International Science and Technology Journal
المجلة الدولية للعلوم والتقنية
Volume 36 العدد
Part 1 المجلد
المجلة الدولية للعلوم والتقنية
International Science and Technology Journal
ISTJ

http://www.doi.org/10.62341/hsab1116

Received	**2024/11/25**	**تم استلام الورقة العلمية بتاريخ**
Accepted	**2024/12/29**	**تم قبول الورقة العلمية بتاريخ**
Published	**2024/12/31**	**تم نشر الورقة العلمية بتاريخ**

Advancements in Big Data Analytics Tools A Comparative Examination of Performance, Usability, and Applications

Hala Shaari
Faculty of Information Technologies- University of Tripoli
Tripoli, Libya
Email: h.shaari@uot.edu.ly
ORCID No: 0000-0003-0973-1398

Abstract

Numerous sectors, including science, business, social sciences, humanities, and finance, heavily rely on data-driven methodologies. Organizations place a high priority on extracting pertinent insights and patterns from massive datasets produced by sensor data, financial transactions, and human behavior. Big data is the term for large, intricate databases that are too big for standard techniques to manage. Through the application of big data analytics technology, these databases may yield insightful information that improves decision-making and fosters innovation and operational efficiency. It is essential to compare technologies like as Hadoop, Spark, and Flink to ascertain which one best suit certain data characteristics and processing needs. A thorough comparison of big data analytics technologies is included in this paper. It showcases the most recent developments in big data analytics, such as cloud-based solutions, machine learning integration, real-time processing, and the capacity to handle a wider range of data types. To fully utilize their data, businesses must comprehend these technologies and how they are developing. Additionally, this paper offers suggestions to consider when selecting the most effective big data analytics tools.

Keywords: Big Data, Big data Analytics, Big Data Analytics Technologies, Hadoop, Spark, Flink.

International Science and Technology Journal
المجلة الدولية للعلوم والتقنية
Volume 36 العدد
Part 1 المجلد
http://www.doi.org/10.62341/hsab1116

التطورات في أدوات تحليل البيانات الضخمة: دراسة مقارنة للأداء وقابلية الاستخدام والتطبيقات

هالة الشاعري
كلية تقنية المعلومات– جامعة طرابلس– ليبيا
h.shaari@uou.edu.ly
ORCID No: 0000-0003-0973-1398

تعتمد العديد من القطاعات، بما في ذلك العلوم والأعمال والعلوم الاجتماعية والإنسانية والمالية، بشكل كبير على منهجيات تعتمد على البيانات. تضع المنظمات أولوية عالية لاستخراج الأفكار والأنماط ذات الصلة من مجموعات البيانات الضخمة التي تنتجها بيانات الاستشعار والمعاملات المالية والسلوك البشري. البيانات الضخمة هو المصطلح المستخدم لوصف قواعد البيانات الضخمة المعقدة التي يصعب على التقنيات القياسية إدارتها. من خلال تطبيق تقنية تحليل البيانات الضخمة، قد تنتج قواعد البيانات هذه معلومات مفيدة تعمل على تحسين عملية اتخاذ القرار وتعزيز الابتكار والكفاءة التشغيلية. من الضروري مقارنة التقنيات مثل Hadoop وSpark وFlink لتحديد أي منها يناسب خصائص البيانات واحتياجات المعالجة بشكل أفضل. يتضمن هذا البحث مقارنة شاملة لتقنيات تحليل البيانات الضخمة. ويعرض أحدث التطورات في تحليلات البيانات الضخمة، مثل الحلول المستندة إلى السحابة، وتكامل التعلم الآلي، والمعالجة في الوقت الفعلي، والقدرة على التعامل مع مجموعة أوسع من أنواع البيانات. للاستفادة الكاملة من بياناتها، يجب على الشركات أن تفهم هذه التقنيات وكيفية تطورها. بالإضافة إلى ذلك، تقدم هذه الورقة اقتراحات يمكن أخذها في الاعتبار عند اختيار أدوات تحليل البيانات الضخمة الأكثر فعالية.

الكلمات المفتاحية: البيانات الضخمة، تحليلات البيانات الضخمة، تقنيات تحليل البيانات الضخمة، Hadoop، Spark، Flink.

Introduction

The phrase "big data" lacks clarity and definition. Other than expressing the idea of magnitude, the statement lacks precision and specificity. The question of how "big" is gigantic and "small" is tiny (Smith, 2013) depends on time, place, and circumstance, therefore

http://www.doi.org/10.62341/hsab1116

the adjective "big" is too general. The scope of "Big Data" is always evolving from an evolutionary perspective. Even if "big data" has become more and more popular, opinions on what it truly means are divided. Large dataset extraction, transformation, and loading (ETL) are commonly associated with the phrase "Big Data" by many professional data analysts. Three essential data attributes—volume, velocity, and variety, or 3Vs—are the foundation of a popular definition of big data. However, it falls short of fully capturing all aspects of big data.

In order to provide a comprehensive explanation of Big Data, we shall examine the term's historical development from its initial meaning to its contemporary meaning. Providing a historical overview of big data and proving that it is 3^2Vs or 9Vs instead of only 3Vs is the main objective of this section. These additional Big Data characteristics show why Big Data analytics (BDA) is really necessary. These expanded features, in our opinion, address certain key questions about the nature of big data analytics, such as which problems big data can help and which problems shouldn't be confused with big data analytics. This section examines these issues by summarizing some significant historical developments.

The study of extracting knowledge and insights from large, complex datasets is known as big data analytics. Big data presents businesses with previously unheard-of opportunities and challenges. It is characterized by volume, velocity, variety, veracity, and value. Companies that employ advanced analytical techniques can improve operations, obtain a competitive advantage, and identify emerging patterns to support decision-making.

The big data analytics technologies are thoroughly compared in this study. It highlights current advances in big data analytics, including cloud-based solutions, machine learning integration, real-time processing, and the ability to handle a larger variety of data types. To fully utilize their data, businesses must have a thorough understanding of these technologies and their background. How to choose the best big data analytics solutions is also covered in this paper.

