International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V72I12P113

Volume 72 Issue 12, 108-120, December 2024
© 2024 Seventh Sense Research Group®

Original Article

Big Data Engineering on Cloud Platforms

Shrikaa Jadiga

Independent Researcher, USA.

Coresponding Author : shrikaajadiga@Gmail.com

Received: 31 October 2024 Revised: 28 November 2024 Accepted: 13 December 2024 Published: 31 December 2024

Abstract *- With the rapid proliferation of data within the digital landscape, a significant amount of dossier is available, ushering in "Big Data" concepts, embracing massive volume, velocity, and variety of information generated daily. The Phenomenon of Big Data presents significant challenges and opportunities for individuals, business organizations, and all entities that can use the concept to achieve strategic advantages. Extensive data engineering is a critical discipline focusing on designing, developing, and managing systems and architectures that enable effective data handling and analysis. The paper's objective is to explore Big Data evolution and the role of Big Data in informed decision-making in the real world. By integrating enhanced technologies such as Apache, Hadoop, and Spark, cloud computing platforms have changed the data processing landscape, allowing individuals, organizations, and entities to extract meaningful data for decision-making. The paper has examined the current trends in big data, including machine learning and artificial intelligence, and the increasing concerns about data security, privacy and availability, and the rising edge of computing. Additionally, the journal discusses challenges such as managing diverse data sources, ensuring data quality, and addressing the skills gap in the workforce. From such a multilevel lens, the paper offers a nuanced understanding of Big Data Engineering, OOP, and Cloud Platforms and their significant impact on organizational strategic decisions and performance. The paper reveals that integrating Big Data concepts supports an organization's innovation, enhances decision-making processes, and gains a competitive edge, ultimately reshaping the future of data management and analysis. Accordingly, the study's findings underscore notable significant data engineering investment, unlocking capabilities for valuable data as an asset in an increasingly complex world.*

Keywords *- Analytics, Big data, Cloud platforms, Data engineering and Hadoop.*

1. Introduction

The exponential data growth in the current digital era has created unprecedented challenges and opportunities for individuals, business organizations, and entities. With the undeniable evolution of data, the era of "Big Data" and revolution, a dossier of sheer volume, velocity, and variety of data necessitates innovative data management and analysis. Big Data Engineering is one crucial emergent in Big Data that equips organizations with a framework and tools to support big datasets, enhancing data-driven decisions and strategic advantages.

1.1. Big Data Comprehension

Big data defines a significant amount of data with underpinning sophistication that requires additional modem software, tools, and applications to unearth patterns, trends, and visualization for decision-making (Chen et al., 2020). "Big Data" encapsulates volume, speed, velocity, and variability, accommodating data formats, types, and diversity with the Big Data Engineering discipline. According to the authors, the Big Data concept has evolved from simple data collection to sophisticated data collection, storage, analysis, and visualization for readability, meaningfulness, and informed decisions. Fundamentally, significant data evolution resonates with the existing value of data as a critical asset for any organization making data-driven decisions. In understanding Big Data Concepts, velocity, volume, and variety underscore practical articulations, addressing pivotal challenges including big data (Shankar & Sahu, 2020). Volume defines a significant amount of datasets collected each second, with a projection demonstrating that by 2025, there will be 175 zettabytes of global data (Shankar & Sahu, 2020). Velocity is a significant aspect of big data, defining data transmission speed from one platform to another and encapsulating IoT, social media, and transactional systems. Variety describes significant data types as structured, unstructured, or semi-structured data from various sources.

1.2. The Importance of Big Data Engineering

With the need to embrace customer needs, demands, preferences, tastes, satisfaction, and experience, extensive data engineering has become vital in the current organizational settings. With a focus on data design, development, and management, the discipline is essential in collecting, analyzing, interpreting, and visualizing big data for business organizations and making informed decisions. Establishing data pipelines for adequate groundwork for data collection, storage, processing, analysis, and visualization is poignant in Big Data Engineering and deliverability. Big Data analysis provides a unique platform for business organizations to acquire enormous data and draw insightful information, patterns, and trends from the data, tailoring customer needs in goods and services delivery, profitability, and sustainability through data-driven business strategies (Minelli et al., 2020). In the real world, Big Data Engineering concepts enable organizations to collect information concerning customers' needs and experience with products and services. Later, business organizations analyze and interpret such data for customer-centric needs and expertise, supporting improved economies of scale and deliverability (Ranjan, 2012). Techniques such as predictive analytics create a platform for efficiency, cost-effectiveness, and seamless operations, hence Big Data Engineering applications to tailor a business organization's strategies toward a compelling competitive landscape in a digital space.

1.3. The Evolution of Big Data Technologies

Recent technological advancements in Big Data Engineering have demonstrated underpinning transformation in Big Data and the digital landscape. Distributed computing networks have created a platform for business organizations to collect data from various sources. One notable technology is Apache Hadoop, introduced in the 2000s, which created a critical revolution in big data processing, allowing seamless data collection, storage, processing, and analysis to support cost-effective data processing. Hadoop accommodates significant tools, including Hive and Pig, supporting data querying and transformation for business organizations' big data effective management and decision-making (Minelli et al., 2020). With the Apache Hadoop introduction, there has been real-time big data processing suitable for rapid and effective data processing and valuable contribution to business organizations. Additionally, increasing computing transformation in the Big Data Engineering discipline has created an exceptional platform for scalable data sources, accommodating changing data landscape and workloads, hence flexibility. Cloud computing platforms such as Microsoft Azure, Google, and Amazon Web Services have recently provided secure, cloud-based platforms for organizations to seamlessly store big data at cost-effective rates (Gandomi et al., 2020). Therefore, Big Data Engineering has accommodated vital evolution requiring practical understanding, especially Big Data Engineering, the Hadoop platform, and cloud-based data collection, storage, and processing areas.

1.4. Big Data Engineering and Industries

Big Data Engineering accommodates practical and real-world applications in almost all industries due to a desire to embrace data-driven decisions in all sectors. Specifically, Big Data Engineering concepts and techniques are applicable in healthcare, finance, transportation, innovative city management, agriculture, and technology, among other industries. (Zulkernine & Khelifa, 2016). For example, the healthcare industry uses predictive analytics by collecting and analyzing patient records, patterns, and trends to provide real-time diagnosis and effective patient-centric intervention and healthcare setting management, reducing errors toward increased and improved patient outcomes. In the finance industry, many transaction systems reoccurring require effective big data engineering techniques. For example, analyzing big transitional data in finance is pivotal in detecting fraud, managing risk, and optimizing business strategies. Moreover, using Machine learning and algorithms to identify patterns in financial transactions is critical in supporting data-driven financial decisions and projections (Zulkernine et al., 2016). Business organizations dealing in goods and services leverage big data by studying customers' behavior patterns, preferences, and experiences through predictive analytics, formulating marketing techniques, business strategies, and effective inventory management aligned with customers’ needs and executions, hence profitability, productivity, and sustainability.

