applied sciences

MDPI

Article

The AI-Powered Evolution of Big Data

Yulia Kumar [1], Jose Marchena [1], Ardalan H. Awlla [2], J. Jenny Li [1] and Hemn Barzan Abdalla [3,*]

[1] Department of Computer Science and Technology, Kean University, Union, NJ 07083, USA; ykumar@kean.edu (Y.K.); marchenj@kean.edu (J.M.); juli@kean.edu (J.J.L.)
[2] Department of Computer Science, Cihan University Sulaimaniya, Sulaymaniyah City 46001, Iraq; ardalan.husain@sulicihan.edu.krd
[3] Department of Computer Science, Wenzhou-Kean University, Wenzhou 325060, China
* Correspondence: habdalla@kean.edu

Abstract: The rapid advancement of artificial intelligence (AI), coupled with the global rollout of 4G and 5G networks, has fundamentally transformed the Big Data landscape, redefining data management and analysis methodologies. The ability to manage and analyze such vast and varied datasets has exceeded the capacity of any individual or organization. This study introduces an enhanced framework that expands upon the traditional four Vs of Big Data—volume, velocity, volatility, and veracity—by incorporating six additional dimensions: value, validity, visualization, variability, volatility, and vulnerability. This comprehensive framework offers a novel and straightforward approach to understanding and addressing the complexities of Big Data in the AI era. This article further explores the use of 'Big D', an AI-driven, RAG-based Big Data analytical bot powered by the ChatGPT-4o model (ChatGPT version 4.0). This article's innovation represents a significant advance in the field, accelerating and deepening the extraction and analysis of insights from large-scale datasets. This will enable us to develop a more nuanced and comprehensive understanding of intricate data landscapes. In addition, we proposed a framework and analytical tools that contribute to the evolution of Big Data analytics, particularly in the context of AI-driven processes.

Keywords: AI-enhanced data analytics; RAG-based AI agents; Big D analytical bot; AI in Big Data management

Citation: Kumar, Y.; Marchena, J.; Awlla, A.H.; Li, J.J.; Abdalla, H.B. The AI-Powered Evolution of Big Data. *Appl. Sci.* **2024**, *14*, 10176. https://doi.org/10.3390/app142210176

Academic Editor: Luis Javier García Villalba

Received: 18 September 2024
Revised: 25 October 2024
Accepted: 1 November 2024
Published: 6 November 2024

1. Introduction

This research expands on the well-known 4 Vs framework of Big Data, suggested by Doug Laney [1] in 2001. This research presents the design and analysis of a novel 10-component model called "The spectrum of Vs". This is designed, among other uses, to evaluate Big Data, including AI-generated counterfactual data that have big implications for the design of relevant aggregator tools, especially as the MATLAB-centric models we depend upon increasingly lose relevance given the increasing complexity and scale of tasks. This change has been driven by the increasing popularity of Large Language Models (LLMs) and other AI technologies, which in turn mandates a revisit of conventional concepts from the literature to adapt them to the current evolving era of AI.

This research critically examines current Big Data tools and practices, assessing AI's ongoing and potential impacts within this rapidly evolving field. This study addresses three primary research questions:

RQ1: How does the proposed "spectrum of Vs" framework deepen the understanding of Big Data management in the context of AI-driven analytics?

RQ2: In what ways is AI already transforming Big Data analytics, and how can existing AI tools further contribute to this evolution?

RQ3: How can RAG-based AI agents, such as the proposed "Big D" analytical bot, enhance the efficiency and depth of insight extraction from vast and complex datasets?

Figure 1 illustrates the integration of Big Data and AI within this expanded framework, positioning AI at the core to symbolize its central role in processing and interpreting vast and complex datasets.