While much research has been conducted to build tools and strategies for big data operations in data-driven sectors, there have been few contemporary studies of how these implementations are functioning. For example, (Mohamed et al., 2020) offered a complete evaluation of several open-source big data analytics technologies, including frameworks and applications. Similarly,

http://www.doi.org/10.62341/hsab1116

(Cui et al., 2020) examined how big data is employed in manufacturing, emphasizing applications and technology needs. They did, however, leave out certain important topics, such as batch and stream processing, fusion models, and a consideration of the big data ecosystem's strengths and drawbacks in smart manufacturing. (Nguyen et al., 2020) conducted a detailed examination of big data in a different industry sector (the oil and gas industry), investigating variables that influence technological adoption. Even yet, they did not thoroughly investigate the benefits and drawbacks of big data analytics for Industry 4.0.

(Adewusi et al., 2024) recently examined how Business Intelligence and Big Data may work together to help organizations achieve greatness. However, their investigation did not delve into the technical specifics that decision-makers would require to select the appropriate tools. (Ochuba et al., 2024) investigated techniques for leveraging big data and analytics to drive corporate success, concentrating on trends, obstacles, and best practices. However, they did not give practical methods or instructions for implementing these best practices, which might help organizations leverage big data and analytics to promote innovation and success in today's digital environment.

This review adopts an alternative methodology. For industrial big data initiatives, it deconstructs different data sources and lists their advantages, disadvantages, and examples. It also identifies workable ways that governments, corporations, and industries might address big data analytics issues. It also examines contemporary techniques, applications, and trends, going over their advantages and disadvantages. Lastly, the assessment provides suggestions for future lines of inquiry that may advance the area and enhance industrial performance.

The reminder of the paper will be arranged as follows: The "Background" section gives context for the historical definition and appraisal of dig data during the previous two decades, with an emphasis on its colorations in current technological breakthroughs. The "Big Data Analytics" part explains big data analytics principles and their relevance in today's decision-making paradigm by introducing the most well-known data analytics technologies. The "A Comparison of Big Data Analytics Technologies" section gives a detailed examination of the introduced data analytics technologies as well as assessment criteria. The "Discussion" section discusses

http://www.doi.org/10.62341/hsab1116

essential features lacking from the major toolkits, while "Conclusion" section 5 provides concluding thoughts and advice.

Background

1.1 Big Data Definition

There are several definitions for big data. Large data is defined by someone as data that is "big" instead of "easy-going," making it challenging to collect, store, manage, and analyze. The acronym "3V," which stands for volume, variety, and velocity, is frequently used. The following definitions are included in addition to many others that have been cited: *"Big Data can be defined as volumes of data available in varying degrees of complexity, generated at different velocities and varying degrees of ambiguity, that cannot be processed using traditional technologies, processing methods, algorithms, or any commercial off-the-shelf solutions."* (Krishnan, 2013). In addition, according to McKinsey (Manyika et al., 2011): "Big Data is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures and analytics to enable insights that unlock new sources of business value."

Furthermore, Three qualities stand out as distinctive to big data, as defined the authors of (Services, 2015): "Huge Volume of data: Rather than thousands or millions of rows, big data can be billions of rows and millions of columns." "Complexity of data types and structures: big data reflects the variety of new data sources, formats and structures, including digital traces being left on the web and other digital repositories for subsequent analysis." "Speed of new data creation and growth: big data can describe high velocity data, with rapid data ingestion and near real time analysis."

1.2 Different Attributes of Big Data Definitions

Gartner — 3Vs definition

Big Data has been characterized since 1997 by a number of traits, such as the Gartner 3Vs, which were introduced in 2004. The phrase was coined in 2001 by Douglas Laney (Laney, 2001) to describe the expansion of data in three dimensions: volume, velocity, and variety. Variety is the quantity of inconsistent and incompatible data forms and structures, velocity is the rate at which data is used for interaction, and volume is the amount of incoming and accumulating data stream. Although many people consider Laney's

http://www.doi.org/10.62341/hsab1116

3Vs formulation to be the "common" characteristics of big data, he did not define "big data" using these characteristics.

IBM — 4Vs definition

As part of the 4Vs of Big Data, IBM has expanded on Douglas Laney's 3Vs by adding a fourth "V" dimension. Volume, velocity, variety, and veracity are these dimensions. The reason for the inclusion, according to Zikopoulos et al. (Zikopoulos et al., 2012), was that clients were having issues with sources and quality in their Big Data initiatives.

Microsoft — 6Vs definition

Microsoft enhanced the 3Vs (variability, veracity, and visibility) attributes of Douglas Laney to 6Vs. Variability is the complexity of the data set; veracity is the trustworthiness of the data source; visibility is the emphasis of having a thorough understanding of the data for well-informed decision-making; volume is the scale of the data; velocity is the term used to describe streaming data analysis; variety is the variety of data forms. This addition is intended to improve data analysis and interpretation.

Additional Vs relating big data

A five-variable Big Data paradigm was proposed by Yuri Demchenko (Demchenko et al., 2014) in 2013. In the IBM 4Vs definition, he added the value dimension. More "Vs," up to 11, have been issued since Douglas Laney published 3Vs in 2001 (Elliott, 2013).

The fundamental goal of each of these definitions—including the 3Vs, 4Vs, 5Vs, and even 11Vs-is to convey a certain property of the data. Most definitions focus on data, but they fail to do an adequate job of articulating how Big Data relates to the BDA principles (Buyya et al., 2016). In order to understand the main idea, we must first define what data is.

1.3 TERMINOLOGY OF BIG DATA: 3 Vs - 3^2 Vs

By empowering decision-makers to make well-informed decisions based on projections, big data analytics seeks to deliver business intelligence (BI). Clarifying new big data qualities and demonstrating how they relate to three domain knowledge aspects—data domain (pattern detection), business intelligence domain (prediction), and statistical domain (assumption making) (Buyya et al., 2016) is essential to achieving this.

Once we have defined each of the three Vs features from three distinct perspectives, we can construct a diagram showing how they

http://www.doi.org/10.62341/hsab1116

are related. This is now the accepted definition of big data, as illustrated in Figure. 1, and it is comprehensive enough to cover all of its components.

Each circle in the diagram, as shown in Figure. 1, is associated with every characteristic of the three Vs in a single aspect. Moreover, the hierarchical diagram may be created by combining the three crucial characteristics of each circle. It summarizes the core notion of big data, according to (Services, 2015).

