR

Amazon Elastic MapReduce (AWS EMR) is an Amazon Web Services managed big data analytics platform, which facilitates the creation and execution of distributed data processing systems on the cloud [Mohan, J. S., & Shanmugapriya, P. 2016]. The scalability of batch and real-time analytics is made easier with its support for prominent open-source technologies like Apache Hadoop, Apache Spark, Apache Hive, and Apache HBase, as shown in Figure 4. AWS EMR enables companies to handle huge amounts of structured and unstructured information utilizing elastic clusters that may be expanded or scaled down in response to workload demands. EMR works perfectly with Amazon S3 buckets to support low-cost storage of data and high-throughput access to the information required to perform analytics. The platform lowers the operating costs through automated cluster provisioning, configuration, monitoring and fault tolerance. Moreover, EMR is a system that promotes security and governance using IAM-based access control and encryption.

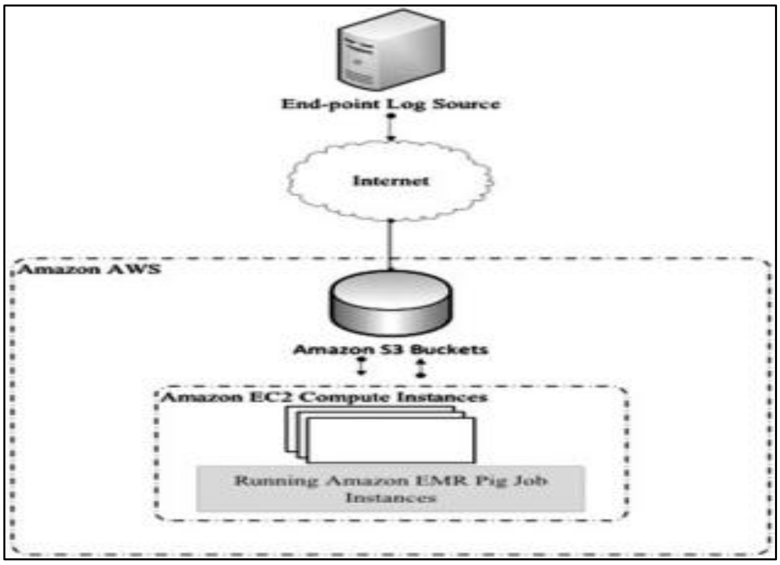

Fig 5: AWS EMR Flow [Guide, D. 2010]

The comparison Table I shows the differences between big data analytics tools and platforms in terms of functional role, scope, integration, scalability, deployment, management, and examples, with a focus on their complementary nature and unique operational features in the analytics workflow.

Publisher: SARC Publisher

Table1: Comparison of Big Data Analytics Tools and Platforms

/	Big Data Analytics Tools	Big Data Analytics Platforms
Functional Role	Execute specific analytics operations such as data storage, processing, querying, or visualization within a big data pipeline	Provide a unified and coordinated environment to design, execute, and monitor complete big data analytics workflows
Scope	Focus on individual tasks or stages of the analytics process with limited cross-functional coverage	Encompass the entire analytics lifecycle, including data ingestion, storage, processing, analytics, and result delivery
Integration	Require manual configuration or custom connectors to interact with other tools and data sources	Offer built-in integration, workflow orchestration, and interoperability among multiple components
Scalability	Support scalability at the component or service level through distributed execution	Enable system-wide, elastic scalability across clusters and cloud infrastructures
Deployment	Can be deployed independently or combined into customized toolchains	Deployed as integrated solutions in on-premise, cloud, or hybrid environments
Management	Demand separate administration, monitoring, and maintenance efforts for each tool	Provide centralized management, monitoring, and automation to reduce operational complexity
Examples	Apache Hadoop, Apache Spark	Hadoop Ecosystem, AWS EMR

ANALYTICAL TECHNIQUES FOR BIG DATA PROCESSING

"Big data analytics" means analyzing large and complicated data sets to find previously unseen connections, trends, and insights that can help in decision-making. Traditional data processing methods aren't necessarily up to the task of handling the volume, velocity, and variety of big data. As a result, there has developed specialized methods of analysis to overcome such obstacles so that organizations gain value out of their data effectively and correctly.