1.5. Current Trends in Big Data Engineering

Big Data Engineering accommodates undeniable and evolving trends, shaping the future of Big Data and data-driven decisions among industries. Integrated artificial intelligence and machine learning in big data processing is a significant trend pivotal to the current techniques, tools, and frameworks of big data processing and understanding. According to the researchers, big data requires AI and ML algorithms that accommodate training and validation in the big data engineering platforms for effective data management (Hashem et al., 2015). Integrating AI and ML supports automated processes and intelligent technologies in organizations, enhancing rapid decision-making processes.

Moreover, there is an increased focus on data security and privacy, embracing policies, frameworks, and security tools for safe and secure data collection, storage, analysis, and transmission. For example, frameworks such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) permeate the Big Data Engineering discipline, focusing on personal data management and protection. Big Data Engineers comply with such frameworks and regulations to ensure data confidentiality, integrity, and availability in real-world applications, enforcing privacy and security concerns (Zulkernine & Khalifa, 2016). Edge computing is another emergent and trend concern in extensive data engineering involving data processing techniques, tools, and platforms. Edge computing is a concept that undercrosses data processing to get closer to data sources without overdependence on centralized data sources and servers. Significantly, the proliferation of Internet of Things devices facilitates real-time data collection, requiring Big Data

Engineers' adaptability, aligning with trends and future extensive data management through real-time collection, storage, processing, analysis, and transmission.

1.6. Challenges in Big Data Engineering

Big Data Engineering and applications in industries and sectors present far-reaching benefits and advantages. Despite such undeniable benefits, there are notable shortcomings that require practical acknowledgment to understand big data concepts and cloud-based platform Hadoop applications. Specifically, big data accommodates sophisticated and diverse data sources that require enhanced techniques, tools, and frameworks for effective data management and practices. Integrating data from various sources creates notable challenges for business organizations, hindering practical data analysis. As a concern in Big Data Engineering, there is a need for techniques for consolidating and harmonizing data from various sources to achieve unified data analysis, which is the primary problem of the paper.

Similarly, data quality is a critical shortcoming within Big Data Engineering that compromises considerable data accuracy and reliability for insightful decision-making. Adapting techniques, tools, and frameworks to support accuracy and reliability is a pivotal aspect for prominent data engineers indispensable for data validation, cleaning, and support to ensure insightful information. Big Data engineering disciplines require expertise and skills that are experiencing a gap, causing organizational difficulties. For example, there needs to be more skilled prominent data professionals, techniques, technologies, and methodologies vital for effective data management (IBM, 2021). Therefore, the background information requires organizations, scholars, and practitioners to train and develop to bridge existing gaps and challenges. In a nutshell, Big Data Engineering is a pivotal discipline, responding to exponential growth in big data due to increased use of technologies and digital platforms. Navigating through Big Data opportunities, challenges, Hadoop, and cloud platforms is an accessible role of prominent data engineers and scholars, providing practical implications and applications.

Harnessing such discrepancies in Big Data Engineering underscores an effective organization’s extensive data application to foster innovation, data-driven decisions, and increased competition within the digital space. Accordingly, understanding big data is an evolving process that unlocks possibilities and insightful information for individuals, business organizations, and other entities (Zulkernine & Khalifa, 2016). Therefore, the paper intends to examine the current trends in big data, including machine learning and artificial intelligence, and the increasing concerns about data security, privacy availability, liability, and the rising edge of computing. Additionally, the journal discusses challenges such as managing diverse data sources, ensuring data quality, and addressing the skills gap in the workforce.

2. The Rise of Big Data Engineering

With the organization's concern about big data and the extraction of insightful information from such enormous datasets, Big Data Engineering has emerged as a critical discipline. Specifically, the emergence of Big Data Engineering is a direct response to the increasing demand for managing vast datasets, prompting more sophisticated data processing, storage, collection, analysis, interpretation, and visualization (Katal et al., 2023). Section two of the paper provides an in-depth exploration of historical development, key influencers, technological increment, and the significant role of Big Data Engineering in shaping industries, establishing a unique platform for Hadoop and Cloud platforms background and development in Big Data Engineering, Hadoop, and cloud platforms.

2.1. Historical Development of Big Data Engineering

The root of big data is traceable beyond the 2000s, with extensive data engineering taking shape in the 2000s. With the digital space of business organizations in the 20th century, there has been a notable appreciation of big data by business organizations, accommodating transactions, customer interactions, and internal processes started to grow at an unprecedented rate (Laney, 2001). In the initial stages of extensive data engineering, data accommodated relational databases suitable for small and medium business organizations with undeniable challenges. Doug Laney introduced the three Vs. of big data, accommodating volume, velocity, and variety, creating a notable arena of increasing difficulties in data management due to rising datasets in the digital and technological space, encapsulating data generation speed, data diversity, and types, and amount of data available within the digital space.

From the Big Data management challenges, the field accommodates the need for new technologies, techniques, tools, and methodologies in collecting, processing, and transmitting data for data-driven decision-making in various organizations and industries. Distributed computing networking established a turning point in Big Data Engineering with notable types such as Apache Hadoop, describing an open-need framework source allowing the distribution and processing of big data across computer clusters. Introducing the Hadoop revolutionized data processing, creating a platform for business organizations to process a significant amount of data with seamless, cost-effective, and efficient tools, frameworks, and techniques.

2.2. Main Influencers of Big Data Engineering Increment

With the advent of technology and the dawn of the digital space, there is a discourse in the field of extensive data engineering as a response to the need for seamless tools, techniques, and approaches to managing big data. According to the authors, various factors contribute to Big

Data Engineering response, including data exploration, advancement in computing power, increasing importance of data-driven decisions, Internet of Things, AI, and ML advancement (Katal et al., 2013). Thus, effective articulation and discussion on such drivers in promoting the field of extensive data engineering aligns with the paper's objectives, creating a practical understanding of extensive data engineering, Hadoop, and cloud platforms in big data analysis and processing toward informed decision-making

2.2.1. The Explosion of Data

The increasing amount of data in the digital space is the main contributor to the development of extensive data engineering. According to the data, in 2024, approximately 97 zettabytes of data will be generated globally (Statista, 2023). The increasing access to social media, e-commerce, IoT devices, cloud computing, and smartphones has led to increased data generation, causing enormous amounts of data available on such platforms. For example, social media interaction, such as likening and positing of videos, creates a platform for data storage and can be analyzed to provide insightful information. Importantly, business organizations embrace modern data processing, leaving traditional data processing methods that need to be improved. Therefore, there is a surging need for prominent data engineers to design, build, and maintain big data management platforms.