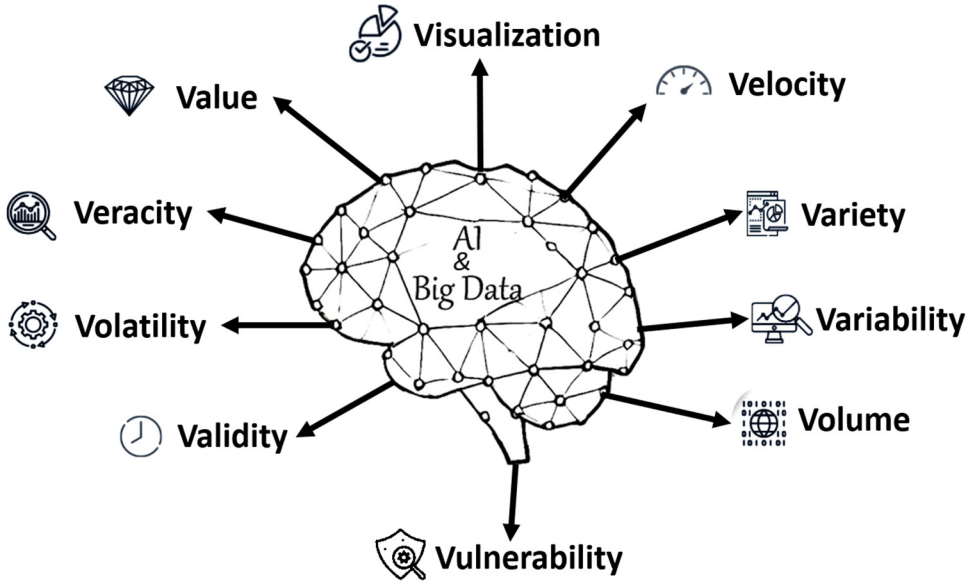


Figure 1. The spectrum of Vs.

The "spectrum of Vs" encapsulates ten critical dimensions of Big Data management and analytics:

- Value: the actionable insights derived from data that enhance decision-making.
- Veracity: the trustworthiness and quality of data are crucial for reliable outcomes.
- Volatility: the rate of change and unpredictability of data, challenging stability.
- Validity: the accuracy and relevance of data for specific purposes, ensuring utility.
- Vulnerability: the exposure of data to security risks, emphasizing the need for robust protection.
- Volume: the sheer scale of data being generated and processed, a hallmark of Big Data.
- Variability: the inconsistency of data over time complicates analysis.
- Variety: the diversity of data types and sources, enriching but complicating analytics.
- Velocity: the speed at which data are generated and need to be processed, demanding real-time solutions.
- Visualization: representing data insights in comprehensible formats, making complex data actionable.

The important role of visualization is emphasized as a bridge between complex Big Data and actionable insights derived from AI. Effective visualization is essential to translating complex data patterns into transparent formats, making the connection between Big Data and AI not only accessible but also actionable. This will allow stakeholders to make informed decisions by viewing the results of AI-driven analytics, thereby demonstrating the critical interaction between Big Data and AI in modern analytics.

Section 3 compares the "spectrum of Vs" framework to other recent studies in Big Data analytics. It discusses how the proposed framework expands upon traditional models by incorporating dimensions like validity, vulnerability, and visualization. Each study referenced is analyzed for its focus and contributions, showing how the "spectrum of Vs" provides a more comprehensive and adaptable approach to Big Data challenges across various industries. Section 4, or the Materials and Methods section, details the methodologies used to develop and validate the "spectrum of Vs" framework. It discusses the integration of AI tools like the ChatGPT-4o model and the RAG-based "Big D" analytical bot to enhance Big Data analytics. This section elaborates on the theoretical framework, including the conceptualization and operationalization of each of the ten Vs. It describes the system-

atic approach used to test the framework's efficacy in handling complex datasets. Section 5 examines the implications of adopting the 'spectrum of Vs' framework. It explores how the framework can influence Big Data management practices, highlighting the importance of robust data governance and the integration of advanced technologies like AI and machine learning. This section also addresses potential challenges such as data security, the ethical use of Big Data, and the need to adapt the framework to continuously keep up with technological advancements. Section 6 is entitled Conclusion and Future Work. Section 7 critically analyzes the limitations and broader implications of using the 'spectrum of Vs' framework within Big Data analytics. It explores the practical challenges of implementing the framework, such as the technological demands of managing high-volume and high-velocity data and the need for specialized skills to leverage AI and machine learning tools effectively. Additionally, ethical considerations are discussed, especially relating to data privacy and the potential for bias in AI algorithms. The section emphasizes the importance of robust data governance and ethical guidelines in mitigating these risks. Furthermore, it examines the societal implications of Big Data analytics, considering how they affect employment, data sovereignty, and access to information. This section aims to provide a balanced view by acknowledging constraints while highlighting the transformative potential of the 'spectrum of Vs' in terms of shaping future data practices.