Descriptive Analytics

Big data processing begins with descriptive analytics, which seek to understand about what has been happening historically by summarizing the prior data. Some of the techniques in this category are statistical analysis, aggregation, data visualization and reporting [Wang, G. *et al*., 2016]. HDFS and SQL based big data applications are common tools that enable analysts to manipulate and analyze structured and semi-structured data.

Predictive Analytics

Using machine learning algorithms and past data, predictive analytics may foretell what's to come. Some of the techniques frequently used include regression analysis, DTs, RFs, SVMs and NNs. Predictive models are applicable in big data environments, when working with high-dimensional and massive data sets in frameworks such as Apache Spark MLlib or TensorFlow. Some of the applications are customer churn prediction, demand forecasting, fraud detection and healthcare diagnostics.

Prescriptive Analytics

Prescriptive analytics extends past forecasting the future by prescribing viable plans of action to maximize the outcomes. This method frequently combines optimization methods, simulation modelling and sophisticated machine learning methods to propose the most optimum course of action [Krumeich, J. *et al*., 2016]. At the same time, such techniques as linear programming, genetic algorithms, and reinforcement learning are typical. In such sectors as logistics, finance, and energy management.

Text and Sentiment Analytics

Narrative data comprise a large part of the big data, including social media, customer feedback, emails, or the content of websites. Text analytics techniques like topic modeling, sentiment analysis, and NLP can glean useful insights from various sources [Sharef, N. M. *et al*., 2016]. The use of such tools as NLTK, spaCy, and BERT-based models can enable organizations to analyze the trends, opinion of customers, and new topics.

Graph Analytics

Graph analytics is a relationship-oriented analytics that represents entities in the form of nodes and connections in the form of edges. Network analysis, community detection and shortest path algorithms are some of the techniques that are used to comprehend complex relationships [Nisar, M. U. *et al*., 2013]. Graph databases like Neo4j and

Publisher: SARC Publisher

Spark frameworks like GraphX support scalable processing. Social network analysis, recommendation systems, supply chain optimization, and fraud detection are some of its applications.

Hybrid and Advanced Techniques

The new generation of big data analytics tends to integrate a variety of techniques to gain a deeper insight. Hybrid methods combine machine learning or deep learning and AI-based algorithms with old-fashioned statistical techniques to process highly-complex and heterogeneous data. Reinforcement learning, batch processing, picture data CNNs, and sequential data RNNs.

CHALLENGES IN BIG DATA ANALYTICS

Big Data is still a very new and undeveloped area of study, as is well known. Big Data is easy to understand and implement because of its inherent qualities, yet these same qualities have also become a source of difficulty. These traits make it difficult to extract valuable information from. Despite these obstacles, researchers, data scientists, and organizations must work tirelessly to turn Big Data into a scientific revolution. Problems with data, processing, and administration are the three basic types of big data obstacles.

Data Challenges

- The Big Data is the result of the enormous amount of data that is obtained every second. Nowadays, petabytes and zettabytes are the units of data generation. This exponential growth in data output necessitates state-of-the-art methods and technologies, paving the way for Big Data. Data scientists have a significant issue when faced with such massive amounts of data.
- Business systems examine social networking websites to gather messages that trigger the verdict to acquire or sell shares in milliseconds, as can see from the millions of images submitted by Facebook users every second and the rapidity with which messages become viral on social media. There is no way to inspect the data as it is being transmitted due to the Big Data streaming processing approach; all data must be stored in a database at some point.
- Photographs, songs, texts, e-mails, medical records, weather reports, log files, and a plethora of other data kinds comprise the backbone of Big Data [Fan, J. *et al.*, 2014]. Thus, the data generated falls into several types, including raw, unstructured, structured, and semi-structured data, all of which appear to be extremely challenging to manage.
- Decision making, knowledge extraction, analytics, and statistical procedures are where the value is concentrated. Without analysis and processing, the data that is created is meaningless.

Processing Challenges

Processing data is something that every part of a company does. Today, data is gathered from a wide variety of sources, such as social media, messaging apps, mobile phones, sensors, security cameras, etc. The greater the amount of data gather, the better the results that can be optimized. It takes cutting-edge equipment, sophisticated algorithms, and a wide range of methods to handle this mountain of data. Data processing includes data collection, repairing similarities discovered in different sources, data refining to a kind suitable for examination, and output characterization [Ji, C. *et al.*, 2012]. Let's examine the many difficulties associated with Big Data processing.