The 3^2 Vs (or 9Vs) convey the semantic meaning of Big Data (data, BI, and statistical correlations), while the original 3Vs data attributes represented the syntactic or logical meaning of the term. Three circles' characteristics combine to provide a higher level 3Vs for machines, which makes the 3^2 Vs a hierarchical model for complex problems or applications. The phrase "machine learning" is essential to big data analytics because, without a machine (computer), learning from massive data would not be feasible.

The aforementioned descriptions illustrate why handling Big Data involves such a high degree of complexity. Ambiguity, viscosity, and virality coexist with big data (Buyya et al., 2016). Lack of information, such descriptions or diagrams, is ambiguity. The data's latency time with respect to the event it describes is called viscosity. The pace at which information is shared inside a network—like Twitter—where tweets spread from their original source is known as virality. These elements aid in improving our comprehension of the dynamics of data sharing in networks.

2. Methodology

2.1 Big Data Analytics

The practice of examining vast volumes of data to provide historical, contemporary, and prospective statistics and insights that may be applied to enhance business decisions is known as big data analytics. Finding hidden patterns, relationships, and significant discoveries involves examining massive amounts of data. Because it can transform unprocessed data into meaningful insights, it is relevant. Massive datasets may be used by organizations to streamline processes, find new opportunities, and make better choices. Businesses may tailor their offerings and boost customer satisfaction and loyalty by using data to better understand consumer behavior.

http://www.doi.org/10.62341/hsab1116

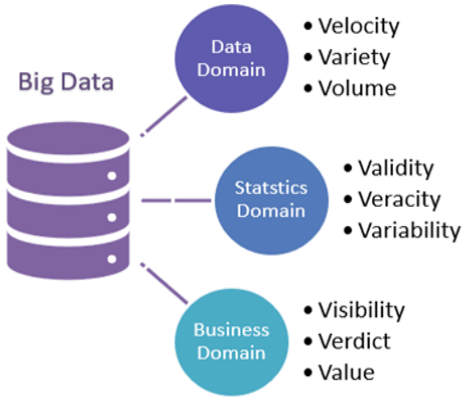

Figure 1. 3² Vs Characteristics diagrams.

Additionally, big data analytics helps to foster innovation by helping to recognize new trends and making it easier to develop innovative products and services. Ultimately, the effective utilization of big data analytics provides a competitive edge by enabling organizations to make better decisions faster than their competitors. Data science and analytics are closely connected fields of study that together make up the two main areas of big data analytics. With a focus on recent and historical statistics, data analytics is the study of gathering and analyzing data. By employing exploratory analytics to provide recommendations based on models built from historical and present data, data science, on the other hand, is futuristic in nature. While 3² Vs represent the philosophical meaning of big data, big data analytics illustrates its pragmatic meaning, according to (Buyya et al., 2016).

The Venn diagrams for big data and big data analytics in Figure 2 may be compared from a computational perspective. Arthur Samuel states that the original definition of machine learning was "the field of study that gives computers (or machines) the ability to learn without being explicitly programmed" (Samuel, 1959). Learning from data, pattern recognition, data science, data mining, text mining, and even business intelligence (BI) has all historically been used to refer to the same idea as machine learning (ML). Recent research has emphasized how crucial Big Data Analytics technologies are to the growth of technology across several sectors. Big data analytics and AI and machine learning must be combined in order to improve data processing capabilities, especially in predictive analytics and real-time decision-making (Udeh et al.,

http://www.doi.org/10.62341/hsab1116

2024). These technologies are become more and more crucial for evaluating and managing the massive volumes of data produced by Internet of Things devices, boosting output, and discovering pertinent information. Authors in (Stojanov & Daniel, 2024) also highlights the growing significance of automation in big data analytics, as it expedites data processing and increases accessibility to sophisticated insights for everybody. These tools are essential in today's data-driven industry.

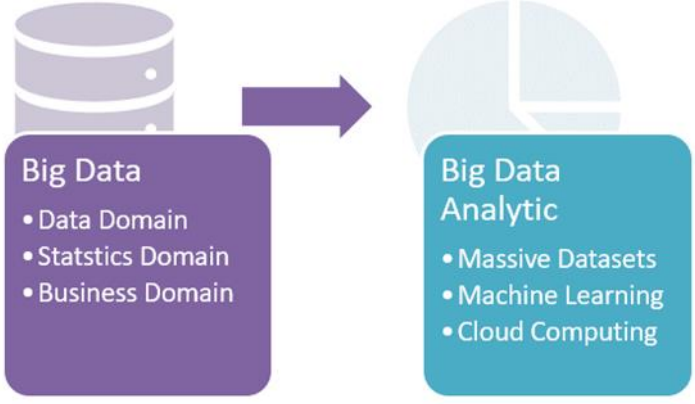

Figure 2. The semantic concept of Big Data has transformed into a practical knowledge of Big Data analytics.

2.2 Types of Big Data Analytics

Big data analytics involves different methods to uncover valuable insights from large sets of data. There are four main types of analytics, each designed for a specific purpose: descriptive, diagnostic, predictive, and prescriptive analytics. Descriptive analytics looks at past data to answer the question, "What happened?" It uses methods like data summaries, charts, and graphs to highlight trends, patterns, and important metrics. For example, a retailer might analyze past sales data to identify customer buying habits, providing a clear picture of the current situation and helping businesses make informed decisions based on what has already occurred (Hung et al., 2023; Wolniak, 2023).

Building on descriptive analytics, diagnostic analytics goes a step further to answer the question, "Why did this happen?" It delves deeper into the data, examining relationships and correlations to uncover the root causes of specific outcomes or problems. For instance, a bank might analyze demographic and economic data to

http://www.doi.org/10.62341/hsab1116

understand why there has been an increase in account closures (Hung et al., 2023).

Predictive analytics takes things further by focusing on the future, addressing the question, "What could happen?" It uses historical data along with statistical algorithms to identify patterns and trends, allowing businesses to anticipate future events. For example, predictive analytics might be used to forecast patient readmission risks based on clinical data or to predict customer churn by analyzing usage patterns (Hung et al., 2023; Sharma et al., 2022).