2. Evolution of Big Data and Revolution in AI

The rapid and massive rise of artificial intelligence (AI) and the rapid expansion of digital connectivity have ushered in an era dominated by Big Data. The origins of this change go back to the technological infrastructure that developed in the mid-twentieth century, especially during the 1940s. During this period, the rise of relational database management systems (RDBMSs) and the rapid increase in information laid the foundation for a new data-driven world. The advent of groundbreaking developments such as the ENIAC and UNIVAC computers [2] marked the beginning of electronic computing and laid the foundation for the basic technologies that would shape future computing. In the 2000s, the idea of Big Data arose due to the emergence of Web 2.0 [3–5] and the widespread use of social media such as Facebook (now known as Meta) and Google. The Hadoop distributed file system (HDFS) stores data across large clusters, enabling applications to manage data using a distributed file system. It represents a revolution in distributed file systems and the MapReduce framework within Hadoop enables effective data storage and processing. Apache released Hadoop (Version 3.4.0 (https://hadoop.apache.org/)) as open-source software in 2006, after which it was utilized for web indexing by Yahoo and for data handling by Facebook. More importantly, Hadoop has paved the way for the further development of distributed analytics frameworks in the cloud, as well as AI-driven (artificial intelligence) analytics systems [6], which are important for managing and processing large datasets. The term "Big Data" was coined in this period [7], reflecting the growing acceptance of the volume and importance of data.

Significant progress was made in the 2010s with the advent of the iPhone, which completely transformed the field of mobile technology and enabled the swift expansion of the Internet of Things (IoT) [8]. This period witnessed the extensive use of data science and analytics, as well as the expansion of cloud computing. Collectively, these developments revolutionized the methods of data collection, storage, and analysis. The massive volumes of data generated globally have become central to AI and machine learning. Natural language processing has significantly advanced by integrating transformer models, such as BERT and the GPT family (e.g., GPT-2 and GPT-3) [9,10]. The broader applications of the transformer architecture make AI tools like ChatGPT into powerful 'Big Calculators' for use in data analysis, even though they are not directly derivatives of BERT or sentence transformers. This development also affected data ethics, privacy concerns, and information needs. The COVID-19 pandemic accelerated the adoption of digital technologies and pushed the boundaries of AI and ML development [11]. During the pandemic, Big Data technologies were adopted by the retail, healthcare, banking, and finance sectors, to name

a few. Amazon optimized supply chain logistics using AI, Walmart utilized AI capabilities as well [12,13], and IT departments in the healthcare sector improved their research methods and then started to create models in order to fast-forward development, as we can see [14–16]. The banking and finance sectors deploy this for completely digital transactions, leading to real-time risk avoidance [17]. Some well-known examples include Deutsche Bank, BBVA Fujitsu, and Hokuhoku Financial Group [17]. The COVID-19 pandemic accelerated the digitization of many industries, leading to rapid growth in the tech/IT sector, including in the area of AI. The rise of AI during pandemics is detailed in Appen's 2020 State of AI and Machine Learning Report, which can be found online [18,19].