- Heterogeneity. The term "heterogeneity" is used to describe the existence of several types of data in relation to Big Data. Massive amounts of multidimensional data are produced by Big Data's endless range of data sources. Many different sources and formats contribute to the data. In Big Data, heterogeneity means dealing with several types of data at the same time.
- Timeliness. Quickness refers to the rate of data processing and analysis. It undoubtedly take more time to evaluate a larger data collection that needs refining. Nevertheless, there are instances where the evaluation's result is required promptly.
- Complexity. A data set's rigidity is a measure of its complexity. Unstructured data is its primary emphasis. Zikopoulos and Eaton state that there are three distinct kinds of multidimensional data: semi-structured, structured, and unstructured and [Zheng, Z. *et al.*, 2015]. Structured data is generated automatically by devices like computers or sensors without any input from the user; it is

Publisher: SARC Publisher

homogenous and has predefined lengths. Reviews, images, and other forms of unstructured multimedia are examples of complicated unstructured data.

- Scalability. A lot of data is being created every day, and it's getting bigger and bigger every day. Enhancing the processing speed can slow down this sudden surge in data flow. Unfortunately, both the amount of data and the processing speeds of computers are growing at an exponential rate. Need to construct a highly scalable system that can accommodate all of these factors, including persistent storage, fast processing speeds, and the ability for a single node to share several hardware resources.

Management Challenges

Data management systems are in dire need of an overhaul since they are ill-equipped to handle the massive amounts of data produced by Big Data and the exponential growth in the amount of data stored. And because data is inherently diverse, current algorithms fall short when it comes to efficiently storing data that has been directly retrieved from several sources. What follows is a list of some of the most pressing issues in Big Data management:

➢ Security. There has to be a system in place to detect maliciously implanted data since Big Data is outstripping the data sources it can consume. This system must be able to trust and rely on each source [Gil, D., & Song, I. Y. 2016]. When massive amounts of data are connected, analyzed, and retrieved for meaningful patterns, information security is quickly becoming an important Big Data problem.

➢ Data sharing. The era of Big Data is here now. These days, everyone is sharing everything on social media. With just one click on "Google," can get information on practically anything. Countless pieces of information are readily available to every person and business, ready to be utilized according to their needs and goals. The only way for any form of content to be easily accessible is for everyone to share it. However, there has to be a clear demarcation between private and shareable information.

➢ Failure handling. Creating a system that is completely dependable is definitely not an easy undertaking. It is possible to bring systems up to a point where the likelihood of failure is within an acceptable range. Starting a process necessitates setting up several nodes, which makes the whole working out procedure cumbersome. The two most difficult things to do are to set a failure threshold and to maintain control points. When discussing Big Data, this is referred to as fault tolerance.

LITERATURE REVIEW

The literature review examines the methods, tools, and uses of big data analytics, paying special attention to the evolving concerns of scalability, architectural limitations, and knowledge gaps in various data-driven research domains.

Nwaimo, Oluoha and Oyedokun (2019) discuss the technologies underpinning the BDA, such as distributed computing and advanced machine learning, which allow identifying insights in large volumes of information. The integration of structured and unstructured data, storage (HDFS) and processing (Apache Spark) engines have been discussed in the paper adding more value to real-time analytics. The study covers both achievements and failures of the existing BDA practices with respect to data governance and ethical aspects. Also, it points to future trends, such as quantum analytics and edge computing, which might transform the approach to big data [Nwaimo, C. S. *et al*., 2019].

Rao *et al*. (2019) case studies on distributed ML technologies like Mahout, Spark MLlib, and FlinkML are probed in this study. Analyze data types, domains, and applications to categorize analytics; compare visualization solutions according to three criteria: features, analytical prowess, and development environment compatibility. Additionally, look at big data technologies and tools, including cloud-based and distributed stream processing tools, in depth and compare them. Moreover, go over the features of several SQL Query tools on Hadoop according to 10 factors. At last, make a few important observations regarding the prospects and paths for study that are related to the present trend of big data. may learn more about how various tools and technologies work to tackle real-world problems by looking at big data infrastructure tools with current advancements [Rao, T. R. *et al*., 2019].