2.2.2. Advancements in Computing Power

Increased computing power is another vital reason for the rise of extensive data engineering as a field and discipline. Distributed Computing frameworks such as Hadoop and current technologies such as Apache Spark provide a platform for business organizations, individuals, and government entities to process significant amounts of data, calling for prominent data engineers. Technologies such as Hadoop provide a platform for parallel processing of data with data from different sources that are analyzable differently, creating an effective and seamless platform for data processing. Access to cloud computing platforms such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud has created a platform for adequate data storage and processing without high-cost hardware investment, triggering the need for extensive data engineering to understand and maintain such platforms (Shankar & Sahu, 2020). Cloud computing offers flexible, scalable, and adequate data workload management and performance.

2.2.3. The Increasing Importance of Data-Driven Decision Making

In the current business platform, data is a critical asset to organizations seeking insight from existing data to perform data-driven decisions. Companies use data from various sources to make informed answerability aligned with organization objectives to meet personalized customer recommendations, supporting performance, profitability, and sustainability. For example, business organizations use social media posts and customer comments on their products and services to tailor marketing strategies that meet customers' needs, expectations, and experiences. The decision-making shift from traditional to data-driven models has triggered the rise of extensive data engineering as a field for effective big data processing and management.

Big data analytics, such as predictive, provide insightful information for business organizations to effectively manage inventory by analyzing customers' behavior, such as purchase patterns, to improve efficiency (Shankar & Sahu, 2020). Similarly, healthcare organizations use predictive analytics to inform patient-centric models and decisions. Therefore, increased data-driven decision-making has influenced the rise of extensive data engineering as a discipline for effective data processing and management.

2.2.4. Growth of the Internet of Things (IoT)

With the growing Internet of Things, a significant amount of data is collected, stored, and transmitted, calling for Big Data Engineering and techniques in connecting devices and data flow (Messaoudi et al., 2024). IoT Devices, wearable and smart home devices generate big data that require modern and enhanced collection, storage, analysis, and transmission techniques achievable through Big Data Engineering techniques, tools, and methods, hence the rise in Big Data Engineering.

Noteworthy, prominent data engineers play a significant role in managing influx dossier from IoT devices, designing and implementing real-time systems for effective data management, hence the need for prominent engineers to harness big data for predictive performance, real-time monitoring, and increased decision-making.

2.2.5. Advancements in Machine Learning and Artificial Intelligence

Machine Learning and Artificial Intelligence are two underpinning occurrences that triggered the growth of Big Data Engineering as a discipline. Machine Learning accommodates algorithms requiring significant data for model training, prediction, and classification. Available big data and computational create an exceptional platform for developing AI models performing in natural language processing, image recognition, and predictive analysis (Shankar & Sahu, 2020). Admittedly, integrating concepts in extensive data engineering is crucial in AI and ML, creating data pipelines, labeling data, and ensuring positive outcomes and deliverability. Therefore, an increasing demand for AI and Machine Learning promotes the need for prominent data engineers.

2.3. Main Technologies and Tools in Big Data Engineering

Big Data Engineering resonates with technological developments and tools applicable in effectively processing and managing big data. Several tools and techniques are applicable in Big Data Engineering, including Apache Hadoop, Apache Spark, cloud platforms, data lakes, and warehouses. These are essential in big data engineering performance and deliverability in data collection, storage, analysis, transformation, and visualization towards informed decisions.

2.3.1. Apache Hadoop

Hadoop is one significant technological development in extensive data engineering, defining a distributed computing network and allowing ample data storage and processing from different computer clusters (Zikopoulos & Eaton, 2011). Notably, Hadoop components include the Hadoop Distributed File System (HDFS) and MapReduce programming model, enabling parallel processing of data and making it possible to handle massive datasets that would be impossible to process on a single machine. Hadoop accommodates an open-source nature and scalability, which is appropriate for most organizations. Moreover, Hadoop has an ecosystem that encapsulates various tools such as Apache Hive (data querying), Apache Pig (Data Transformation), and Hbase (for real-time data processing (White, 2015). These tools enhance data querying, transformation, and real-time application capabilities in the Big Data Engineering platform.

2.3.2. Apache Spark

In the current Big Data Engineering atmosphere, Apache Spark is one notable technological development alternative to Hadoop's MapReduce, which is applicable in processing large datasets. Apache Spark provides a batch and real-time data processing platform suitable for rapid and in-memory computations in data processing and management (White, 2015). Moreover, Apache Spark provides an exceptional arena for speed and flexibility, hence applicable for big data applications, real-time processing, and effective visualization through machine learning, graph processing, and stream processing. Apache Spark presents scalability and integrative nature with Hadoop and other cloud platforms, hence extended adaptation by business organizations in batch and real-time data processing.

2.3.3. Cloud Platforms

In the development of extensive data engineering, cloud platforms such as AWS, Google Clouds, and Microsoft Azure permeate the current extensive data analysis and storage on cloud-based models. Cloud platform offers seamless areas for storing, processing, and analyzing big data without investing in physical hardware for data storage and processing. For example, Amazon S3 and Google Cloud Storage are pivotal in data storage in cloud-based platforms (Zhang et al., 2020). Technologies such as Amazon EMR, Azure HDInsight, and machine learning are essential in data processing on cloud platforms. Similarly, AWS SageMaker and Google AI platforms are crucial to real-time data processing and visualization for informed business decisions. Cloud-based platforms posit significant advantages over traditional methods, including scalability, flexibility, and seamless data migration. Therefore, the Cloud platform offers managed services, reducing data collection and processing sophistication and focusing on data analysis rather than infrastructure management.

2.3.4. Data Lakes and Data Warehouses

Data storage is one critical aspect that requires practical management within Big Data Analytics and performance. Specifically, efficient data storage solutions that accommodate data lakes and warehouses are critical technological developments for real-time solutions for extensive data management and applications. A data lake is a centralized repository that stores raw, unstructured, and semi-structured data in its native format, allowing organizations to store massive amounts of data at a lower cost. This data can then be processed and analyzed using big data tools.

Conversely, data warehouses are optimized for structured data, analytical queries, and reporting. Modern cloud-based data warehouses, such as Amazon Redshift and Google BigQuery, offer fast query performance and scalability, making them suitable for business intelligence and analytics use cases. Similarly, Data warehouse describes data warehouses as optimized for structured data and are used for analytical queries and reporting. Modern cloud-based data warehouses, such as Amazon Redshift and Google BigQuery, offer fast query performance and scalability, making them suitable for business intelligence and analytics use cases.

2.4. The Role of Big Data Engineering Across Industries

Extensive data engineering has transformed various industries, enabling organizations to unlock new opportunities and improve their operations. Some of the critical sectors where extensive data engineering has made a significant impact include:

2.4.1. Healthcare

In the healthcare industry, extensive data engineering has enabled the analysis of vast amounts of patient data to improve treatment outcomes, reduce costs, and enhance the overall quality of care. By analyzing Electronic Health Records (EHRs), genomic data, and medical imaging, healthcare organizations can identify patterns and trends that inform diagnosis and treatment decisions (Zorba et al., 2020). Additionally, extensive data engineering has enabled the development of predictive analytics models that can identify patients at risk of developing chronic conditions, allowing for early intervention and personalized care.