Prescriptive analytics extends predictive analytics by not only estimating future events but also recommending the right actions to attain optimal results. It assesses several prospective scenarios and their consequences, allowing companies to make well-informed, data-driven decisions. For example, energy businesses may utilize prescriptive analytics to detect the factors that influence oil and gas prices, allowing them to better manage risks.

Each sort of analytics contributes significantly to better decision-making by providing a clear and complete perspective of data to help strategic planning across sectors (Hung et al., 2023; Sharma et al., 2022).

2.3 Big Data Analytics' Challenges

Large volumes of digital data are now available to organizations, providing strategic options. However, significant ethical concerns regarding the sharing and use of this data have been raised by the quick development and application of big data. Businesses must behave responsibly by putting privacy and security first, obtaining consent before processing personal data, and treating stakeholders fairly while upholding the integrity of their data operations, according to the Institute of Big Ethics (Ethics, 2016).

Businesses also have a need to guard against the abuse of personal information and to be open and honest. According to the Institute of Big Ethics (Ethics, 2016), neglecting these obligations can have a negative effect on their income as well as their reputation and confidence with partners and consumers.

Big data analytics presents a number of obstacles for organizations seeking to fully realize the potential of their data. One of the most important challenges is data privacy and security, which focuses on protecting personal information. As businesses acquire huge volumes of data, the danger of breaches and illegal access increases significantly. Maintaining the integrity of sensitive information while assuring its security is a significant challenge. Organizations

http://www.doi.org/10.62341/hsab1116

must have strong security mechanisms to avoid data misuse, as well as comply with data privacy rules (Himeur et al., 2023).

Big data's diversity, which includes structured, semi-structured, and unstructured data, presents a significant challenge—managing unstructured data. Traditional data warehouses and relational databases often struggle to process unstructured data effectively, limiting their ability to support comprehensive analysis. Developing advanced tools and technologies to handle and extract insights from unstructured datasets remains a pressing need (Komalavalli & Laroiya, 2019).

Another critical concern is the ethical implications of big data analytics, which are receiving increasing attention. Bias in algorithmic decision-making, for example, can lead to discriminatory outcomes and reinforce social inequalities. To maintain public trust and ensure fairness, organizations must actively identify and address these biases in their analytics processes (Komalavalli & Laroiya, 2019). Additionally, ethical data use requires transparency in how data is collected, analyzed, and utilized—a challenging goal in complex organizational systems.

Integrating data from various sources is another significant challenge in big data analytics. Organizations often work with disparate systems, making it difficult to combine and analyze data seamlessly. This fragmentation can hinder the ability to gain a complete understanding of the data and lead to inefficiencies in decision-making (Kaur, 2023).

Furthermore, the increasing demand for real-time analytics has added pressure on systems to process high-velocity data streams. Traditional data processing methods are often inadequate for real-time analysis, creating bottlenecks in delivering timely insights that are essential for staying competitive in fast-paced markets (Kaur, 2023). By overcoming these challenges, organizations can unlock the full potential of big data analytics, enabling smarter decisions and improving operational efficiency.

2.4 Big Data Analytics Technologies

The big data ecosystem includes various technologies designed to handle the vast volume and complexity of modern data. Tools like Apache Hadoop and Spark are widely used for storing, processing, and analyzing large datasets. NoSQL databases such as MongoDB and Cassandra, along with the Hadoop Distributed File System (HDFS), efficiently manage diverse data types. For real-time data streams, technologies like Apache Kafka and Storm are key, while

http://www.doi.org/10.62341/hsab1116

Apache Pig and Hive provide user-friendly interfaces for extracting insights.

Visualization and analysis are supported by platforms like Tableau, Power BI, and Qlik, which offer interactive dashboards and reports. Cloud providers like Google Cloud Platform (GCP), Azure, and Amazon Web Services (AWS) also integrate big data capabilities, simplifying adoption for businesses. Ultimately, the choice of tools depends on specific business needs and data characteristics. This section highlights major tools like Apache Hadoop, Spark, and Flink, frequently discussed in the realm of big data processing. (Bajaber et al., 2016; Singh & Reddy, 2015).

Apache Hadoop

Apache Hadoop has become a widely used standard for accessing and sharing data and computational resources (Vavilapalli et al., 2013). As an open-source, scalable computing platform, it enables the distribution of computation tasks across multiple servers, even those with standard processors (White, 2012). Its core components include the Hadoop Distributed File System (HDFS) for data storage and the MapReduce engine for task execution.

Flexibility, scalability, cost-effectiveness, and reliability in managing and processing massive amounts of organized and unstructured facts are some of Hadoop's advantages. In order to balance workloads, resources, and data, it also provides job scheduling. Hadoop evolved into YARN, which architecture assigns numerous scheduling tasks to per-application components and separates the programming paradigm from the resource management infrastructure (Vavilapalli et al., 2013).

Apache Spark

A unified engine for distributed data processing is Apache Spark (White, 2012). Resilient Distributed Datasets, or RDDs, are an abstraction for data sharing that is added to a programming architecture that is comparable to MapReduce. SQL, streaming, machine learning, and graph processing are just a few of the processing tasks that Spark can now do that previously required the usage of separate engines. Spark was developed to address disk I/O limitations and enhance the performance of earlier systems (Zaharia et al., 2016). One of Spark's primary features is its in-memory computation capability. It removes the disk overhead barrier for iterative operations in the YARN by allowing data to be cached in memory.

http://www.doi.org/10.62341/hsab1116

Apache Flink

Derived from the Stratosphere project (Alexandrov et al., 2014), Apache Flink is an open-source framework for stream and batch processing intended for high-performance, distributed applications (Carbone et al., 2015). It is based on the notion that pipelined fault-tolerant data flows may be used to develop and execute a variety of data processing applications, including historical data processing, real-time analytics, continuous data pipelines, and iterative algorithms. Flink may run on top of YARN and HDFS or as a stand-alone framework. It increases runtime execution efficiency by utilizing in-memory storage. A distributed data flow runtime that uses pipelined streaming execution for batch and stream workloads, exactly-once state consistency via lightweight checkpointing, native iterative processing, and advanced window semantics that support out-of-order processing are among Flink's key innovations over prior Big Data technologies.