Figure 2 provides a timeline that depicts the major technology and data science milestones throughout the evolution of data management—starting with early computing systems in the early 1940s, and ending with current Big Data and AI technologies. The following is a brief timeline that considers the advent of relational database management systems (RDBMSs) in the 1980s, the development of World Wide Web Internet and web 2.0 technologies at the end of the 1990s, the early 2000s interlude, and the introduction of AI-driven analytics over the last couple of years. These events were developmentally key "pivot points", and these developments provided a personal business environment that enabled the growth of different suppliers. The advent of RDBMSs [20] and early computers like UNIVAC and ENIAC was important [2]. This was followed by the invention of AI, the emergence of ethical entanglements, and the COVID-19 epidemic, which further sped up the adoption of technology. Figure 2 is forward-looking, showing the growth rate of innovation accelerating as we head into 2023 and beyond. For example, federated learning [21], blockchain [22], and quantum computing [23] are some of the new technologies that will have a significant impact on the landscape. For people who want to keep their business or organization on par with the competition in this data-centric world, only the growth of data generation, edge computing, data fabric, and regulations will define these evolutions as it relates to the integration of AI and ML futures; however, this is very important since it puts into perspective the different phases of technological advancement over the years, which have all led to where Big Data analytics stand today [24]. Every technological advance established the foundation for the next, thereby progressively improving data processing abilities. For instance, initial databases facilitated the ability to process vast amounts of structured data, which is necessary for Big Data operations. Web technologies were adapted, helping to enable the explosion of unstructured data as they outgrew what traditional databases could manage (paving the way for Big Data platforms like Hadoop). The new marriage of AI and cloud technology represents a monumental move towards Big Data analytics with more dynamic real-time processing capabilities. These developments underpin the "spectrum of Vs" framework supported by this paper, as well as its embrace of additional dimensions in Big Data management [25,26].

Federated learning is a decentralized form of machine learning, allowing multiple edge devices or servers to participate in the training process of a shared model and retain the data within the device. It can potentially transform how Big Data are analyzed, sufficiently reducing any privacy and security risks associated with having to localize sensitive data. It allows real-time analytics and model improvements across distributed networks while respecting data sovereignty. Nonetheless, due to the requirement to synchronize updates and handle communication overhead in huge-scale distributed systems, federated learning also brings scalability challenges. Furthermore, this has the drawback of a high dependence on the quality and diversity of local data, making federated learning susceptible to biased models. There are also regulatory concerns regarding data governance and compliance across jurisdictions, making the deployment of federated models challenging [27].

Quantum computing will revolutionize the speed at which extremely complex calculations can be performed by harnessing a property of quantum bits, called the superposition, that allows them to make multiple computations simultaneously. This will be optimized for solving Big Data analytics problems—involving things like optimization or

cryptography, or maybe just a lot of data processing—and it can perform some calculations related to that much faster than classical models. This could enable breakthroughs with regard to the hardest problems, which necessitate power-on-demand and real-time analytics, whether performing genomic analyses, climate modeling, or financial modeling [28]. Quantum computing is still in its infancy, and it currently suffers from many technological challenges—most importantly, high error rates and the low coherence times of qubits—which severely limit the practicality of existing quantum devices. Scalability is still problematic, as keeping qubits stable for long periods is extremely difficult. Furthermore, there are large security considerations, such as that quantum computing might break most of the current cryptographic protocols and hence would need new quantum-resistant cryptography [29].

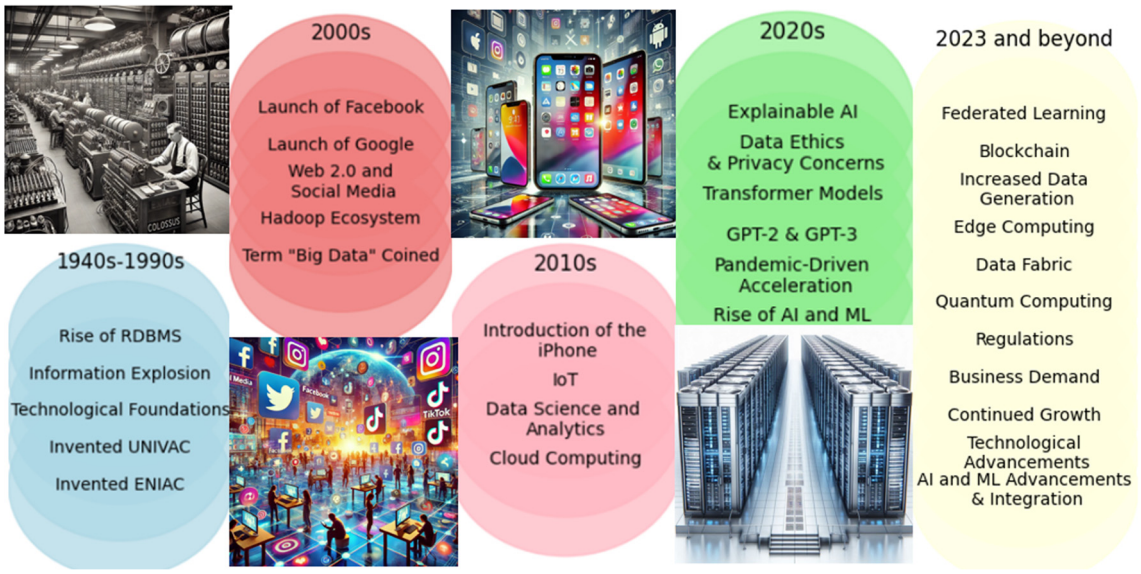


Figure 2. Big Data roadmap at a glance.

