PERFECT
JOURNAL OF SMART ALGORITHMS

E-ISSN: E-ISSN: 3064-0377
Volume 3, **Number** 1, January 2026
https://doi.org/10.62671/perfect.v3i1.217

A Comprehensive Review on Data Science Frameworks for Big Data Analytics

Hassan Raza[1], Tsendayush Erdenetsogt[2], A Singh[3], Mazhar Farooq[4], Muhammad Mohsin Kabeer[5*], Muhammad Shahrukh Aslam[6]
[1] Washington university of science and technology, USA
[2] University of the Potomac, USA
[3] University of North America (UoNA), USA
[4] Southern New Hampshire University
[5] Gannon University
[6] Concordia University, USA
[1]hr968182@gmail.com, [2]Tsendayush.Erdenetsogt@student.potomac.edu, [3]ankursingh.30dec@gmail.com ,
[4]Mazhar.farooq@snhu.edu, , [5]Mohsinkabeer86@gmail.com,[6]shahrukhaslam81991@gmail.com

open access

***Corresponding Author**

Article History:
Submitted: 09-12-2025
Accepted: 29-12-2025
Published: 06-01-2026

Keywords:
Big Data; Data Science Frameworks; Hadoop; Spark; Real-Time Analytics

ABSTRACT

The importance of big data analytics is now essential in deriving insights in large and complex information in various industries. This review discusses major data science frameworks, such as Apache Hadoop, Spark, Flink, and Storm, their architecture, capabilities, and a relative advantage of processing batches and in real-time. It also presents major challenges that can affect the framework efficiency, including scalability, latency, and heterogeneity of data, security, and the complexity of operational, among others. Lastly, the new trends such as the adoption of AI, cloud-native architecture, real-time streaming, and intelligent automation are discussed to demonstrate the changing environment. This review gives an in-depth insight into the concept of big data frameworks and how they facilitate the achievement of effective analytics.

INTRODUCTION

The fast development of digital technologies, interconnectivity, and web-based systems has resulted in an unprecedented increase in the volume of the data produced every day. This massive growth, which is commonly known as the big data era, has changed the manner in which organizations acquire, control and make use of information (Cao, 2017). Within this context, data science has become a critical multidisciplinary area that integrates statistical procedures, computational approaches, expert knowledge, and smart algorithms to learn significant information about large and complicated datasets. The more data is accumulated, the more its volume and complexity, the more efficient, scalable and well-structured frameworks are needed (Sakr & Elgammal, 2016).

Data science models are important in facilitating researchers, analysts, and organizations to operate with big data. These models offer standard tools, libraries and workflow models that direct the overall analysis process including data ingestion and preprocessing to modeling, evaluation and deployment. Without these, the problems that come with big data, namely the scale and complexity of dealing with distributed data, the large scale computations, and the real time analytics would be much harder to handle (Ahmed et al., 2023). The number of available tools has also wildly increased, and it now provides fine-grained solutio ns that are machine learning focused, deep learning focused, and stream processing focused, as well as cloud-based analytics solutions and automated workflow solutions (Galetsi et al., 2019).

Due to the dynamic and rapid technological development, it might be difficult to select the appropriate framework. All frameworks are different in terms of architecture, scalability, performance features, support of programming and appropriateness in certain applications. That is why it is required to have a thorough examination that would assist the readers in perceiving the landscape of data science frameworks applied to big data analytics (Szymańska, 2018). This kind of review not only reveals the opportunities and weaknesses of popular tools but also gives an idea of how these structures are compatible with new technologies, such as artificial intelligence, cloud computing, and distributed systems (Abuqabita et al., 2019).

The aim of this review is to present a systematic analysis of the largest data science architecture serving big data analytics. It examines their fundamental functionalities, compares how they perform regarding industry use cases and

PERFECT
JOURNAL OF SMART ALGORITHMS

E-ISSN: E-ISSN: 3064-0377
Volume 3, **Number** 1, January 2026
https://doi.org/10.62671/perfect.v3i1.217

how they mitigate some of the major challenges associated with processing large-scale datasets. Moreover, the review establishes the current trends that can determine the future of big data analytics, including automation, real-time decision-making, and the growing adoption of cloud-native infrastructures. All in all, this introduction preconditions the comprehension of the reasons why data science frameworks have become essential in the modern data-driven environment and how they are still developing to address the requirements of modern analysis (Abuqabita et al., 2019).

BASICS OF BIG DATA ANALYTICS

Big data analytics can be defined as the systematic study of large, heterogeneous, and ever-increasing data volumes with a view to revealing any latent patterns, correlations, and insights that might be employed to enhance improved decision-making. In contrast to the conventional datasets that can be processed with the help of the traditional tools, big data demands new technologies and models of analysis that can manage the peculiarities of such data (Acharjya & Ahmed, 2016). These attributes can be widely described by the famous so-called V model that consists of Volume, Variety, Velocity, Veracity and Value. Volume refers to the vastness of data generated with the help of sensors, social media, transactions, and IoT devices. Diversity brings to the fore the existence of structured, semi-structured and unstructured information. Velocity is concerned with the pace of the creation of new data and its required processing. Veracity is concerned with the data quality and uncertainty whereas Value is concerned with the potential insights that can be made out (Ahn et al., 2022).