Desai (2018) describes big data as a massive, heterogeneous, and fast-expanding object that needs new technologies to manage it. The paper

Publisher: SARC Publisher

emphasizes its importance in decision making in several sectors, scientific exploration, business output and development of the services to the people. As much as big data provides a lot of opportunities including improvement of healthcare to national security, it also creates challenges because it is diverse, vast, and fast. The paper discusses the principles of big data, its uses and challenges that are inherent in it [Desai, P. V. 2018].

Khullar *et al*. (2018) introduces novel analytical and computational obstacles, such as measurement inaccuracies, storage bottlenecks, and noise build-up. Big Data files are stored in HDFS, due to a particular quality of the data (HDFS). Hadoop, on the other hand, is objectively complicated. Users of Hadoop are in their infancy, thus this study delves into the major problems encountered throughout data mining and Hadoop adoption [Khullar, R. *et al*., 2018].

Grover and Kar (2017) examine big data studies published in high-quality business management journals, together with the big data analysis and visualization tools, to extract the dominant themes and relationships. Their primary method of data separation is industry; specifically, they zero in on discretionary spending by consumers and government spending. The articles centre around the methodological foundation of machine learning, text mining, and social media analytics with an aim of supply chain management and marketing. They find however a hole in the presentation of the tools of analysis. The paper fills this gap by talking about the development and the various types of big data tools, classifying them into a platform, database, programming language, search tools and aggregation tools and its uses [Grover, P., & Kar, A. K. 2017].

Park *et al*. (2017) An HPC database analytics system is presented in this paper. The Apache Spark framework allows for quick in-memory processing of the log data, while distributed NoSQL database technology provide scalability and high availability. System administrators and end users can benefit from the analytics framework's ability to extract various system data points, which helps them meet their individual needs. This article describes the author's experience with using the analytics framework to extract system behavior insights from Titan supercomputer logs at Oak Ridge National Laboratory [Park, B. H. *et al*., 2017].

Table II summarizes the previous research studies by the focus, methods, contribution, challenges, and recommendations, stating that a significant problem of scaling remains unsolved.

Table 2: Comparative Analysis of Big Data Analytics Studies

Author(s)	Focus	Methods	Contributions	Challenges Identified	Recommendations
Nwaimo, Oluoha & Oyedokun (2019)	Core BDA technologies and applications	Conceptual analysis and case-based review	Comprehensive overview of BDA tools, architectures (HDFS, Spark), and cross-domain applications	Limited empirical validation and scalability benchmarking	Need for quantitative evaluation frameworks and performance-driven scalability studies
Rao *et al*. (2019)	Distributed ML and big data tools	Comparative analysis of platforms and tools	Detailed comparison of ML libraries, Hadoop ecosystem, and visualization tools	Lack of real-world deployment and stress testing	Development of unified decision models for tool selection under scalability constraints
Desai (2018)	Fundamentals and applications of big data	Literature-based conceptual study	Summarizes big data characteristics, applications, and general	High-level discussion without technical depth	Empirical studies linking big data challenges to concrete analytics performance metrics

Publisher: SARC Publisher

			challenges		
Khullar *et al*. (2018)	Hadoop-based big data processing	Analytical study of Hadoop architecture	Identifies computational, storage, and noise-related challenges in HDFS	Over-reliance on Hadoop; limited architectural diversity	Exploration of alternative and hybrid architectures for scalable analytics
Grover & Kar (2017)	Big data research trends and tools	Bibliometric and thematic analysis	Categorization of big data tools and business use cases	Weak emphasis on technical scalability mechanisms	Integration of business analytics needs with scalable system design
Park *et al.* (2017)	Scalable log analytics	Framework implementation using Spark and NoSQL	Demonstrates scalable analytics in HPC environments	Domain-specific focus limits generalizability	Need for cross-domain scalable analytics frameworks

CONCLUSION & FUTUREWORK

Big data analytics have become an important part of modern data science. They help businesses find useful information in large amounts of different types of data. The review further reveals that tools like Hadoop, Apache Spark, and cloud-based services such as AWS EMR are very important in managing, processing, and analyzing massive data. Descriptive, predictive, prescriptive, text, sentiment, and graph analytics represent analytical methods that provide powerful solutions to the discovery of patterns, trends, and correlations. Although these developments have been made, it has some major challenges especially when it comes to scaling, data heterogeneity, timeliness, security and fault tolerance. The solution to these challenges is to keep advancing innovations in distributed computing systems, hybrid analytics systems, and automated data management systems. The future studies ought to look into coming up with scalable and real time analytics solutions that can be used to integrate various data sources without compromising on privacy, accuracy and efficiency. The new technologies like edge computing, quantum analytics, and AI-based adaptive systems hold potential opportunities in addressing the existing constraints. Organizations in many fields, including healthcare, finance, business intelligence, and scientific research, can benefit from big data analytics by filling in these gaps and using the raw data as a strategic resource. This, in turn, leads to innovations, improved operations, and competitive advantages.