Overall, the Big Data field is undergoing significant transformations, driven by a variety of emerging trends. Organizations are increasingly relying on real-time analytics and data processing to gain immediate insights, supported by technologies like stream processing and edge computing, which enhance speed and efficiency, particularly for IoT data analytics. The seamless integration of AI and machine learning is propelling AI-driven analytics forward, with advancements in natural language processing (NLP) and explainable AI (XAI) making complex models more understandable. Automation through augmented data management, augmented analytics, and AutoML is leading to hyper-automation in data workflows, while advanced analytics and data storytelling are providing deeper insights and enabling the more effective communication of findings. Data democratization is expanding access to information across organizations, and there is a strong focus on ethical data collection and usage in order to address AI biases. Ensuring data privacy and governance remains crucial, and heightened efforts are being made to comply with regulations and protect sensitive information. Additionally, the shift towards cloud-native data architectures, hybrid and multi-cloud environments, and serverless computing is providing scalable and flexible solutions. Innovations such as data lakes, data warehouses, data lakehouses, data mesh, decentralized data management, and data fabric architectures are fostering more agile and scalable data infrastructures. Emerging technologies like graph databases, blockchain for secure data management, and quantum computing offer new capabilities. The industry also embraces open-source and community-driven innovations, prioritizing sustainability and green computing, and leveraging Big Data for social good.

To effectively navigate these trends, a comprehensive suite of tools is essential. Foundational platforms like HDFS and Hadoop offer robust distributed storage and process-

ing frameworks, while MapReduce, Hive, and Pig simplify data querying and manipulation. With its components Spark SQL, Spark Streaming, and GraphX, Apache Spark provides powerful in-memory processing capabilities. NoSQL databases, such as HBase and Couchbase, along with graph databases like Neo4j, are ideal for handling unstructured and relationship-intensive data. Real-time data processing is facilitated by tools like Apache Beam, Flume, Samza, Ignite, and Pulsar. Data governance is managed through Apache Atlas, ensuring compliance and effective metadata management. For visualization and exploration, Apache Superset and Zeppelin offer interactive analytics interfaces. In the realm of machine learning and AI, frameworks like Apache MXNet, TensorFlow, PyTorch, Keras, and Scikit-learn are vital for developing sophisticated models. Data scientists rely on libraries such as Pandas, NumPy, Matplotlib, Seaborn, and Bokeh for data analysis and visualization. Business intelligence platforms like Tableau, Power BI, SAS, IBM Watson, SAP HANA, Oracle Business Intelligence, Google BigQuery, Amazon Redshift, Microsoft Azure Synapse Analytics, and Snowflake provide comprehensive solutions for enterprise analytics. Data preparation and integration are streamlined with tools like Datameer, Trifacta, Talend, Informatica, Alteryx, and RapidMiner. Together, these tools support the entire Big Data lifecycle, enabling organizations to maximize the value of their data assets.

3. Related Work

The related work is presented in Table 1, where it is compared to the proposed spectrum of Vs framework and Big D app (v1) in this study.

Table 1. Comparison of related work.