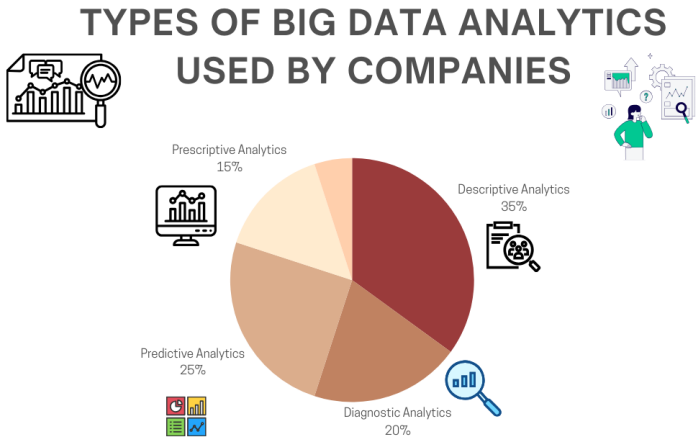

Figure 1. Types of big data analytics used by companies

The process of big data analytics has a lifecycle in general consisting of several stages. It starts with the data collection stage that will receive information through the high number of sources databases, streaming platforms, logs, and web services. This is then succeeded by data storage, which can be based on scalable distributed storage systems such as HDFS or a cloud-based storage system, which is designed to be able to process large quantities of data reliably (Arowoogun et al., 2024). Data processing is the next step, and the big data systems of Apache Spark, Hadoop MapReduce, or Flink can be applied to process and prepare data to analyze it. Once processed, data analysis occurs with the help of statistical techniques, machine learning models, or a deep learning technique in order to derive meaningful information. Lastly, data visualization and reporting can be used to summarize the findings in a way that can be interpreted and used to make decisions (Akil et al., 2017).

Notwithstanding its potential, big data analytics is associated with a lot of challenges. The processing of large volumes of data requires a high level of processing power and streamlined algorithms. The combination of heterogeneous data of various types can be complicated. Another significant issue is the security and privacy of the data, particularly when it is associated with sensitive or personal information. Moreover, companies are frequently faced with the lack of competent specialists who would be capable of working with big data tools and technologies (Al-Omoush et al., 2024).

It is important to grasp the basics of big data analytics since it forms the basis on which the contemporary data science models are rooted. These frameworks are specially made to tackle the problem of scale, speed, complexity and reliability- eventually letting organizations convert raw and unorganized data to useful information. With industries generating more data than ever before, the importance of mastering these fundamentals continues to increase in order to be useful in analytics and make effective decisions (Al-Sai et al., 2022).

PERFECT
JOURNAL OF SMART ALGORITHMS

E-ISSN: E-ISSN: 3064-0377
Volume 3, **Number** 1, January 2026
https://doi.org/10.62671/perfect.v3i1.217

CLASSIFICATION OF DATA SCIENCE FRAMEWORKS

Data science frameworks are very crucial tools which offer organized surroundings in which big data is processed, analyzed, and insights are retrieved. They are aimed at making complex tasks easier, enhance reproducibility, and optimization. Frameworks could be divided into various categories in accordance with their use cases, functionality and architecture (Al-Salim et al., 2018). The knowledge of such categories assists organizations and researchers in choosing the most suitable framework regarding their particular analytical requirements. The main categories of them will be programming-based frameworks, distributed computing frameworks, machine learning and deep learning frameworks, cloud based frameworks, and workflow or pipeline management tools (Ali & Hariprasad, 2023).

PROGRAMMING-BASED FRAMEWORKS

Based on programming, frameworks are aimed at delivering libraries, APIs, and tools that support data manipulation, statistical analysis, and visualisation. They are normally language specific, and the most popular ones are Python libraries such as Pandas, NumPy, and Tidyverse in R. These frameworks enable data scientists to do exploratory data analysis, preprocessing and simple modeling. Although they are not specifically aimed at large scale distributed computing, their simplicity and flexibility are ideal in small to medium size data sets and prototyping models before scaling to large data sets (Alosert et al., 2022).

FRAMEWORKS: DISTRIBUTED COMPUTING FRAMEWORKS

The distributed computing systems are also on-demand systems that handle large amounts of data with a large number of machines, as it is highly performative and resilient to failures. Apache Hadoop and Apache Spark are the best examples. Hadoop is based on the paradigm of MapReduce and HDFS to process in the batch, whereas Spark offers the possibilities of in-memory computing and supports both batch and real-time analytics (Altuwairiqi, 2023). These frameworks can help in scaling up therefore giving the organization the ability to process petabytes of data with efficiency. They are especially handy when intensive computing is needed like in the analysis of logs, ETL (extract, transform, load) processes and massive date conversions (Alwadi et al., 2023).