2.4.2. Finance

The finance industry has leveraged extensive data engineering to gain insights into customer behavior, detect fraudulent activities, and optimize trading strategies. Financial institutions analyze transaction data, social media sentiment, and market trends to develop predictive models that inform investment decisions and risk management strategies. Additionally, extensive data engineering has enabled the development of real-time fraud detection systems that monitor transactions for suspicious activity.

2.4.3. Retail

Extensive data engineering has enabled companies to analyze customer behavior and preferences in the retail sector, leading to more personalized marketing and improved customer experiences. Retailers use data from online transactions, loyalty programs, and social media interactions to gain insights into customer preferences, optimize pricing strategies, and improve inventory management. By analyzing historical sales data, retailers can forecast demand and adjust their inventory levels accordingly, reducing the risk of stockouts or overstocking.

2.4.4. Manufacturing

Extensive data engineering has also played a critical role in optimizing manufacturing operations. By analyzing data from sensors embedded in machinery, manufacturers can monitor equipment performance in real-time, identify potential issues, and perform predictive maintenance. This reduces downtime, improves efficiency, and extends the lifespan of equipment. Additionally, extensive data engineering has enabled manufacturers to optimize their supply chains by analyzing data from suppliers, logistics providers, and production lines.

2.4.5. Transportation and Logistics

Extensive data engineering has enabled companies to optimize route planning, reduce fuel consumption, and improve delivery times in the transportation and logistics industry. Logistics companies can identify the most efficient routes and adjust their operations in real-time by analyzing data from GPS devices, traffic sensors, and weather reports. Additionally, extensive data engineering has enabled the development of predictive maintenance systems for vehicles, reducing the risk of breakdowns and ensuring timely deliveries.

2.5. Big Data Engineering Future

Advent technologies and innovations create an effective platform for significant data engineering evolution. Artificial Intelligence, Machine learning integration, and edge computing are possible drivers for the next evolution within extensive data engineering. They benefit business organizations in terms of real-time data processing and performance. In the future of extensive data engineering, enormous amounts of data will be accessible, and there are implications for ethical considerations and privacy that will drive the need to adhere to such ethics in collecting, analyzing, processing, and transmitting data. In summary, the rise of extensive data engineering accommodates data exploration, computing power advancement, the increased need for data-based decisions, IoT, Machine Learning, and AI applications. Therefore, business organizations require prominent data engineers to manage and process data, accommodating modern and real-time tools, techniques, and platforms for effective and secure data collection, storage, processing, and transmission.

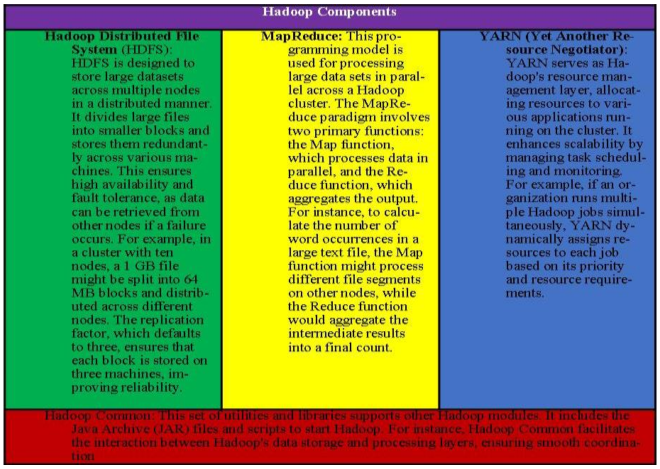
3. Overview of Big Data Engineering

3.1. Overview of Big Data Engineering

Big Data Engineering involves systematically designing, developing, and deploying infrastructures to store, process, and analyze vast amounts of structured and unstructured data. It encompasses various technologies, tools, and methodologies, including Hadoop, Spark, and cloud platforms like AWS and Microsoft Azure. This section comprehensively summarizes the critical materials and methods for implementing extensive data systems.

3.2. Hadoop Ecosystem

The Hadoop ecosystem is one significant aspect of big data engineering projects with the open-structure framework, allowing distributed storage and big data processing within computer clusters (White, 2015). Significantly, a practical understanding of Hadoop components, including Hadoop Distributed File System, MapReduce, YARN, and Hadoop Common, underscore effective practice and application in the data engineering discipline. The figure below illustrates Hadoop components for practical understanding and implications in extensive data engineering and applications.



Hadoop Components

Hadoop Distributed File System (HDFS): HDFS is designed to store large datasets across multiple nodes in a distributed manner. It divides large files into smaller blocks and stores them redundantly across various machines. This ensures high availability and fault tolerance, as data can be retrieved from other nodes if a failure occurs. For example, in a cluster with ten nodes, a 1 GB file might be split into 64 MB blocks and distributed across different nodes. The replication factor, which defaults to three, ensures that each block is stored on three machines, improving reliability.

MapReduce: This programming model is used for processing large data sets in parallel across a Hadoop cluster. The MapReduce paradigm involves two primary functions: the Map function, which processes data in parallel, and the Reduce function, which aggregates the output. For instance, to calculate the number of word occurrences in a large text file, the Map function might process different file segments on other nodes, while the Reduce function would aggregate the intermediate results into a final count.

YARN (Yet Another Resource Negotiator): YARN serves as Hadoop's resource management layer, allocating resources to various applications running on the cluster. It enhances scalability by managing task scheduling and monitoring. For example, if an organization runs multiple Hadoop jobs simultaneously, YARN dynamically assigns resources to each job based on its priority and resource requirements.

Hadoop Common: This set of utilities and libraries supports other Hadoop modules. It includes the Java Archive (JAR) files and scripts to start Hadoop. For instance, Hadoop Common facilitates the interaction between Hadoop's data storage and processing layers, ensuring smooth coordination

Fig. 1 Hadoop components

3.3. Tools and Frameworks in the Hadoop Ecosystem

Hadoop accommodates significant tools and frameworks, extending functionality, performance, and deliverability. Notably, Apache Hive, Apache Pig, Apache Hbase, Apache Flume, and Kafka underscore effective performance and functionality.

3.3.1. Apache Hive

Hive is a data warehousing solution that allows users to write SQL-like queries to process data stored in Hadoop. It abstracts the complexity of writing MapReduce programs, making it easier for developers and analysts to interact with large datasets (White, 2015). For example, an analyst could use Hive to run a SQL query on a dataset stored in HDFS to analyze sales trends over time.

3.3.2. Apache Pig

Pig is a high-level platform for creating MapReduce programs. Its scripting language, Pig Latin, simplifies the process of writing complex data transformations and analyses (Choudhury, 2021). For example, a data engineer could use Pig to filter, sort, and join data from multiple sources before performing a statistical analysis.

3.3.3. Apache HBase

HBase is a distributed, scalable NoSQL database that provides real-time read and write access to large datasets (Choudhury, 2021). It is built on top of HDFS and supports random access to data, making it ideal for applications that require quick access to specific pieces of information. For instance, a telecommunications company might use HBase to store and retrieve millions of customers' call detail records (CDRs).