Research Work	Year	Focus/Contribution	Comparison with the Proposed Study	Ref.
Aminul	2024	The impact of Big Data analytics on digital marketing strategies and customer engagement.	Both works explore the practical applications of Big Data. The spectrum of Vs goes beyond digital marketing.	[30]
Sargiotis	2024	The incorporation of AI and Big Data into virtual infrastructures, focusing on educational environments.	Both examine AI's role in reshaping Big Data applications. The scope of the spectrum of Vs spans beyond education.	[31]
Tosi et al.	2024	A comprehensive summary of 15 years of Big Data research, focusing on patterns, challenges, and prospects.	Both offer forward-looking perspectives on Big Data and its evolution, but this study introduces new dimensions to the framework; Big D focuses on enhancing real-time data analysis.	[32]
Arachchige	2023	The evaluation of LLMs' accuracy in medical settings, highlighting Big Data's role in enhancing precision.	Both studies emphasize the role of Big Data in enhancing AI precision. The spectrum of Vs has applications beyond medical applications.	[33]
Cui et al.	2023	Innovative methods for encoding text using LLMs.	Both explore advanced AI techniques. The spectrum of Vs goes beyond text encoding.	[34]
DATA et al.	2023	Advanced AI models and large-scale remote sensing data are utilized to improve Earth observation.	Both studies apply AI to enhance data analysis. The spectrum of Vs offers a broader application framework beyond Earth observation.	[35]
Dida et al.	2023	The integration of ChatGPT with Big Data for improved text-to-speech processes.	Both explore the integration of AI with Big Data. The spectrum of Vs offers a broader framework beyond text-to-speech.	[36]
Fahim	2023	The influence of fast engineering on AI language models, focusing on extensive training datasets.	Both works highlight the importance of data quality in AI training. The spectrum of Vs incorporates it into its framework.	[37]
Rashid et al.	2023	The impact of Big Data analytics and AI on green supply chain practices, emphasizing sustainability.	Both studies emphasize sustainability within Big Data and AI. Big D's real-time processing and analytics provide dynamic support for sustainability initiatives.	[38]
Sardi et al.	2023	Correlation between Big Data and performance assessment research.	Both explore performance assessment within Big Data. Big D's real-time processing enables continuous performance monitoring.	[39]
Zhang et al.	2023	Analyzing meaning using LLMs.	Both explore the application of advanced AI techniques. Big D applies these techniques practically in real-time Big Data analytics.	[40]

Table 1. *Cont.*

Research Work	Year	Focus/Contribution	Comparison with the Proposed Study	Ref.
Ram and Verma	2022	Comparative examination of AI-based chatbots, emphasizing Big Data's role in enhancing performance.	Both studies examine AI-driven performance improvements. The spectrum of Vs offers a broader framework for evaluating these improvements across different sectors. Ram and Verma's work is chatbot-specific.	[41]
Sundu et al.	2022	Data-driven innovation, with a focus on AI and Big Data in organizational innovation.	Both works explore the intersection of AI and Big Data in fostering innovation. The spectrum of Vs provides a structured framework to evaluate and enhance these innovations systematically.	[42]
Bormida	2021	Examine Big Data progression, its dimensions, impacts, and research challenges.	Both explore the core dimensions of Big Data, but this study also addresses new challenges in the AI era. Big D's real-time processing goes beyond static analysis.	[43]
Raubenheimer	2021	Examine Big Data progression, its dimensions, impacts, and research challenges.	Both explore the core dimensions of Big Data, but the spectrum of Vs expands them further.	[44]
Raban and Gordon	2020	The bibliometric analysis of research developments in data science and Big Data.	Both aim to map out the research landscape, but Big D's capability for real-time data processing and AI-driven analysis contrasts with Raban and Gordon's bibliometric focus.	[45]
Chae	2019	A methodology for examining digital innovation ecosystems, with a focus on Big Data.	Both explore the role of Big Data in digital ecosystems, but previous work has no AI aspect and no real-time application.	[46]
Nadal et al.	2019	An ontology for regulating the development of Big Data ecosystems, focusing on integration and adaptation.	Both focus on the structure and management of Big Data ecosystems, but previous work has no AI aspect and no real-time application.	[47]
Bonner et al.	2017	Technological advancements in Big Data as a scientific and research field.	Both analyze the evolution of Big Data and leverage technological advancements in data analysis.	[48]
Gu et al.	2017	The visualization of Big Data research's knowledge structure and evolution in healthcare informatics.	Both explore the visualization of Big Data. The spectrum of Vs adds AI to these visualizations.	[49]
Salminen et al.	2017	The potential of digitization and Big Data in responsible corporate co-evolution.	Both consider the ethical implications of Big Data, but Big D's real-time data processing offers practical solutions.	[50]
Halevi and Moed	2012	The historical progression of Big Data as a research and scientific field.	Both analyze the evolution of Big Data, but previous work has no AI aspects and no real-time application.	[51]

As can be seen from Table 1, some of the most recent studies do emphasize AI's impact on Big Data analysis, but most of them have a narrower scope than the spectrum of Vs framework.