MACHINE LEARNING AND DEEP LEARNING STRUCTURES

Machine learning and deep learning systems include existing algorithms, neural network models and training tools of predictive modeling and pattern recognition. Good examples are TensorFlow, PyTorch and Scikit-learn. These frameworks hasten the construction of models through provided optimized implementations of classification, regression, clustering and deep learning architectures including convolutional or recurrent neural networks (Amalina et al., 2019). Their automation, GPUs acceleration and repeatability are invaluable to AI-based applications such as image recognition, natural language processing and recommendation engines (Ben Atitallah et al., 2020).

CLOUD-BASED FRAMEWORKS

Cloud-based frameworks make use of cloud infrastructure to offer scalable, flexible and cost-effective data processing services. On-premise hardware does not require users to invest enormous sums of money in deploying, training, and running models, as platforms such as Google Cloud AI, AWS SageMaker, and Microsoft Azure Machine Learning enable users to do so (Ayvaz & Alpay, 2021). The following functions are provided by these frameworks: automated resource provisioning, integration with other cloud services and serverless computing. Cloud-based architectures prove to be most advantageous when companies need to have elastic computing, team work, and straightforward implementation of high-scope analytics software (Backhoff & Ntoutsi, 2016).

PROCESS AND WORKFLOW MANAGEMENT TOOLS

Apache Airflow, Luigi, and Kubeflow are workflow and pipeline management frameworks, aimed at automating workflows of intricate machine learning pipelines and data processing pipelines. They guarantee reproducibility, schedule, track data workflow execution and deal with the dependencies between data streams. They are critical in the process of administering the end-to-end data science procedures effectively, minimize human error, and simplify the implementation in production settings (Bansal et al., 2020). The data science frameworks are sorted into categories that indicate their specialized functions in data processing, analytics, machine learning, and cloud computing as well as workflow management. The choice of the suitable framework is based on the amount of data, calculation needs, the skills of the teams, and the purpose of the project, which helps organizations to utilize the potential of big data and sophisticated analytics to their utmost (Ben Hamida et al., 2021).

PERFECT
JOURNAL OF SMART ALGORITHMS

E-ISSN: E-ISSN: 3064-0377
Volume 3, **Number** 1, January 2026
https://doi.org/10.62671/perfect.v3i1.217

DATA SCIENCE FRAMEWORKS ROLE

Data science frameworks are central to the contemporary analytics since they offer structured environments, tools and processes that facilitate the process of extracting insights on complex and large datasets. In the modern data-driven world, organizations are creating huge amounts of structured, semi-structured and unstructured data in the form of social media, IoT devices, transactional systems and sensors. In the absence of the strong frameworks, it would be almost impossible to effectively manage, process and analyze this information (Bhatia & Kumar, 2018). The frameworks serve to keep the raw data and actionable insights in close touch, therefore, data scientists can concentrate on modeling, analysis, and decision-making as opposed to infrastructure and low-level programming issues (Brendel et al., 2022).

Scalability is one of the main positive aspects of the data science frameworks. Such frameworks as Apache Hadoop and Apache Spark enable companies to use petabytes of information on distributed clusters. Such horizontal scaling ability has made it possible to scale data analytics in line with the growing data volumes without affecting its performance. Since datasets are growing, frameworks automatically allocate resources, distribute tasks and withstand faults, it is simpler to ensure that consistent results are achieved when working with large-scale workloads of data. Another important role of these frameworks is efficiency (Briard et al., 2023). The in-memory computing systems, including Spark, significantly decrease the time of data processing in contrast to the more traditional disk-based systems. Data manipulation, statistical analysis and machine learning libraries are optimized, which minimizes computation to provide faster insights. Frameworks are also connected with databases, cloud computing, and real-time streams, which will require less time to prepare and integrate data (Calude & Longo, 2017).