3.3.4. Apache Flume and Kafka

These are data ingestion tools used to collect, aggregate, and move large amounts of log data or streaming data into Hadoop (Choudhury, 2021). For example, Flume can collect logs from web servers and transport them to HDFS, while Kafka enables real-time processing of streaming data, such as sensor readings from IoT devices.

3.4. Apache Spark Integration

Apache Spark is applicable within the Hadoop ecosystem with in-memory processing capabilities, offering improved performance than traditional MapReduce (Choudhury, 2021). Spark provides effective batch and stream processing, hence versatile applications and integrations in Big Data applications.

Noteworthy, Apache Spark presents significant features, including Resilient Distributed Datasets (RDDs), Data Frame API and SparkSQL, Spark Streaming, and MLlib (Machine Learning Library).

3.4.1. Resilient Distributed Datasets (RDDs)

RDDs are fault-tolerant collections of objects that can be processed in parallel across a cluster (Apache Software Foundation., 2020). For example, Spark can store an entire dataset in memory across the cluster and perform real-time transformations such as filtering and mapping.

3.4.2. DataFrame API and SparkSQL

These APIs allow users to interact with structured data using SQL queries (Apache Software Foundation., 2020). For instance, a data scientist might use SparkSQL to query a large dataset of customer transactions stored in HDFS to identify spending patterns.

3.4.3. Spark Streaming

This enables real-time processing of data streams (Apache Software Foundation., 2020). For example, an organization might use Spark Streaming to monitor and analyze real-time social media activity, identifying trending topics and customer sentiment as they emerge.

3.4.4. MLlib (Machine Learning Library)

Spark's MLlib provides a scalable machine learning platform, supporting various classification, regression, clustering, and recommendation systems algorithms. For example, a retail company could use MLlib to build recommendation models that suggest products to customers based on their past purchases.

3.5. Cloud Platforms for Big Data Engineering

With increasing technologies and advancement, Big Data Engineering accommodates cloud platforms vital for real-time data storage and processing for seamless decision-making. Specifically, Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP) are leading cloud platforms in extensive data engineering and application, creating an effective platform for data processing and management. Specifically, a cloud platform offers on-demand computing services, resources, and scalability necessary for organization data and processing.

3.5.1. Amazon Web Services (AWS)

Amazon We-Service offers myriad services, such as Amazon EMR, which supports a user interface for running Hadoop and Spark clusters within cloud-based platforms. EMR abstracts the complexity of managing hardware and provides pre-configured environments for running big data applications (White, 2015). For instance, a company could use EMR to process petabytes of log data and generate real-time analytics dashboards.

3.5.2. Microsoft Azure

Azure provides significant services, including Azure HDInsight, a fully managed cloud service that allows users to run open-source frameworks like Hadoop, Spark, and Hive. HDInsight supports seamless integration with other Azure services, such as Azure Data Lake and Power BI, for end-to-end data processing and visualization (Baesens & Van Huffel, 2020). Therefore, Azure is one significant Microsoft cloud-based platform that supports practical, extensive data engineering applications and integrations.

3.5.3. Google Cloud Platform (GCP)

GCP provides Google Cloud Dataproc, a managed Spark and Hadoop service that allows users to process large datasets using familiar Hadoop ecosystem tools. Dataproc offers fast cluster provisioning and integrates with Google BigQuery for scalable data analytics. For example, a media company could use Dataproc to process video metadata and generate recommendations for personalized content.

3.6. Data Processing Workflow

The data processing workflow in big data engineering projects typically involves several stages, from data ingestion to transformation and analysis (Hashem, 2015). The workflow in Hadoop and cloud platforms involves stepwise stages, including data ingestion, storage, transformation, analysis, visualization, and reporting.

3.6.1. Data Ingestion

Data is collected from various sources, such as databases, logs, sensors, and external APIs. Tools like Apache Flume or Kafka stream data into the Hadoop cluster or cloud storage systems like Amazon S3 or Azure Data Lake. For example, a transportation company might use Kafka to ingest real-time GPS data from its fleet of vehicles.

3.6.2. Data Storage

After ingestion, data is stored in distributed file systems such as HDFS or cloud storage services. The data is partitioned and replicated across multiple nodes to ensure fault tolerance and availability.

For example, a social media platform might store user activity logs in HDFS to support large-scale analysis of user behavior.

3.6.3. Data Transformation

Data is cleaned, filtered, and transformed into a suitable format for analysis. This step often involves writing MapReduce jobs, Pig scripts, or Spark transformations. For instance, an e-commerce company might use Spark to filter out incomplete transactions and aggregate sales data by region.

3.6.4. Data Analysis

Once the data is prepared, it is analyzed using tools such as Apache Hive, SparkSQL, or machine learning algorithms in Spark's MLlib. For example, a financial services firm could use Hive to run SQL queries on transaction data and detect fraudulent activities.

3.6.5. Data Visualization and Reporting

The analysis results are visualized and reported using tools like Tableau, Power BI, or custom-built dashboards. For example, a healthcare provider might create a real-time dashboard to monitor patient admissions and resource utilization.

4. Case Study: Retail Industry Use of Hadoop and Cloud Platforms

A real-world example of the application of Hadoop and cloud platforms can be seen in the retail industry. A leading global retailer implemented a big data solution using Amazon EMR to process large volumes of transaction data from its online and physical stores. The retailer used Hadoop and Spark on EMR to analyze customer purchasing patterns, enabling the company to offer personalized recommendations and optimize inventory management. The workflow began with data ingestion, where transaction logs were streamed from the Point-of-Sale (POS) systems to Amazon S3 (Jin et al., 2021). The data was then processed using Spark on EMR, which performed data cleaning, filtering, and aggregation transformations. The final dataset was stored in Amazon Redshift, a data warehousing solution, where analysts could run complex SQL queries and generate reports on customer behavior. Through this implementation, the retailer significantly improved customer satisfaction and operational efficiency. The scalable nature of the cloud platform allowed the company to handle seasonal spikes in data volume without investing in additional on-premise infrastructure.

4.1. Overview of Results

Big Data Engineering on Hadoop and cloud platforms transform industries such as healthcare, Retail, finance, entertainment, and others, calling for another active channel. In the paper, results are determinable through improved data processing efficiency, enhanced decision-making capabilities, cost savings, and the ability to handle larger datasets than traditional methods. Specifically, the main findings in the systematic review improved scalability and data proceeding through Hadoop's distributed computing framework, coupled with cloud infrastructure, allows organizations to process terabytes and even petabytes of data in a fraction of the time previously required by traditional databases (Jin et al., 2021).