REFERENCES

1. Nerella, V. M. L. G. "Automated cross-platform database migration and high availability implementation." Turkish Journal of Computer and Mathematics Education (TURCOMAT) ISSN 3048 (2018): 4855.
2. Acharjya, D. P., & Ahmed, K. "A survey on big data analytics: challenges, open research issues and tools." International Journal of Advanced Computer Science and Applications 7.2 (2016): 511-518.
3. Gopinathan, S. "A Review on Concepts, Applications, Challenges and Future Scope in Big Data." (2018).
4. Hu, H., Wen, Y., Chua, T. S., & Li, X. "Toward scalable systems for big data analytics: A technology tutorial." IEEE access 2 (2014): 652-687.
5. Kushwaha, A., Pathak, P., & Gupta, S. "Review of Optimize Load Balancing Algorithms in Cloud." International Journal of Distributed & Cloud Computing 4.2 (2016).
6. Garg, S. "Predictive analytics and auto remediation using artificial inteligence and machine learning in cloud computing operations." Available at SSRN 5267117 (2019).
7. Gupta, P., Sharma, A., & Jindal, R. "Scalable machine-learning algorithms for big data analytics: a comprehensive review." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 6.6 (2016): 194-214.
8. Pouchard, L. "Revisiting the data lifecycle with big data curation." International Journal of Digital Curation 10.2 (2015): 176-192.

Publisher: SARC Publisher

9. Vassakis, K., Petrakis, E., & Kopanakis, I. "Big data analytics: applications, prospects and challenges." Mobile big data: A roadmap from models to technologies (2017): 3-20.
10. Tong, H. "Big Data Classification." (2014): 275-286.
11. Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. "How 'big data'can make big impact: Findings from a systematic review and a longitudinal case study." International journal of production economics 165 (2015): 234-246.
12. Malallah, S., Zalah, Y., & Karne, R. "An Analysis of the Advanced Encryption Standard and Threats Associated." (2018)
13. Hannan, S. A. "An overview on big data and hadoop." International Journal of Computer Applications 154.10 (2016).
14. Mukhopadhyay, D., Agrawal, C., Maru, D., Yedale, P., & Gadekar, P. "Addressing name node scalability issue in Hadoop distributed file system using cache approach." 2014 International Conference on Information Technology. IEEE, (2014).
15. V. K. S. V, "Comparative Analysis of Hadoop and Apache Spark," Int. J. Trend Res. Dev., (2019): 47–50,
16. Salloum, S., Dautov, R., Chen, X., Peng, P. X., & Huang, J. Z. "Big data analytics on Apache Spark." International Journal of Data Science and Analytics 1.3 (2016): 145-164.
17. Landset, S., Khoshgoftaar, T. M., Richter, A. N., & Hasanin, T. "A survey of open source tools for machine learning with big data in the Hadoop ecosystem." Journal of big data 2.1 (2015): 24.
18. Mohan, J. S., & Shanmugapriya, P. "PROGRESSIVE DATA ANALYTICS IN HEALTH INFORMATICS USING AMAZON ELASTIC MAPREDUCE (EMR)." ICTACT Journal on Soft Computing 6.3 (2016).
19. Guide, D. "Amazon elastic mapreduce." (2010).
20. Wang, G., Gunasekaran, A., Ngai, E. W., & Papadopoulos, T. "Big data analytics in logistics and supply chain management: Certain investigations for research and applications." International journal of production economics 176 (2016): 98-110.
21. Krumeich, J., Werth, D., & Loos, P. "Prescriptive control of business processes: new potentials through predictive analytics of big data in the process manufacturing industry." Business & Information Systems Engineering 58.4 (2016): 261-280.
22. Sharef, N. M., Zin, H. M., & Nadali, S. "Overview and Future Opportunities of Sentiment Analysis Approaches for Big Data." J. Comput. Sci. 12.3 (2016): 153-168.
23. Nisar, M. U., Fard, A., & Miller, J. A. "Techniques for graph analytics on big data." 2013 IEEE International Congress on Big Data. IEEE, (2013).
24. Fan, J., Han, F., & Liu, H. "Challenges of big data analysis." National science review 1.2 (2014): 293-314.
25. Ji, C., Li, Y., Qiu, W., Jin, Y., Xu, Y., Awada, U., & Qu, W. "Big data processing: Big challenges and opportunities." Journal of Interconnection Networks 13.03n04 (2012): 1250009.
26. Zheng, Z., Wang, P., Liu, J., & Sun, S. "Real-time big data processing framework: challenges and solutions." Applied Mathematics & Information Sciences 9.6 (2015): 3169.
27. Gil, D., & Song, I. Y. "Modeling and management of big data: challenges and opportunities." Future Generation Computer Systems 63 (2016): 96-99.
28. Nwaimo, C. S., Oluoha, O. M., & Oyedokun, O. Y. E. W. A. L. E. "Big data analytics: technologies, applications, and future prospects." Iconic Research and Engineering Journals 2.11 (2019): 411-419.
29. Rao, T. R., Mitra, P., Bhatt, R., & Goswami, A. "The big data system, components, tools, and technologies: A survey." Knowledge & Information Systems 60.3 (2019).
30. Desai, P. V. "A survey on big data applications and challenges." 2018 second international conference on inventive communication and computational technologies (ICICCT). IEEE, (2018)
31. Khullar, R., Sharma, T., Choudhury, T., & Mittal, R. "Addressing challenges of Hadoop for big data analysis." 2018 International conference on communication, computing and internet of things (IC3IoT). IEEE, (2018).
32. Grover, P., & Kar, A. K. "Big data analytics: A review on theoretical contributions and tools used in literature." Global Journal of Flexible Systems Management 18.3 (2017): 203-229.
33. Park, B. H., Hukerikar, S., Adamson, R., & Engelmann, C. "Big data meets hpc log