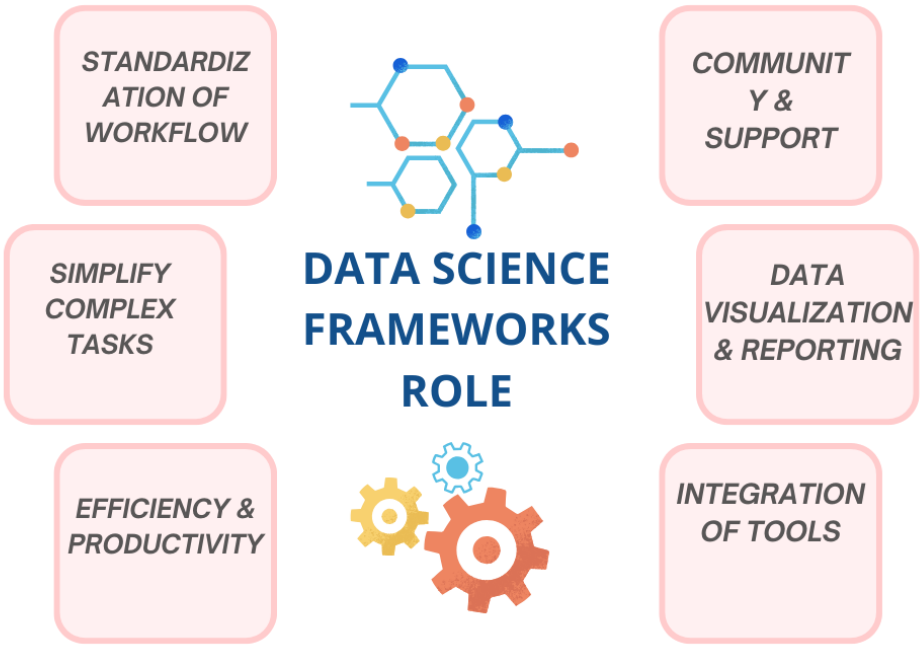

Figure 2. Data science frameworks role

One of the major features offered by data science frameworks is automation. Apache Airflow and Kubeflow are workflow management systems that enable the scheduling, tracking, and coordination of pipelines of complex tasks without human intervention. Automated preprocessing, model training, hyperparameter optimization, and deployment pipelines minimise human error, enhance reproducibility and speed up the development-production pipeline (Carbone et al., 2015). Frameworks guarantee re-reproducibility of the data analytics and machine learning experiments. Reliable inter-rater reliability Standardized APIs, code and dataset version control and uniform execution environments aid in the replication of results, which is a critical aspect of scientific studies, regulatory guidelines, and group initiatives. Reproducibility also enables organizations to test models, contrast methods and keep the quality check over various teams (Chen & Zhang, 2014).

The foundations of the current analytics are built on the principles of data science frameworks which allow to be scaled, enhance efficiency, automation, and reproducibility. They simplify the process of working with big data and leave data scientists to concentrate on deriving meaningful insights. These structures are essential in converting raw data into practical knowledge by means of a well-organized and streamlined setting to aid in informed decision-making and innovation in a variety of industries (Chen et al., 2023).

PERFECT
JOURNAL OF SMART ALGORITHMS

E-ISSN: E-ISSN: 3064-0377
Volume 3, **Number** 1, January 2026
https://doi.org/10.62671/perfect.v3i1.217

COMPARISON OF HIGH-POPULAR FRAMEWORKS AND BIG DATA ANALYTICS

The past few years have seen the rate of growth of big data analytics rise and as a result a variety of frameworks have been developed to handle, process and analyze large amounts of data in an efficient manner. Some of the most popular frameworks are Apache Hadoop, Apache Spark, Apache Flink, and Apache Storm among others. Both the frameworks have use cases and architectural benefits, and it is important to compare them in terms of performance, scalability, fault tolerance, and integration simplicity (Chopra et al., 2022).

COMPARISON OF HIGH-POPULAR FRAMEWORKS AND BIG DATA ANALYTICS

Feature	High-Popular Frameworks	Big Data Analytics Frameworks
Purpose	Build applications (web, mobile, backend)	Process and analyze massive datasets
Data Handling	Small to medium data, handled in memory or standard databases	Huge volumes of structured/unstructured data (TBs to PBs)
Scalability	Limited; mostly vertical or simple horizontal scaling	Highly scalable horizontally across clusters
Performance	Optimized for app responsiveness and UI/UX	Optimized for distributed processing and batch/stream analytics
Examples	React, Angular, Vue, Django, Flask, Spring Boot	Hadoop, Apache Spark, Apache Flink, Apache Kafka, Presto
Use Case	Web apps, mobile apps, APIs, microservices	Big data analytics, real-time streaming, machine learning, ETL
Complexity	Moderate; focuses on app development	High; requires knowledge of distributed computing
Data Storage	Uses relational or NoSQL databases	Uses distributed file systems (HDFS) or big data stores (Cassandra, HBase)

Figure 3. Comparison of high popular frameworks and big data analytics

One of the first and most popular systems of big data processing is apache Hadoop. It is based on the Hadoop Distributed File System (HDFS) to store the massive data on many nodes and apply the MapReduce programming model to complete batch processing. Hadoop is structured to be very scalable thus capable of processing petabytes of structured and unstructured data. It has a fault tolerance system where failure by a node would automatically be reallocated into tasks by the system (Dean & Ghemawat, 2008). Though, it is also evident that Hadoop has been criticized because of its reduced processing speed mostly in processing real-time data analytics as a result of the overhead of writing intermediate data to disk between map and reduce jobs (Deepa et al., 2022).

Spark is a Apache-built system designed to overcome the latency of Hadoop and is also in-memory computing and thus does a great job in speeding up the processing speed. Spark is appropriate in processing complex data-sets and machine learning as well as graph computing since it supports both analytic batch and real-time. It is also compatible with Hadoop HDFS where it adds to it and better performance is achieved despite being backward compatible. The popularity of Spark among data scientists is also boosted by the fact that it is easy to use due to its high-level APIs in Java, Scala, Python (Dhifli et al., 2017).

Another framework used in real-time processing of the data streams is Apache Flink. In contrast to batch-oriented solutions, Flink is the most effective when the streams of data are continuous and have low latency. It offers high consistency guarantees and event-time processing, therefore, needed to guarantee correct real-time analytics. Flink also supports stateful calculations by default, which is why it is a resilient option in the case that needs to monitor continuously and make decisions dynamically. Apache Storm is directly specialized in event-driven applications in real-

PERFECT
JOURNAL OF SMART ALGORITHMS

E-ISSN: E-ISSN: 3064-0377
Volume 3, **Number** 1, January 2026
https://doi.org/10.62671/perfect.v3i1.217

time (Dicuonzo et al., 2019). It can handle unlimited data streams with very little latency which makes it suitable in applications such as fraud detection, live monitoring and online recommendation systems. On the one hand, it has a high level of reliability, but on the other hand, Storm is more complicated to install and maintain than Spark and Flink (Diouf et al., 2018).