Moreover, there are increased data analysis capabilities with significant tools, including Apache Spark, Hive, and HBase, with myriad functionalities in batch and real-time data processing platforms, acknowledging machine learning and big data analytics. The application of Big Data Engineering on Hadoop and cloud platforms has produced transformative results in numerous industries, including healthcare, Retail, finance, and entertainment. These results can be measured in terms of improved data processing efficiency, enhanced decision-making capabilities, cost savings, and the ability to handle larger datasets than traditional methods (Kumar et al., 2019). This section delves into specific results and comprehensively discusses the broader implications. Moreover, extensive data engineering and Hadoop applications on cloud platforms result in cost-effectiveness and flexibility through pay-as-you-go and cloud-based storage systems, eliminating hardware investment costs and flexibility in significant data migration through cloud-based platforms (Jin et al., 2021). Equally, strategic and rapid decisions within business organizations and industries through Apache Kafka and Spark Streaming support real-time analytics for organizations to rapidly collect, analyze, and process data, enhancing data-driven decision-making.

4.2. Hadoop and Cloud Platform Applications

With the need to embrace customer needs, demands, preferences, tastes, satisfaction, and experience, extensive data engineering has become vital in the current organizational settings. With a focus on data design, development, and management, the discipline is essential in collecting, analyzing, interpreting, and visualizing big data for business organizations and informed decisions (White, 2015). Establishing data pipelines for adequate groundwork for data collection, storage, processing, analysis, and visualization is poignant in Big Data Engineering and deliverability. Big Data analysis provides a unique platform for business organizations to acquire enormous data and draw insightful information, patterns, and trends from the data, tailoring customer needs in goods and services delivery, profitability, and sustainability through data-driven business strategies (Minelli et al., 2020).

In the real world, Big Data Engineering concepts enable organizations to collect information concerning customers' needs and experience with products and services. Later, business organizations analyze and interpret such data for customer-centric needs and expertise, supporting improved economies of scale and deliverability. Techniques such as predictive analytics create a platform for efficient, cost-effective, and seamless operations, hence Big Data Engineering applications to tailor a business organization's strategy toward a compelling competitive landscape in a digital space.

4.2.1. Healthcare

In the healthcare industry, adopting big data solutions has enabled providers to make significant strides in patient care, research, and operational efficiency. One example is using Hadoop and cloud platforms to process vast amounts of medical data, such as Electronic Health Records (EHRs), genomic data, and sensor data from wearable devices. For instance, large hospitals and research institutions have employed Hadoop and Spark to analyze patient records, identifying patterns that may indicate the onset of chronic diseases such as diabetes and heart disease (Chen et al., 2014). By leveraging machine learning algorithms, healthcare providers can predict patient outcomes more accurately and offer personalized treatment plans. Moreover, cloud platforms have facilitated the secure storage and sharing of medical data. Providers can now collaborate across institutions and geographies more effectively, enabling advances in medical research. For example, using cloud-based big data platforms, researchers could analyze genomic datasets of cancer patients and identify novel gene mutations linked to different cancer types.

4.2.2. Retail and E-Commerce

Retailers have also reaped significant benefits from big data solutions. Hadoop and cloud platforms have enabled companies to process customer transaction data, social media interactions, and website logs to gain deeper insights into consumer behavior. Companies can optimize their marketing strategies by analyzing these datasets, personalizing customer experiences, and improving inventory management. For instance, Amazon leverages big data to analyze millions of users' browsing and purchasing patterns. Using this data, Amazon's recommendation engine can suggest products customers will likely buy based on their previous purchases and interactions on the platform (Messaoudi et al., 2024). Similarly, large-scale retailers like Wal-Mart have implemented real-time analytics systems using Hadoop and Spark to monitor customer transactions and adjust pricing dynamically based on supply and demand. The retail sector also relies on big data for predictive analytics, helping companies anticipate market trends and adjust their inventory accordingly. This predictive ability minimizes the costs associated with overstocking or understocking products.

4.2.3. Finance

Financial institutions are among the most significant users of big data technologies. Hadoop and cloud platforms analyze massive volumes of transactional data, customer profiles, and market trends. By leveraging real-time analytics, banks can identify fraudulent transactions, assess credit risks, and enhance customer service offerings. One prominent use case in finance involves risk management. Large banks and hedge funds use Hadoop to process historical financial data and apply machine learning algorithms to model and predict market risks. These insights allow traders and analysts to make informed decisions and minimize potential losses. In the insurance industry, big data platforms are used to calculate premiums more accurately (Messaoudi et al., 2024). By analyzing a customer's historical data and behavior patterns, insurers can offer personalized policies and assess risks more precisely. Furthermore, cloud platforms ensure that data is securely stored and accessible to authorized users, ensuring compliance with regulations like the General Data Protection Regulation (GDPR).

4.2.4. Entertainment and Media

The entertainment and media industries have also benefited from big data technologies. Companies such as Netflix, Spotify, and YouTube rely heavily on Hadoop and cloud-based systems to deliver personalized recommendations to their users. For example, Netflix processes massive amounts of viewer data, including search queries, viewing times, and ratings, to make recommendations that cater to individual preferences (Messaoudi et al., 2024). Using machine learning models on Spark, Netflix can continuously improve its recommendation engine, keeping users engaged and increasing subscription retention rates.

4.3. Comparing Hadoop and Cloud-Only Platforms

Hadoop platforms accommodated significant use of distributed computing networks, allowing data processing from computer clusters. Cloud platforms, including Amazon EMR, Google BigQuery, and Microsoft Azure, effectively changed extensive data engineering. Importantly, comparative analysis of Hadoop and Cloud Platform accommodates performance and scalability, cost, security and compliance, and maintenance and management.

4.3.1. Performance and Scalability

In Hadoop, data processing tasks can be distributed across clusters of computers through nodes. Conversely, Cloud-based platforms provide unlimited and virtual scalability (Zhang et al., 2020). For example, Amazon EMR and Google BigQuery allow users to spin up clusters on demand**,** ensuring that computational resources are always available when needed. This flexibility is particularly beneficial for organizations experiencing fluctuating workloads.

4.3.2. Cost Considerations

Cost is vital to Big Data Engineering, especially Hadoop and Cloud-based platforms. Hardware costs are needed in the on-premises Hadoop application, and long-term management and maintenance are expensive. Cloud platforms operate on a pay-as-you-go model, allowing organizations to avoid large capital expenditures.

Instead of investing in hardware, companies only pay for the resources they use, reducing overall costs (Kumar et al., 2019). However, long-term cloud usage can become expensive, especially for businesses with continuous high-volume data processing needs.

4.3.3. Security and Compliance

Both Hadoop and cloud platforms face security challenges, but cloud services have made significant strides in providing robust security features. Cloud providers offer encryption, secure authentication mechanisms, and compliance with industry standards such as GDPR and the Health Insurance Portability and Accountability Act (HIPAA).

Hadoop clusters require meticulous configuration to ensure data security (Zhang et al., 2020). Organizations must implement encryption at rest and in transit, secure user access controls, and monitor for unauthorized activity. In contrast, cloud providers offer built-in security tools and monitoring systems, reducing the burden on IT teams.