2.5 Case Studies in specific domains

Healthcare Domain

Big data analytics has revolutionized the healthcare industry by empowering institutions to handle enormous volumes of data for better patient care and decision-making. Advanced technologies are necessary to manage healthcare data successfully due to its growing complexity, which includes everything from medical imaging and electronic health records to real-time data from wearables (Ogundipe, 2024). At the vanguard of this change are tools like Apache Hadoop, Apache Spark, and Apache Flink, which offer strong options for processing, storing, and analyzing data.

Because of its effectiveness in processing and storing huge datasets, Apache Hadoop is widely utilized. It assists medical institutions in analyzing enormous volumes of patient data, spotting patterns of illness, and refining treatment plans. However, because of its in-memory processing capabilities, Apache Spark performs exceptionally well in terms of performance. Spark is perfect for real-time applications such as clinical trial data analysis and wearable device monitoring. For predictive analytics in personalized medicine and drug development, its machine learning libraries are extensively utilized. Apache Flink, on the other hand, provides unmatched real-time data processing, which makes it essential for vital applications like telemedicine assistance or critical care patient monitoring (Shaari et al., 2022). By acting swiftly on information,

http://www.doi.org/10.62341/hsab1116

these tools help healthcare professionals improve patient outcomes and operational efficiency.

Tools like Hadoop, Spark, and Flink are essential for maximizing the potential of big data as the healthcare industry becomes more and more data-driven. They streamline processes, cut expenses, and improve the quality of treatment. But enormous power also comes with responsibility, therefore it's imperative to protect data privacy and utilize it ethically. The healthcare industry can use these technologies to create a future where patient trust and enhanced treatment are given top priority by striking a balance between innovation and strong data control.

Higher Education Domain

By allowing institutions to handle large datasets for better decision-making, individualized learning, and operational efficiency, big data analytics is revolutionizing the higher education industry. Advanced technologies are needed to manage the amount, diversity, and velocity of data sources such student records, learning management systems (LMS), campus IoT devices, and research outputs. For organizing and analyzing big information, technologies like Apache Hadoop, Apache Spark, and Apache Flink offer strong solutions that spur innovation in research, education, and administration.

An essential technology for batch processing and scalable data storage in higher education is Apache Hadoop. Petabytes of structured and unstructured data, including library records, campus infrastructure data, and historical student performance measures, can be stored by universities using the Hadoop Distributed File System (HDFS). Hadoop can process this data using MapReduce to find trends, including identifying students who are likely to fail based on participation, attendance, and grade criteria. Institutions can implement targeted intervention programs by analyzing multi-year student data to identify key factors that influence academic achievement (Ikegwu et al., 2022). Additionally, Hadoop plays a crucial role in aligning education with job market needs by enabling the integration of various data, such as employment trends, to better match curricula with labor demands.

Adaptive learning platforms and advanced analytics in education can greatly benefit from Apache Spark's high-speed data processing and real-time application support. For instance, Spark can analyze student engagement data from learning management systems (LMS) like Moodle or Blackboard to create personalized learning paths. By examining clickstream data and the amount of time spent on various

http://www.doi.org/10.62341/hsab1116

resources, Spark can recommend specific study materials or tasks tailored to each student's progress (Stojanov & Daniel, 2024). This approach helps ensure that the learning experience is both relevant and individualized for each student.

In addition, Spark's capabilities extend to social network analysis through its GraphX library, which allows institutions to identify key influencers in academic forums or study collaboration patterns in student projects. Spark's machine learning libraries also play a critical role in predictive modeling tasks. For example, they can be used to predict trends in student enrollment or help optimize resource allocation during peak times, enabling institutions to manage their resources more efficiently (Stojanov & Daniel, 2024). These advanced analytical tools make Spark a powerful asset in educational settings, enhancing decision-making and student outcomes.

Apache Flink is increasingly being used in dynamic, time-sensitive situations, thanks to its expertise in real-time stream processing. For example, universities are adopting Flink to monitor classrooms equipped with Internet of Things (IoT) devices. By processing real-time data from smart sensors, Flink helps control energy efficiency, temperature, and lighting, creating a more sustainable and comfortable environment for students and staff (Stojanov & Daniel, 2024). Additionally, throughout the semester, Flink can analyze data from multiple sources, including real-time attendance records, learning management system (LMS) interactions, and campus resource usage. This enables educators to gain valuable insights that help them adjust strategies proactively, enhancing student success initiatives.

Flink's real-time event detection capabilities are also highly effective in campus security systems (Fang & Jiyuan, 2024) . It can quickly process data from access records or surveillance feeds to identify potential safety threats, allowing for rapid responses. By analyzing these data streams in real-time, Flink helps ensure that both the academic and security aspects of campus life run smoothly and efficiently, contributing to a safer and more responsive university environment.

Higher education organizations are using big data to improve student learning, expedite administrative procedures, and facilitate cutting-edge research by integrating tools like Hadoop, Spark, and Flink. Proactive solutions are made possible by these technologies, which provide a detailed awareness of operational and academic

http://www.doi.org/10.62341/hsab1116

difficulties. However, in order to address ethical and privacy concerns and make sure that innovation in education is in line with trust and accountability, its implementation must be supported by strong data governance processes.

3. Study Results

3.1 A Comparison of Big Data Analytics Technologies

Big data evaluation may be done in a variety of methods that sometimes overlap and cross disciplinary boundaries. Neural networks, data mining, machine learning, and pattern recognition are among the disciplines. An approach from one or more disciplines is used to extract relevant insights from huge data. This section explores the main characteristics of Apache Hadoop, Apache Spark, and Apache Flink, aiming to compare them across different challenges, as outlined in Table 1.

By examining these tools in various scenarios, we can clearly see the advantages and disadvantages of each. Hadoop has long been a strong solution for large-scale data processing, with its distributed file system (HDFS) offering a reliable framework for handling massive data volumes. However, as the demand for real-time data processing increases, Hadoop's batch-oriented MapReduce architecture and reliance on disk-based storage present significant drawbacks. Despite its robust ecosystem and effectiveness for batch processing, Hadoop struggles with high latency and complexity, making it less ideal for the real-time analytics needs of modern applications.