The application needs are important in determining the type of big data framework to be used. Hadoop is a trusted option where processing large volumes of data in batch mode is required, Spark is also adaptable when it comes to mixed workloads, Flink is the best option when a consistent stateful stream processing is required, and Storm is suitable when real-time event processing is needed. It is also important to consider the factors like speed, fault tolerance, scalability, and integration capabilities and therefore choose the best framework to use in big data analytics (Domann et al., 2016).

BIG DATA FRAMEWORK CHALLENGES AND LIMITATIONS

Although the frameworks of big data are rapidly evolving, some of them still face threats and limitations that affect their efficacy, scale, and adoption in general. With the growing use of big data models by organizations to make informed decisions that are based on data, it is important to learn about these limitations to make effective decisions when developing a strong big data solution (Dundar et al., 2007).

Scalability: Although the current systems such as Hadoop, Spark, and Flink are developed to handle thousands of nodes, resource management turns out to be a challenge. Massive deployments necessitate the sophisticated configuration and optimization to prevent data storage, network bandwidth, and compute bottlenecks (Elser & Montresor, 2013). Lack of proper resource allocation may cause imbalance in the distribution of loads thus causing underutilization of hardware and underutilization process. Additionally, the dynamism in cloud environments may introduce cost implications because it may be costly to sustain high availability and performance of huge databases (Emmanuel & Stanier, 2016).

Information Diversity and Unification: There are several types of big data, including structured, semi-structured, and unstructured, and they may be provided by the social media, IoT devices, and enterprise databases. The difficulty of integrating heterogeneous data into one framework is due to the fact that most frameworks were originally optimized to process either a batch or a stream. The efficient processing of unstructured data, including images, videos, and text, may mandate further processing, which may elevate the complexity of processing and the latency (Tandon et al., 2020).

Fault Tolerance: Such frameworks as Hadoop and Spark are fault tolerant but are not resistant to failures, particularly when operating in large-scale distributed configuration. Data may be lost or re-computed due to failures of nodes, network interruptions or bugs in the software programs which impact on the overall performance. The exact-once processing semantics in the context of the failure conditions are considered especially complicated in real-time streaming systems such as Storm and Flink and may restrict the reliability in the applications with high stakes (Imran et al., 2021).

Latency Problems and Performance Problems: Although Spark and Flink provide in-memory processing to enhance faster processing, latency in the event of huge data volumes or high-throughput data streams can still be an issue. Hadoop is disk-based and is problematic with real-time analytics. Furthermore, overhead of coordination of tasks, movement of data between nodes and distributed storage may affect throughput responsiveness (Khanra et al., 2020).

Difficulty and Adequacy of Skills: Big data deployment and maintenance entails expert expertise in distributed computing, cluster management, and programming. Organizations may also struggle to hire the staff with Spark, Hadoop, or Flink expertise, which may restrain the adoption and raise the cost of operation (Mohamed et al., 2020).

Security and Privacy Issue: Dealing with sensitive data in distributed nodes creates vulnerability in security. Guaranteeing the encryption and control of data as well as adherence to regulations introduce extra complexity that is not inherently supported by most structures. The potential offered by big data structures is enormous in analytics, yet they are characterized by scalability problems, data integration issue, reliability, latency, operational complexity, and security problems (Shahnawaz & Kumar, 2025). All these shortcomings should be considered to maximize performance and deliver cost effective and reliable big data solutions.

TRENDS IN THE FUTURE OF DATA SCIENCE ARCHITECTURE IN BIG DATA ANALYTICS

The sphere of big data analytics is changing at great pace, which is conditioned by the growth of technologies, accumulation of data, and the necessity to receive real-time information. As a result of this, the data science structures of big data will also experience dramatic changes in the foreseeable future. The trends are critical to comprehend by the organizations and researchers intending to implement the latest solutions (Olaniyi et al., 2023).

The combination with Artificial Intelligence and Machine Learning: In the future, big data frameworks will have more and more in-built support of machine learning and artificial intelligence (AI) algorithms. Although existing platforms such as Apache Spark already provide ML libraries, the direction of the trend is towards greater integration, allowing the frameworks to automatically plan data processing pipelines to predictive analytics, anomaly detection and

PERFECT
JOURNAL OF SMART ALGORITHMS

E-ISSN: E-ISSN: 3064-0377
Volume 3, **Number** 1, January 2026
https://doi.org/10.62671/perfect.v3i1.217

recommendation systems. This will minimize the use of independent platforms and shorten the process of data reception and usable information (Rane et al., 2024).

Real-time and streaming analytics: Real-time analytics will be in demand, especially in such fields as financial services, e-commerce, and IoT applications. Apache Flink and Spark streaming frameworks will probably keep developing, with ultra-low latency processing, improved state management and support of event-driven architecture. This will also allow organizations to make real time decisions using streaming data, enhancing their responsiveness and operational efficiency (Ochuba et al., 2024).