4.3.4. Maintenance and Management

One of the critical advantages of cloud platforms is that they abstract much of the maintenance and management burden from users. Cloud services automatically handle software updates, security patches, and infrastructure management, allowing organizations to focus on their data engineering tasks. Hadoop, on the other hand, requires constant attention. Clusters must be monitored for performance issues, disk space needs to be managed, and periodic upgrades must be performed to keep the system running smoothly. This management overhead can significantly disadvantage smaller organizations with limited IT resources.

4.4. Challenges of Big Data Engineering on Hadoop and Cloud Platforms

Despite promising and far-reaching benefits of Hadoop and Cloud-based platform applications in extensive data engineering, there are undeniable setbacks, including data integration, skill gap, data security, privacy, and infrastructure complexities that impact overall practice, application, and implementation in extensive data engineering. Big Data Engineering and applications in industries and sectors present far-reaching benefits and advantages. Despite such undeniable benefits, there are notable shortcomings that require practical acknowledgment to understand big data concepts and cloud-based platform Hadoop applications. Specifically, big data accommodates sophisticated and diverse data sources that require enhanced techniques, tools, and frameworks for effective data management and practices.

Integrating data from various sources creates notable challenges for business organizations, hindering practical data analysis (Zorba et al., 2017). As a concern in Big Data Engineering, there is a need for techniques for consolidating and harmonizing data from various sources to achieve unified data analysis, which is the paper's primary concern. Similarly, data quality is a critical shortcoming within Big Data Engineering that compromises considerable data accuracy and reliability for insightful decision-making. Adapting techniques, tools, and frameworks to support accuracy and reliability is a pivotal aspect for prominent data engineers indispensable for data validation, cleaning, and support to ensure insightful information.

Big Data engineering disciplines require expertise and skills that are experiencing a gap, causing organizational difficulties. For example, there needs to be more skilled, prominent data professionals, and techniques, technologies, and methodologies are vital for effective data management (Zorba et al., 2017). Therefore, the background information requires organizations, scholars, and practitioners to train, develop, and ridge existing gaps and challenges.In a nutshell, Big Data Engineering is a pivotal discipline, responding to exponential growth in big data due to increased use of technologies and digital platforms. Navigating through Big Data opportunities, challenges, Hadoop, and cloud platforms is an accessible role of prominent data engineers and scholars, providing practical implications and applications. Harnessing such discrepancies in Big Data Engineering underscores an effective organization's extensive data application to foster innovation, data-driven decisions, and increased competition within the digital space. Accordingly, understanding big data is an evolving process that unlocks possibilities and insightful information for individuals, business organizations, and other entities (Zulkernine & Khalifa, 2016). Therefore, the paper intends to examine the current trends in Big Data, including Machine learning, Artificial Intelligence, and increasing concerns on data security, privacy ava, liability, and growing computing edge (Zorba et al., 2017). Additionally, the journal discusses challenges such as managing diverse data sources, ensuring data quality, and addressing the skills gap in the workforce.

4.5. Future Directions and Trends in Big Data Engineering

Big Data Engineering evolves continuously and regularly, impacting overall practices and trends. Machine learning and Artificial Intelligence trends shape the future of extensive data engineering in myriad ways.

4.5.1. AI and Machine Learning Integration

AI and Machine learning are integrated into Big Data Engineering via platforms such as MLlib and TensorFlow on Hadoop, allowing organizations to build advanced models for predictive analytics, natural language processing, and image recognition. Therefore, integrating ML and AI is vital in formulating and forgoing insightful methods in big data analytics platforms.

4.5.2. Edge Computing

Edge computing is another emergent and trend concern in extensive data engineering involving data processing techniques, tools, and platforms. Edge computing is a concept that undercrosses data processing to get closer to data sources without overdependence on centralized data sources and servers. Notably, the proliferation of Internet of Things devices facilitates real-time data collection, requiring Big Data Engineers' adaptability, aligning with trends and future extensive data management through real-time collection, storage, processing, analysis, and transmission.

4.5.3. Ethical Considerations

Advent technologies and innovations create an effective platform for significant data engineering evolution. Artificial Intelligence, Machine learning integration, and edge computing are possible drivers for the next evolution within extensive data engineering. They benefit business organizations in terms of real-time data processing and performance. In the future of extensive data engineering, enormous amounts of data will be accessible, and there are implications for ethical considerations and privacy that drive the need to adhere to such ethics in collecting, analyzing, processing, and transmitting data.

5. Conclusion

The exponential data growth in the current digital era has created unprecedented challenges and opportunities for individuals, business organizations, and entities. With the undeniable data evolution, the era of "Big Data," and revolution, a dossier of sheer volume, velocity, and variety of data necessitates innovative underscores in data management and analysis. Big Data Engineering is one crucial emergent in Big Data that equips organizations with a framework and tools to support big datasets, enhancing data-driven decisions and strategic advantages.

5.1. Results Summary

The paper has explored intricate concerns in Big Data Engineering with an in-depth analysis of underlying technologies in Hadoop and Cloud-based platforms. In the paper, results are determinable through improved data processing efficiency, enhanced decision-making capabilities, cost savings, and the ability to handle larger datasets than traditional methods. Specifically, the main findings in the systematic review improved scalability and data proceeding through Hadoop's distributed computing framework, coupled with cloud infrastructure, allowing organizations to process terabytes and even petabytes of data in a fraction of the time previously required by traditional databases. Moreover, there are increased data analysis capabilities with significant tools, including Apache Spark, Hive, and HBase, with a myriad of functionalities in batch and real-time data processing platforms, acknowledging machine learning and big data analytics

5.2. Implications for Organizations

According to the journal findings, organizations and individuals benefit from extensive data engineering on Hadoop and cloud platforms, facilitating data-driven decisions directly connected to performance, sustainability, ty, and profitability. Significant practical implications of the findings include improved decision-making, operations efficiency, and competitive advantages.

5.2.1. Improved Decision-Making

One of the most significant implications of using Hadoop and cloud platforms is the ability to improve decision-making processes within organizations. Analyzing historical and real-time data allows businesses to identify trends, predict future outcomes, and confidently make informed decisions. For example, in Retail, companies can use big data analytics to optimize inventory management and pricing strategies by predicting consumer behavior. Similarly, in healthcare, big data enables more precise and personalized patient care by predicting disease patterns, treatment outcomes, and resource allocation. Processing large amounts of data quickly allows healthcare providers to act proactively rather than reactively, leading to better patient outcomes and cost savings.

5.2.2. Enhanced Operational Efficiency

Another critical implication is the enhancement of operational efficiency. Organizations can automate and optimize various processes by implementing big data engineering solutions, from supply chain management to marketing campaigns. In the finance industry, for instance, companies have used real-time data analytics to automate fraud detection and risk management systems, resulting in improved security and reduced losses. Hadoop's ability to process unstructured data from multiple sources, combined with the cloud's flexibility, allows businesses to integrate and analyze data seamlessly. These results in streamlined operations and more efficient workflows, reducing time and resources spent on manual processes. In sectors like Manufacturing, predictive maintenance powered by big data helps prevent costly machine downtime by identifying potential failures before they occur.