In contrast, Apache Spark addresses many of Hadoop's limitations through in-memory processing, which significantly reduces latency and enhances speed, particularly for real-time analytics and iterative machine learning tasks. Spark's ability to handle both batch and streaming data more seamlessly gives it a competitive edge in versatility. However, this advantage comes with trade-offs. Spark's reliance on in-memory processing requires a substantial amount of memory, which can complicate large-scale deployments and increase operational costs. While it excels in performance and flexibility, the resource demands can be a limiting factor in certain environments, especially when scaling for very large datasets.

Yet, Apache Spark has become a strong rival that has addressed many of Hadoop's drawbacks. Spark's in-memory processing power significantly reduces latency, increasing processing speed for batch and real-time applications. This performance benefit is particularly

http://www.doi.org/10.62341/hsab1116

helpful in situations where real-time analytics and iterative algorithms are required. Through its broad library, which includes MLlib for machine learning, GraphX for graph processing, and Spark SQL for structured data processing, Spark also enables a wide range of complex analytics. Unfortunately, this excellent performance comes with a high memory resource cost, which can increase system complexity and operating expenses. Additionally, despite Spark's powerful streaming capabilities, its micro-batch processing design introduces small delays that might not be ideal for applications requiring ultra-low latency.

For scenarios that require real-time data processing with ultra-low latency, Flink stands out as a more specialized option. Unlike Spark, which uses a micro-batch approach, Flink is designed for event-driven, continuous streaming. This makes Flink particularly well-suited for applications like real-time fraud detection and live monitoring, where the need for immediate data processing is crucial. Its architecture allows it to handle data streams continuously, providing faster, more efficient processing for time-sensitive tasks.

Table1. A Comparative Analysis of the Hadoop, Spark, and Flink Big Data Analytics tools.

Big Data Aspect	Apache Hadoop	Apache Spark	Apache Flink
Data Collection	Flume (high volume, unstructured), Sqoop (relational databases), Kafka (real-time)	Structured Streaming (real-time, batch), Kafka integration, Flume	Connectors (Kafka, Kinesis, etc.), Flume, SQL-like DDL for creating tables
Data Storage	HDFS (distributed file system), HBase (NoSQL)	In-memory storage, HDFS, External storage (object stores, databases)	State backends (RocksDB), External storage (object stores, databases)
Data Formats	Avro, Parquet, ORC, Text, RCFile	Avro, Parquet, JSON, CSV, Protobuf, binary formats	Avro, Parquet, JSON, CSV, Protobuf, binary formats
Data Transformation	MapReduce, Pig, Hive, Spark SQL integration	Spark SQL, DataFrames, Datasets, RDDs, MLlib pipelines	SQL-like API, DataStream API, CEP, Table API, SQL
Data Enrichment	Joins, aggregations, UDFs, window functions (limited)	Joins, aggregations, UDFs, window functions, machine learning pipelines, feature engineering	Joins, aggregations, UDFs, window functions, stateful computations, CEP

http://www.doi.org/10.62341/hsab1116

Data Analytics	Mahout, Pig, Hive, Spark integration	MLlib, GraphX, Spark SQL, R, Python integration	MLlib (emerging), SQL-like API, CEP, Flink ML
Data Science	R, Python integration (limited)	R, Python integration, MLlib, GraphX	R, Python integration, Flink ML, Gelly (graph processing)
Processing Model	Batch (primarily)	Batch, micro-batch, streaming	Streaming, micro-batch, batch
Fault Tolerance	HDFS replication, task retries	Lineage tracking, checkpointing, fault tolerance mechanisms	Checkpointing, state backends, fault tolerance mechanisms
Performance	High throughput for batch, latency varies	High throughput, low latency, in-memory processing	Very low latency, high throughput, optimized for streaming
Scalability	Horizontal scaling with additional nodes	Horizontal scaling with clusters, cloud-based options	Horizontal scaling with clusters, cloud-based options
Deployment Models	On-premise, cloud (Hadoop YARN)	On-premise, cloud (standalone, YARN, Kubernetes)	On-premise, cloud (standalone, YARN, Kubernetes, Flink SQL)
Strengths	Mature ecosystem, cost-effective for batch processing, reliable storage	Versatility, speed, machine learning capabilities, interactive analytics	Low latency, high throughput, stateful computations, real-time analytics
Weaknesses	Limited for real-time processing, complex setup	Resource intensive, complex API, evolving ecosystem	Steeper learning curve, less mature ecosystem than Spark
Use Cases	Batch processing, data warehousing, ETL	Batch, interactive analytics, machine learning, graph processing, real-time analytics	Real-time analytics, IoT, fraud detection, CEP
Additional Features	YARN for resource management, HBase for NoSQL	Structured Streaming for real-time processing, MLlib pipelines, Spark SQL for interactive queries	CEP, stateful functions, windowing, time-based joins

What distinguishes Flink from Hadoop and Spark is its ability to handle both batch and streaming data within the same environment, along with its robust event-time processing capabilities. This makes Flink particularly powerful for applications that require precise handling of time-sensitive data streams. However, despite its unique features, Flink is still relatively new to the big data landscape, with

http://www.doi.org/10.62341/hsab1116

a smaller user base and fewer third-party application integrations compared to Hadoop and Spark. As such, setting up and running Flink can be challenging, especially for teams that lack experience with stream processing technologies.

Despite these challenges, Flink is widely recognized as a leading tool for real-time analytics, and its importance in the future of big data is expected to grow as the demand for fast, actionable insights increases. When comparing these tools, Spark emerges as a versatile and well-rounded solution for a broad range of analytics tasks, while Hadoop continues to be essential for batch processing. However, when it comes to real-time stream processing, Flink is clearly the frontrunner, offering specialized capabilities that make it the ideal choice for applications that rely on continuous data streams.

Discussion

Hadoop: A Powerful but Outdated Workhorse

Hadoop, once considered the cornerstone of big data processing, offers significant benefits, particularly in terms of scalability and reliability. Its core storage component, HDFS, is well-suited for large-scale batch processing, capable of managing petabytes of data across distributed systems. This makes Hadoop a solid choice for handling massive datasets in traditional, batch-oriented environments. However, Hadoop faces limitations when it comes to real-time data processing, primarily due to its reliance on batch processing techniques like MapReduce and disk-based storage. These methods introduce considerable latency, making Hadoop less efficient for real-time analytics compared to newer frameworks like Spark and Flink, which utilize in-memory processing.