Publisher: SARC Publisher

analytics: Scalable approach to understanding systems at extreme scale." 2017 IEEE International Conference on Cluster Computing (CLUSTER). IEEE, (2017).
34. Gupta, A. K., Polu, A. R., Narra, B., Buddula, D. V. K. R., Patchipulusu, H. H. S., & Vattikonda, N. "Leveraging deep learning models for intrusion detection systems for secure networks." Journal of Computer Science and Technology Studies 6.2 (2024): 199-208.
35. Narra, B., Buddula, D. V. K. R., Patchipulusu, H., Vattikonda, N., Gupta, A., & Polu, A. R. "The integration of artificial intelligence in software development: Trends, tools, and future prospects." Available at SSRN 5596472 (2024).
36. Achuthananda, R. P., Bhumeka, N., Dheeraj Varun Kumar, R. B., Hari Hara, S. P., & Navya, V. "Evaluating machine learning approaches for personalized movie recommendations: A comprehensive analysis." J Contemp Edu Theo Artific Intel: JCETAI-115 (2024).
37. Waditwar, P. "The Intersection of Strategic Sourcing and Artificial Intelligence: A Paradigm Shift for Modern Organizations." Open Journal of Business and Management 12.06 (2024): 4073-4085.
38. Bitkuri, V., Kendyala, R., Kurma, J., Mamidala, J. V., Attipalli, A., & Enokkaren, S. J. "A Survey on Blockchain-Enabled ERP Systems for Secure Supply Chain Processes and Cloud Integration." International Journal of Technology, Management and Humanities 10.04 (2024): 126-135.
39. Mamidala, J. V., Bitkuri, V., Attipalli, A., Kendyala, R., Kurma, J., & Enokkaren, S. J. "Machine Learning Approaches to Salary Prediction in Human Resource Payroll Systems." Journal of Computer Science and Technology Studies 6.5 (2024): 341-349.
40. Waditwar, P. "AI for Bathsheba Syndrome: Ethical Implications and Preventative Strategies." Open Journal of Leadership 13.03 (2024): 321-341.
41. Attipalli, A., Kendyala, R., Kurma, J., Mamidala, J. V., Bitkuri, V., & Enokkaren, S. J. "Privacy Preservation in the Cloud: A Comprehensive Review of Encryption and Anonymization Methods." International Journal of Multidisciplinary on Science and Management IJMSM 1.1 (2024).
42. Tamilmani, V., Maniar, V., Singh, A. A., Kothamaram, R. R., Rajendran, D., & Namburi, V. D. "A Review of Cyber Threat Detection in Software-Defined and Virtualized Networking Infrastructures." International Journal of Technology, Management and Humanities 10.04 (2024): 136-146.
43. Singh, A. A. S., Kothamaram, R. R., Rajendran, D., Deepak, V., Namburi, V. T., & Maniar, V. "A Review on Model-Driven Development with a Focus on Microsoft PowerApps." International Journal of Humanities, Science Innovations and Management Studies 1.1 (2024): 43-56.
44. Padur, S. K. R. "AI-augmented platform engineering: Redefining developer experience through autonomous, self-optimizing enterprise systems." International Journal of Science, Engineering and Technology (2024).