5.2.3. Competitive Advantage

Organizations that embrace extensive data engineering are also likely to gain a competitive edge in their respective markets. By leveraging insights derived from data analytics, companies can develop innovative products and services, improve customer satisfaction, and stay ahead of competitors. In the digital age, the ability to analyze consumer preferences, behaviors, and market trends in real-time is a critical driver of success. For example, Netflix has capitalized on big data by using machine learning algorithms to offer personalized content recommendations, leading to higher user engagement and retention. The use of big data has been a critical differentiator for the company, allowing it to anticipate viewer preferences and deliver tailored content more effectively than traditional media companies.

5.3. Addressing Challenges

The advantages of Big Data Engineering on Hadoop and cloud platforms are precision, offering undeniable, seamless, and feasible big data management platforms; hence, challenges need to be addressed and real-time solutions created. Addressing challenges such as data integration and quality, skills required, security and privacy, and infrastructure complexities. Organizations often need help integrating data from disparate sources, which may be in different formats or stored in various locations. This makes creating a unified view of the data difficult, which is essential for accurate analysis. In addition, ensuring data quality remains a critical issue. Only complete, consistent, or accurate data can lead to flawed insights, undermining the benefits of big data analytics.

Organizations must invest in robust data integration tools and processes that ensure data consistency and accuracy. Solutions like ETL (Extract, Transform, and Load) tools and data lakes can help manage and integrate large datasets. Additionally, establishing data governance frameworks can improve data quality and compliance with regulations. To mitigate this issue, companies must invest in training and upskilling their existing workforce while collaborating with educational institutions to foster the next generation of data professionals. By creating partnerships with universities and offering internships, organizations can tap into emerging talent and fill the skills gap in the long run.

Organizations must implement comprehensive security strategies to protect sensitive information, including encryption, access controls, and regular security audits. Compliance with data privacy regulations, such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA), is critical in maintaining customer trust and avoiding legal consequences. Cloud-native big data platforms can simplify infrastructure management by abstracting much of the complexity. Services like Amazon EMR, Google BigQuery, and Microsoft Azure HDInsight provide managed environments where organizations can run their big data workloads without dealing with the intricacies of cluster management.

5.4. Future Directions

The evolution of Big Data Engineering continuously evolves, impacting extensive data engineering and practices (Luan et al., 2020). Machine learning and artificial intelligence trends shape the future of extensive data engineering in myriad ways.

5.4.1. AI and Machine Learning Integration

AI and Machine learning are integrated in Big Data Engineering via platforms such as MLlib and TensorFlow on Hadoop, allowing organizations to build advanced models for predictive analytics, natural language processing, and image recognition (Luan et al., 2020). Therefore, integrating ML and AI is vital in formulating and forgoing insightful methods in big data analytics platforms.

5.4.2. Edge Computing

Edge computing is another emergent and trend concern in extensive data engineering involving data processing techniques, tools, and platforms. Edge computing is a concept undercrossing data processing closer to data sources without overdependence on centralized data sources and servers (Luan et al., 2020). Significantly, the proliferation of Internet of Things devices facilitates real-time data collection, requiring Big Data Engineers' adaptability, aligning with trends and future extensive data management through real-time collection, storage, processing, analysis, and transmission.

5.5. Final Thoughts

With Big Data Engineering, especially Hadoop and cloud platforms, organizations experience efficient, effective, and seamless data collection, storage, and analysis toward informed decision-making. Hadoop and cloud-based platforms, such as big data engineering platforms, have far-reaching and promising benefits, such as improved decision-making and operational efficiency, competitive advantage, and cost savings. Individuals, organizations, and government entities must fully address data integration, talent shortages, security, and infrastructure complexity challenges in navigating big data analytics and applications. Intergrading AI and ML, edge computing, and ethical data governance is pivotal in current and future considerable data engineering disciplines, supporting innovation and applications. Thus, there is a compelling need to conduct further research on AI, ML, and edge computing and their impacts on extensive data engineering as a discipline and practice.

Conflicts of Interest

In the paper exploring extensive data engineering on Hadoop and cloud-based platforms, the author declares no conflicts of interest concerning the research, authorship, and subsequent publication. Noteworthy, the study is based on publicly available academic literature, articles, and books on extensive data engineering on Hadoop and Cloud-based computing. The author further confirms that the findings and conclusions in this paper are bias-less, reflecting on independent analysis and interpretation, supporting the research's originality, integrity, and credibility.

References

[1] Apache Software Foundation, Hadoop, 2020. [Online]. Available: https://hadoop.apache.org/

[2] Amir Gandomi, and Murtaza Haider, “Beyond the Hype: Big Data Concepts, Methods, And Analytics,” *International Journal of Information Management*, vol. 35, no. 2, pp. 137-144, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[3] Ibrahim Abaker Targio Hashem et al., “The Role of Big Data in Smart City, *International Journal of Information Management*, vol. 35, no. 5, pp. 155-157, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[4] IBM, Big Data and Analytics, 2021. [Online]. Available: https://www.ibm.com/think/topics/big-data-analytics

[5] Avita Katal, Mohammad Wazid, and R. H. Goudar, “Big Data: Issues, Challenges, Tools, and Good Practices,” *2013 Sixth International Conference on Contemporary Computing (IC3)*, Noida, India, pp. 404-409, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[6] Doug Laney, “3D Data Management: Controlling Data Volume, Velocity, and Variety,” *META Group Research Note*, vol. 6, no. 70, 2001. [Google Scholar]

[7] Hui Luan et al., “Challenges and Future Directions of Big Data and Artificial Intelligence in Education,” *Frontiers in Psychology*, vol. 11, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[8] Mohamed Dhiaeddine Messaoudi, Bob-Antoine J. Menelas, and Hamid Mcheick, “Integration of Smart Cane with Social Media: Design of a New Step Counter Algorithm for Cane,” IoT, vol. 5, no. 1, pp. 168-186, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[9] Michael Minelli, Michele Chambers, and Ambiga Dhiraj, *Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses*, Wiley, 1st ed., 2012. [Google Scholar] [Publisher Link]

[10] Statista, Number of connected devices worldwide from 2019 to 2030, Statista, 2021. [Online]. Available: https://www.statista.com/statistics/1194682/iot-connected-devices-vertically/#:~:text=Number%20of%20IoT%20connected%20devices%20worldwide%202019%2D2033%2C%20by%20vertical&text=The%20consumer%20sector%20is%20anticipated,24%20billion%20connected%20devices%20worldwide.

[11] Tom White, *Hadoop: The Definitive Guide*, O'Reilly Media, 4th ed., 2015. [Google Scholar] [Publisher Link]

[12] Paul Zikopoulos, and Chris Eaton, *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, McGraw-Hill Osborne Media, 2011. [Google Scholar]