Serverless Architectures: The big data architectures are shifting towards the cloud-native architecture, which utilizes elasticity, scalability, and managed cloud services. There is an increasing popularity of serverless computing, in which the infrastructure is abstracted. The trend makes operations simpler, enables workload-based dynamic scaling, and enables organizations to manage the costs better when handling huge amounts of data (Thayyib et al., 2023).

Increased Security and Privacy Accounts: As data privacy laws get more stringent across the globe, the frameworks of tomorrow will incorporate superior security features, including end-to-end encryption, differential privacy and auto compliance. This will help organizations to work with sensitive data without affecting performance (Pedro, 2023).

Automation and Smart Orchestration: Automation will also be an important part of new frameworks, such as smart task scheduling, resource scheduling, and error repairing. With the use of AI-based orchestration, frameworks will waste less energy by using big data analytics, and fewer human resources will be required of organizations with less technical resources. The future of data science models of big data analytics is in intelligent, faster, and more secure systems. The trends that will transform the way organizations gain value on large-scale data will include the integration of AI, real-time processing, cloud-native architectures, improved security and intelligent automation like never before (Nazir et al., 2020).

CONCLUSION

The use of big data analytics has established itself as an inseparable part of contemporary organizations, which allows making decisions based on data, predictive modeling, and taking action in different sectors. The increasing rate of data volumes, speed and diversity has made the creation of advanced data science systems that can handle and process large volumes of data effectively. The review has examined the development, functionality, comparisons, challenges and future trends of the frameworks and how vital they are to the big data ecosystem.

Apache Hadoop, Apache Spark, Apache Flink, and Apache Storm are data science frameworks that have largely changed the way data is processed and analyzed. Hadoop created the base with its batch processing that was scaled and fault-tolerant and Spark enhanced it with the use of in-memory computing and support of both batch and real-time analytics. Flink and Storm brought a new layer of real-time stream processing to allow organizations to react fast to real-time streams of data. Both frameworks have their own advantages and the choice of the right solution is determined by variables like data type, processing needs, latency tolerance, and scalability needs. These differences are important to organizations that are keen on maximizing their big data strategies.

In spite of the benefits, big data frameworks have significant threats and constraints. Lack of scalability, complexity of resource management, heterogeneity of data, latency issues and specially required skills can impede the efficient implementation of such systems. Issues of security and privacy are also not to be overlooked, as sensitive information may be spread over several nodes and regions. To overcome these constraints, there is need to continue research, develop more sophisticated engineering solutions and implement emerging technologies strategically.

In the future, the future of data science frameworks is highly connected with the development of artificial intelligence, machine learning, real-time analytics, cloud-native architecture, and intelligent automation. The combination of AI-driven analytics, serverless computing, more robust security measures, and automated orchestration will ensure that big data structures become efficient and more accessible and can manage complex workloads with a minimal amount of human intervention. Such trends are believed to allow organizations to derive more insights, enhance operational efficiency and have a competitive advantage in a world that is becoming more and more data-driven.

Big data analytics data science frameworks are inseparable resources of the contemporary enterprises. Though it is not free of challenges, their performance, scalability and usability are improving with the continued technological development and innovative design solutions. With an adequate choice and optimization of these frameworks, the organizations will be able to maximize the potentials of their data and make informed decisions and generate innovation in a variety of domains. Further development of these frameworks will be a major factor in the future of big data analytics.

PERFECT
JOURNAL OF SMART ALGORITHMS

E-ISSN: E-ISSN: 3064-0377
Volume 3, **Number** 1, January 2026
https://doi.org/10.62671/perfect.v3i1.217

REFERENCES

Abuqabita, F., Al-Omoush, R., & Alwidian, J. (2019). A comparative study on big data analytics frameworks, data resources and challenges. *Modern Applied Science, 13*(7), 1–14.

Abuqabita, F., Al-Omoush, R., & Alwidian, J. (2019). A comparative study on big data analytics frameworks, data resources and challenges. *Modern Applied Science, 13*(7), 1–14.

Acharjya, D. P., & Ahmed, K. (2016). A survey on big data analytics: Challenges, open research issues and tools. *International Journal of Advanced Computer Science and Applications, 7*(2), 511–518.

Ahmed, A., Xi, R., Hou, M., Shah, S. A., & Hameed, S. (2023). Harnessing big data analytics for healthcare: A comprehensive review of frameworks, implications, applications, and impacts. *IEEE Access, 11*, 112891–112928.

Ahn, J. S., Jhung, K., Oh, J., Heo, J., Kim, J.-J., & Park, J. Y. (2022). Association of resting-state theta–gamma coupling with selective visual attention in children with tic disorders. *Frontiers in Human Neuroscience, 16*, 1017703.

Akil, B., Zhou, Y., & Röhm, U. (2017). On the usability of Hadoop MapReduce, Apache Spark and Apache Flink for data science. In *Proceedings of the IEEE International Conference on Big Data* (pp. 303–310). IEEE.