45. Gangineni, V. N., Tyagadurgam, M. S. V., Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., & Chalasani, R. "AI-Powered Cybersecurity Risk Scoring for Financial Institutions Using Machine Learning Techniques (Approved by ICITET 2024)." Journal of Artificial Intelligence & Cloud Computing (2024).
46. Sagili, S. R., Goswami, C., Bharathi, V. C., Ananthi, S., Rani, K., & Sathya, R. "Identification of Diabetic Retinopathy by Transfer Learning Based Retinal Images." 2024 9th International Conference on Communication and Electronics Systems (ICCES). IEEE, (2024).
47. Sagili, S. R., & Kinsman, T. B. "Drive dash: Vehicle crash insights reporting system." 2024 International Conference on Intelligent Systems and Advanced Applications (ICISAA). IEEE, (2024).
48. Padur, S. K. R. "Securing Oracle Integration Cloud ERP ecosystems, zero trust architecture, data governance, and compliance automation." International Journal of Science, Engineering and Technology 12.4 (2024): 10-5281.
49. Sagili, S. R., Chidambaranathan, S., Nallametti, N., Bodele, H. M., Raja, L., & Gayathri, P. G. "NeuroPCA: Enhancing Alzheimer’s disorder Disease Detection through Optimized Feature Reduction and Machine Learning." 2024 Third International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT). IEEE, (2024).

Publisher: SARC Publisher

50. Sagili, S. R., Veeranjaneyulu, K., Puli, B., Sundaramoorthy, P., Murugadoss, R., & Keerthana, N. V. "Advancing Cervical Cancer Identification using Generative-based Adversarial Networks: An Integrative Learning Methodology." 2025 6th International Conference for Emerging Technology (INCET). IEEE, (2025).
51. Routhu, K. K. "Beyond Automation: AI-Powered Employee Engagement Journeys in Oracle HCM Cloud." KOS Journal of AIML, Data Science, and Robotics 1.1 (2024): 1-6.
52. Routhu, K. K. "The future of HCM: Evaluating Oracle's and SAP's AI-powered solutions for workforce strategy." Journal of Artificial Intelligence, Machine Learning & Data Science 2.2 (2024): 2942-2947.
53. Sannapureddy, R., Nadella, V. M., & Nelavelli, S. "Edge–Cloud Continuums for Latency-Sensitive Tasks." International Journal of AI, BigData, Computational and Management Studies 5.4 (2024): 189-201
54. Arigela, A. K., Brahmareddy, A., Sreenivas, T. S., Selvan, M. P., Venu, N., & Lal, D. K. "Optimizing Energy Efficiency and Latency in IoT Devices Through AI-Based Adaptive Protocols in Fog-Edge Computing Environments." Congress on Smart Computing Technologies. Singapore: Springer Nature Singapore, (2024).
55. Nadella, V. M. "Al-Native 6G Network Management." American International Journal of Computer Science and Technology 6.1 (2024): 23-37.

Source of support: Nil; **Conflict of interest:** Nil.

Cite this article as:

Nagumotu, V. T., Kavuluri, H. V. R., Pathani, A. K., Dasari, A., Chilakapati, V. K. & Keshireddy, S. R. "A Review of Big Data Analytics: Tools, Techniques, and Scalability Challenges in Data Science" *Sarcouncil Journal of Engineering and Computer Sciences* 4.7 (2025): pp 1499-1511.

Publisher: SARC Publisher