Furthermore, Hadoop's ecosystem, consisting of various components like Pig, Mahout, and Hive, is complex and requires significant administrative effort to manage. The steep learning curve and maintenance overhead associated with these tools can be daunting for organizations. As a result, many modern big data applications have moved away from Hadoop, driven by the growing demand for real-time data processing and low-latency solutions. While Hadoop remains valuable for batch processing, its shortcomings in handling real-time data and its complexity have made it less appealing in today’s fast-paced analytics environment.

Spark: Speed and Versatility with Resource Intensity

Spark’s primary feature is its use of in-memory processing, which significantly reduces the time spent reading and writing data to disk,

http://www.doi.org/10.62341/hsab1116

resulting in faster processing speeds. This makes Spark an excellent choice for applications that demand low latency, such as machine learning, iterative processing, and real-time data analytics. Spark is highly valued by data scientists and developers for its versatility, as it supports advanced analytics through libraries like MLlib, GraphX, and Spark SQL. These capabilities enable complex data processing tasks to be executed more efficiently, further enhancing Spark's appeal in the big data ecosystem.

However, Spark's reliance on in-memory computation comes with its own set of challenges. While it boosts processing speed, it also requires substantial memory, which can increase operational and maintenance costs, and complicate the management of large-scale clusters. Additionally, while Spark's real-time streaming capabilities are strong, they are based on a micro-batch model, meaning that it operates with relatively low latency compared to true real-time streaming systems like Flink. Although Spark excels in many use cases, its micro-batch architecture may not be suitable for applications that demand ultra-low latency, limiting its effectiveness for certain real-time processing scenarios.

Flink: The Future of Real-Time Stream Processing

When real-time data processing is required, Apache Flink performs exceptionally well. Flink's architecture is meant for low-latency streaming with event-time processing, in contrast to Spark's micro-batching, guaranteeing precise handling of time-sensitive data. Because of this, Flink is perfect for real-time analytics, live monitoring, and fraud detection. Flink is made more appealing by its ability to handle complex event processing and batch and stream processing inside a single paradigm. But compared to Hadoop and Spark, Flink is a more recent addition to the big data landscape, thus it has a smaller ecosystem and fewer integrations with other technologies. Despite its strength, Flink may also be difficult to set up and adjust, particularly for companies who have no prior experience with stream processing. Adoption may be hampered by this complexity, particularly for teams looking for a more straightforward answer. Notwithstanding these problems, Flink's advantages in real-time processing keep it in the forefront, particularly given the growing need for quick insights.

To this end, Spark, Hadoop, and Flink each offer distinct advantages and drawbacks, making them suited for different types of big data applications. Hadoop remains a strong choice for batch processing, handling large datasets effectively, but it falls short in real-time

http://www.doi.org/10.62341/hsab1116

scenarios due to its reliance on disk-based storage and batch processing methods. Spark, known for its speed and flexibility, works well for many analytics tasks but has limitations in ultra-low latency streaming and requires substantial memory, which can add to operational costs. Flink, on the other hand, shines in real-time stream processing, offering low-latency processing and the ability to handle both batch and stream data in one framework. However, its complexity and smaller ecosystem may present challenges for organizations looking for simpler solutions. Despite these challenges, Flink's advanced real-time capabilities make it a strong contender for the future of big data applications.

Conclusion

Big data refers to massive and complex datasets that are challenging to manage using traditional data processing methods. Due to its sheer size, speed, variety, accuracy, and potential value, big data necessitates specialized extraction, processing, and analysis techniques. The selection of an appropriate tool for managing and analyzing big data depends heavily on the specific requirements of the dataset in question. Among the top big data analytics solutions are Hadoop, Spark, and Flink. Each of these tools offers unique capabilities: Hadoop excels in batch processing and data storage for large datasets, but it struggles with real-time analytics due to its higher latency. Spark is versatile, handling both batch and streaming data efficiently and performing well in machine learning and interactive analytics, though it demands significant memory resources. Flink, on the other hand, is optimized for low-latency, stateful applications, making it ideal for real-time processing, but its complexity and lack of community support may pose challenges for some organizations.

When choosing the right tool for big data analytics, it's crucial to consider the specific needs of the application. Factors like the type of data (batch or stream), processing speed requirements, available infrastructure, and the learning curve for each platform all play a role in the decision-making process. For instance, if your application is batch-oriented and involves large datasets that don't require real-time processing, Hadoop could be the best fit due to its cost-effective and robust batch capabilities. However, for scenarios that demand flexibility and the ability to handle a mix of batch and streaming data, Spark offers the adaptability needed. Flink, meanwhile, is particularly suited for real-time applications where

http://www.doi.org/10.62341/hsab1116

low-latency processing is critical, such as fraud detection or live monitoring, but its complexity might make it more difficult to implement and scale.

In addition to these basic considerations, a more thorough comparison of these tools requires examining factors such as integration capabilities, performance benchmarks, and how well each solution meets the unique demands of different use cases. It's important to consider how well each tool integrates with other technologies in your infrastructure, its ability to scale, and how it handles various big data processing needs. For example, Hadoop's ecosystem supports a wide range of related tools like Pig, Hive, and Mahout, which may be beneficial for specific types of batch processing tasks. Spark's flexibility in handling both batch and streaming data gives it an edge in a variety of data environments, while Flink's ability to handle real-time streaming data with low latency makes it a strong choice for cutting-edge applications. Ultimately, the decision depends on the specific business goals and use case requirements.

As technology continues to evolve, the big data analytics landscape is shifting to incorporate new trends such as machine learning, real-time analytics, and advanced data management techniques. These developments are driving big data analytics technologies to become increasingly powerful and essential for businesses looking to innovate and stay ahead in competitive markets. With the rise of new tools and capabilities, the future of big data analytics promises to offer even more opportunities for organizations to gain insights, optimize processes, and make data-driven decisions with unprecedented speed and accuracy. As these technologies become more advanced, they will undoubtedly play an even more crucial role in shaping the success and innovation of businesses in the years to come.

References

Adewusi, A. O., Okoli, U. I., Adaga, E., Olorunsogo, T., Asuzu, O. F., & Daraojimba, D. O. (2024). Business intelligence in the era of big data: a review of analytical tools and competitive advantage. *Computer Science & IT Research Journal*, *5*(2), 415-431.