Ali, I. M. S., & Hariprasad, D. (2023). Hyper-heuristic salp swarm optimization of multi-kernel support vector machines for big data classification. *International Journal of Information Technology, 15*(2), 651–663.

Al-Omoush, K. S., Garcia-Monleon, F., & Mas Iglesias, J. M. (2024). Exploring the interaction between big data analytics, frugal innovation, and competitive agility: The mediating role of organizational learning. *Technological Forecasting and Social Change, 200*, 123188.

Alosert, H., Savery, J., Rheaume, J., Cheeks, M., Turner, R., Spencer, C., Farid, S. S., & Goldrick, S. (2022). Data integrity within the biopharmaceutical sector in the era of Industry 4.0. *Biotechnology Journal, 17*(6), 2100609.

Al-Sai, Z. A., Husin, M. H., Syed-Mohamad, S. M., Abdin, R. M. S., Damer, N., Abualigah, L., & Gandomi, A. H. (2022). Explore big data analytics applications and opportunities: A review. *Big Data and Cognitive Computing, 6*(4), 157.

Al-Salim, A. M., El-Gorashi, T. E. H., Lawey, A. Q., & Elmirghani, J. M. H. (2018). Greening big data networks: Velocity impact. *IET Optoelectronics, 12*(3), 126–135.

Altuwairiqi, M. (2023). Combining extreme learning machine through random projections for dimensional information taxonomy and assembling. In *Proceedings of the IEEE International Conference on Innovations in High Speed Communication and Signal Processing* (pp. 488–491). IEEE.

Alwadi, M., Chetty, G., & Yamin, M. (2023). A framework for vehicle quality evaluation based on interpretable machine learning. *International Journal of Information Technology, 15*(1), 129–136.

Amalina, F., Hashem, I. A. T., Azizul, Z. H., Fong, A. T., Firdaus, A., Imran, M., & Anuar, N. B. (2019). Blending big data analytics: Review on challenges and a recent study. *IEEE Access, 8*, 3629–3645.

Arowoogun, J. O., Babawarun, O., Chidi, R., Adeniyi, A. O., & Okolo, C. A. (2024). A comprehensive review of data analytics in healthcare management: Leveraging big data for decision-making. *World Journal of Advanced Research and Reviews, 21*(2), 1810–1821.

Ayvaz, S., & Alpay, K. (2021). Predictive maintenance system for production lines in manufacturing: A machine learning approach using IoT data in real time. *Expert Systems with Applications, 173*, 114598.

Backhoff, O., & Ntoutsi, E. (2016). Scalable online-offline stream clustering in Apache Spark. In *Proceedings of the IEEE International Conference on Data Mining Workshops* (pp. 37–44). IEEE.

Bansal, M., Chana, I., & Clarke, S. (2020). A survey on IoT big data: Current status, 13 V's challenges, and future directions. *ACM Computing Surveys, 53*(6), 1–59.

Ben Atitallah, S., Driss, M., Boulila, W., & Ben Ghézala, H. (2020). Leveraging deep learning and IoT big data analytics to support smart cities development: Review and future directions. *Computer Science Review, 38*, 100303.

Ben Hamida, S., Benjelloun, G., & Hmida, H. (2021). Trends of evolutionary machine learning to address big data mining. In *Proceedings of the International Conference on Information and Knowledge Systems* (pp. 85–99). Springer.

Bhatia, S., & Kumar, R. (2018). Review of graph processing frameworks. In *Proceedings of the IEEE International Conference on Data Mining Workshops* (pp. 998–1005). IEEE.

Brendel, M., Su, C., Bai, Z., Zhang, H., Elemento, O., & Wang, F. (2022). Application of deep learning on single-cell RNA sequencing data analysis: A review. *Genomics, Proteomics & Bioinformatics, 20*(5), 814–835.

Briard, T., Jean, C., Aoussat, A., & Véron, P. (2023). Challenges for data-driven design in early physical product design: A scientific and industrial perspective. *Computers in Industry, 145*, 103814.

PERFECT
JOURNAL OF SMART ALGORITHMS

E-ISSN: E-ISSN: 3064-0377
Volume 3, **Number** 1, January 2026
https://doi.org/10.62671/perfect.v3i1.217

Calude, C. S., & Longo, G. (2017). The deluge of spurious correlations in big data. *Foundations of Science, 22*(3), 595–612.

Cao, L. (2017). Data science: A comprehensive overview. *ACM Computing Surveys (CSUR), 50*(3), 1–42.

Carbone, P., Katsifodimos, A., Ewen, S., Markl, V., Haridi, S., & Tzoumas, K. (2015). Apache Flink: Stream and batch processing in a single engine. *IEEE Data Engineering Bulletin, 38*(4), 28–38.

Chen, C. L. P., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences, 275*, 314–347.

Chen, Y., Hong, Z., & Yang, X. (2023). Cost-sensitive online adaptive kernel learning for large-scale imbalanced classification. *IEEE Transactions on Knowledge and Data Engineering, 35*(10), 10554–10568.

Chopra, M., Singh, S. K., Gupta, A., Aggarwal, K., Gupta, B. B., & Colace, F. (2022). Analysis and prognosis of sustainable development goals using big data-based approach during COVID-19 pandemic. *Sustainable Technology and Entrepreneurship, 1*(2), 100012.

Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM, 51*(1), 107–113.

Deepa, N., Pham, Q.-V., Nguyen, D. C., Bhattacharya, S., Gadekallu, T. R., Maddikunta, P. K. R., Fang, F., & Pathirana, P. N. (2022). A survey on blockchain for big data: Approaches, opportunities, and future directions. *Future Generation Computer Systems, 131*, 209–226.

Dhifli, W., Aridhi, S., & Mephu Nguifo, E. (2017). MR-SimLab: Scalable subgraph selection with label similarity for big data. *Information Systems, 69*, 155–163.

Dicuonzo, G., Galeone, G., Zappimbulso, E., & Dell'Atti, V. (2019). Risk management 4.0: The role of big data analytics in the bank sector. *International Journal of Economics and Financial Issues, 9*(6), 40–47.

Diouf, P. S., Boly, A., & Ndiaye, S. (2018). Variety of data in the ETL processes in the cloud: State of the art. In *Proceedings of the IEEE International Conference on Innovative Research and Development* (pp. 1–5). IEEE.

Domann, J., Meiners, J., Helmers, L., & Lommatzsch, A. (2016). Real-time news recommendations using Apache Spark. In *Proceedings of CLEF* (pp. 628–641).

Dundar, M., Krishnapuram, B., Bi, J., & Rao, R. B. (2007). Learning classifiers when the training data is not IID. In *Proceedings of the International Joint Conference on Artificial Intelligence* (pp. 756–761).

Elser, B., & Montresor, A. (2013). An evaluation study of big data frameworks for graph processing. In *Proceedings of the IEEE International Conference on Big Data* (pp. 60–67). IEEE.

Emmanuel, I., & Stanier, C. (2016). Defining big data. In *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies* (pp. 1–6).

Galetsi, P., Katsaliaki, K., & Kumar, S. (2019). Values, challenges and future directions of big data analytics in healthcare: A systematic review. *Social Science & Medicine, 241*, 112533.

Imran, S., Mahmood, T., Morshed, A., & Sellis, T. (2021). Big data analytics in healthcare—A systematic literature review and roadmap for practical implementation. *IEEE/CAA Journal of Automatica Sinica, 8*(1), 1–22.

Khanra, S., Dhir, A., Islam, A. K. M. N., & Mäntymäki, M. (2020). Big data analytics in healthcare: A systematic literature review. *Enterprise Information Systems, 14*(7), 878–912.

Mohamed, A., Najafabadi, M. K., Wah, Y. B., Zaman, E. A., & Maskat, R. (2020). The state of the art and taxonomy of big data analytics: View from new big data framework. *Artificial Intelligence Review, 53*(2), 989–1037.

Nazir, S., Khan, S., Khan, H. U., Ali, S., García-Magariño, I., Atan, R. B., & Nawaz, M. (2020). A comprehensive analysis of healthcare big data management, analytics and scientific programming. *IEEE Access, 8*, 95714–95733.

Ochuba, N. A., Amoo, O. O., Okafor, E. S., Akinrinola, O., & Usman, F. O. (2024). Strategies for leveraging big data and analytics for business development: A comprehensive review across sectors. *Computer Science & IT Research Journal, 5*(3), 562–575.

Olaniyi, O. O., Okunleye, O. J., & Olabanji, S. O. (2023). Advancing data-driven decision-making in smart cities through big data analytics: A comprehensive review of existing literature. *Current Journal of Applied Science and Technology, 42*(25), 10–18.

Pedro, F. (2023). A review of data mining, big data analytics, and machine learning approaches. *Journal of Computational and Natural Sciences, 3*, 169–181.

Rane, N. L., Paramesha, M., Choudhary, S. P., & Rane, J. (2024). Machine learning and deep learning for big data analytics: A review of methods and applications. *Partners Universal International Innovation Journal, 2*(3), 172–197.

Sakr, S., & Elgammal, A. (2016). Towards a comprehensive data analytics framework for smart healthcare services. *Big Data Research, 4*, 44–58.

Shahnawaz, M., & Kumar, M. (2025). A comprehensive survey on big data analytics: Characteristics, tools and techniques. *ACM Computing Surveys, 57*(8), 1–33.

PERFECT
JOURNAL OF SMART ALGORITHMS

E-ISSN: E-ISSN: 3064-0377
Volume 3, **Number** 1, January 2026
https://doi.org/10.62671/perfect.v3i1.217

Szymańska, E. (2018). Modern data science for analytical chemical data: A comprehensive review. *Analytica Chimica Acta, 1028*, 1–10.

Tandon, A., Dhir, A., Islam, A. K. M. N., & Mäntymäki, M. (2020). Blockchain in healthcare: A systematic literature review, synthesizing framework and future research agenda. *Computers in Industry, 122*, 103290.

Thayyib, P. V., Mamilla, R., Khan, M., Fatima, H., Asim, M., Anwar, I., Shamsudheen, M. K., & Khan, M. A. (2023). State-of-the-art of artificial intelligence and big data analytics reviews in five different domains: A bibliometric summary. *Sustainability, 15*(5), 4026.