Alexandrov, A., Bergmann, R., Ewen, S., Freytag, J.-C., Hueske, F., Heise, A., Kao, O., Leich, M., Leser, U., & Markl, V. (2014). The

http://www.doi.org/10.62341/hsab1116

stratosphere platform for big data analytics. *The VLDB Journal*, *23*, 939-964.

Bajaber, F., Elshawi, R., Batarfi, O., Altalhi, A., Barnawi, A., & Sakr, S. (2016). Big data 2.0 processing systems: Taxonomy and open challenges. *Journal of Grid Computing*, *14*, 379-405.

Buyya, R., Calheiros, R. N., & Dastjerdi, A. V. (2016). *Big data: principles and paradigms*. Morgan Kaufmann.

Carbone, P., Katsifodimos, A., Ewen, S., Markl, V., Haridi, S., & Tzoumas, K. (2015). Apache flink: Stream and batch processing in a single engine. *The Bulletin of the Technical Committee on Data Engineering*, *38*(4).

Cui, Y., Kara, S., & Chan, K. C. (2020). Manufacturing big data ecosystem: A systematic literature review. *Robotics and computer-integrated Manufacturing*, *62*, 101861.

Demchenko, Y., De Laat, C., & Membrey, P. (2014). Defining architecture components of the Big Data Ecosystem. 2014 International conference on collaboration technologies and systems (CTS),

Elliott, T. (2013, 15 Aug). 7 Definitions of Big Data You Should Know About. https://timoelliott.com/blog/2013/07/7-definitions-of-big-data-you-should-know-about.html

Ethics, I. o. B. (2016). *Business Ethics and Big Data, Business Ethics Briefing, No 52*. 2016. Retrieved 28/12/2024 from https://www.ibe.org.uk/resource/business-ethics-and-big-data.html

Fang, H., & Jiyuan, R. (2024). Analyzing Big Data Professionals: Cultivating Holistic Skills Through University Education and Market Demands. *IEEE access*.

Himeur, Y., Elnour, M., Fadli, F., Meskin, N., Petri, I., Rezgui, Y., Bensaali, F., & Amira, A. (2023). AI-big data analytics for building automation and management systems: a survey, actual challenges and future perspectives. *Artificial Intelligence Review*, *56*(6), 4929-5021.

Hung, C.-C., Tu, M.-Y., Chien, T.-W., Lin, C.-Y., Chow, J. C., & Chou, W. (2023). The model of descriptive, diagnostic, predictive, and prescriptive analytics on 100 top-cited articles of nasopharyngeal carcinoma from 2013 to 2022: bibliometric analysis. *Medicine*, *102*(6), e32824.

Ikegwu, A. C., Nweke, H. F., Anikwe, C. V., Alo, U. R., & Okonkwo, O. R. (2022). Big data analytics for data-driven

http://www.doi.org/10.62341/hsab1116

industry: a review of data sources, tools, challenges, solutions, and research directions. *Cluster Computing*, *25*(5), 3343-3387.

Kaur, J. (2023). Streaming Data Analytics: Challenges and Opportunities. *International Journal of Applied Engineering & Technology*, *5*(S4), 10-16.

Komalavalli, C., & Laroiya, C. (2019). Challenges in big data analytics techniques: a survey. 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence),

Krishnan, K. (2013). *Data warehousing in the age of big data*. Newnes.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META group research note*, *6*(70), 1.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). Big data: The next frontier for innovation, competition, and productivity.

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

Nguyen, T., Gosine, R. G., & Warrian, P. (2020). A systematic review of big data analytics for oil and gas industry 4.0. *IEEE access*, *8*, 61183-61201.

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.

Ogundipe, D. O. (2024). The impact of big data on healthcare product development: A theoretical and analytical review. *International Medical Science Research Journal*, *4*(3), 341-360.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, *3*(3), 210-229.

Services, E. E. (2015). *Data science and big data analytics: discovering, analyzing, visualizing and presenting data*. John Wiley & Sons.

Shaari, H., Durmić, N., & Ahmed, N. (2022). Modern ABI platforms for healthcare data processing. Advanced Technologies, Systems, and Applications VI: Proceedings of the International Symposium on Innovative and Interdisciplinary Applications of Advanced Technologies (IAT) 2021,

http://www.doi.org/10.62341/hsab1116

Sharma, A. K., Sharma, D. M., Purohit, N., Rout, S. K., & Sharma, S. A. (2022). Analytics techniques: descriptive analytics, predictive analytics, and prescriptive analytics. *Decision intelligence analytics and the implementation of strategic business management*, 1-14.

Singh, D., & Reddy, C. K. (2015). A survey on platforms for big data analytics. *Journal of big data*, *2*, 1-20.

Smith, T. P. (2013). *How Big is Big and How Small is Small: The Sizes of Everything and Why*. OUP Oxford.

Stojanov, A., & Daniel, B. K. (2024). A decade of research into the application of big data and analytics in higher education: A systematic review of the literature. *Education and Information Technologies*, *29*(5), 5807-5831.

Udeh, C. A., Orieno, O. H., Daraojimba, O. D., Ndubuisi, N. L., & Oriekhoe, O. I. (2024). Big data analytics: a review of its transformative role in modern business intelligence. *Computer Science & IT Research Journal*, *5*(1), 219-236.

Vavilapalli, V. K., Murthy, A. C., Douglas, C., Agarwal, S., Konar, M., Evans, R., Graves, T., Lowe, J., Shah, H., & Seth, S. (2013). Apache hadoop yarn: Yet another resource negotiator. Proceedings of the 4th annual Symposium on Cloud Computing,

White, T. (2012). *Hadoop: The definitive guide*. " O'Reilly Media, Inc.".

Wolniak, R. (2023). The concept of descriptive analytics. *Zeszyty Naukowe. Organizacja i Zarządzanie/Politechnika Śląska*.

Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., & Franklin, M. J. (2016). Apache spark: a unified engine for big data processing. *Communications of the ACM*, *59*(11), 56-65.

Zikopoulos, P., Deroos, D., Parasuraman, K., Deutsch, T., Giles, J., & Corrigan, D. (2012). *Harness the power of big data The IBM big data platform*. McGraw Hill Professional.