4. Materials and Methods

This study introduces an advanced Big Data framework known as the "spectrum of Vs", which expands upon the traditional four Vs of Big Data [52,53] by incorporating six additional dimensions. These additional components enrich the framework's ability to address the intricate interconnections between Big Data and artificial intelligence (AI), offering a more nuanced understanding of the challenges and opportunities the modern data landscape presents. Each of the ten components of the spectrum of Vs is analyzed in relation to AI, illustrating how these dimensions contribute to the effective utilization of Big Data in various AI applications. In addition to the theoretical framework, this research proposes the development of an AI-driven application named "Big D", a Big Data analysis bot built using the retrieval-augmented generation (RAG) architecture [54]. This application leverages the capabilities of the ChatGPT-4o mini model [55] and uses the OpenAI Assistant API v2 as its backend. Big D is designed to be highly knowledgeable about the spectrum of Vs framework and has this methodology embedded in its operational memory, allowing it to assist users in various analytical tasks. The architecture and functionalities of Big D will be further elaborated later in this paper. Figure 3 represents a global workflow of the study's methodology.

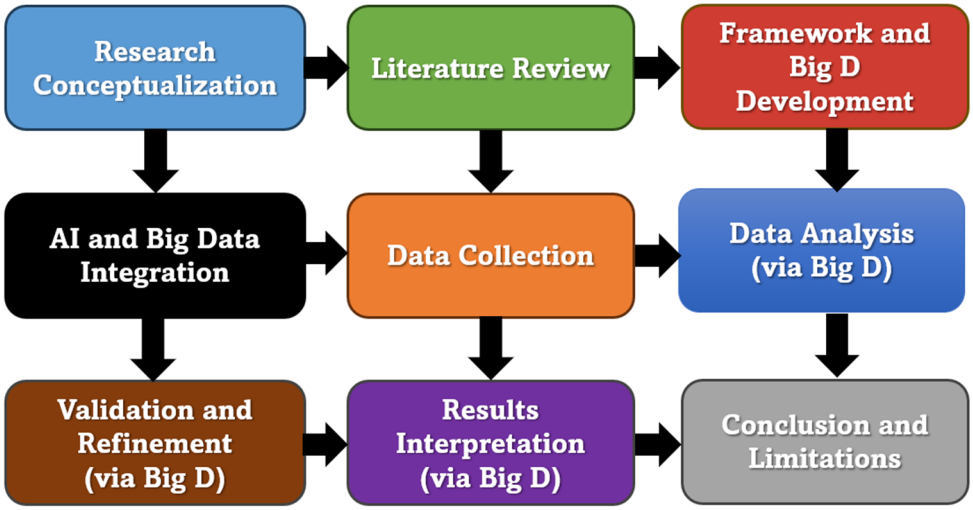


Figure 3. Global methodology workflow.

4.1. Spectrum of Vs

The concept of Big Data is more than just big datasets. A large infrastructure is also required to process, analyze, and store these large data. This infrastructure includes advanced computing platforms, dedicated storage solutions, data centers, and strong security measures. This structure aims to increase the speed and efficiency of data processing, improve decision-making results, and optimize the discovery of possible views. In modern business environments, the strategic application of Big Data is the most important success factor that guides organizations to make informed decisions. Research has shown that organizations that lead in terms of analytics improve their operational performance and lead the market by strategically aligning people, tools, and data capabilities [56]. These organizations are twice as likely to be industry leaders, three times stronger in terms of decision-making, and five times faster than their competitors [57].

As articulated in the abstract, this research extends the traditional 4 Vs Big Data framework—volume, velocity, variety, and veracity [51,52]—by introducing a more comprehensive "spectrum of Vs" framework that adds six additional dimensions. The transition from the 4 Vs to the spectrum of Vs Big Data framework (BDF) is illustrated in Figure 4.

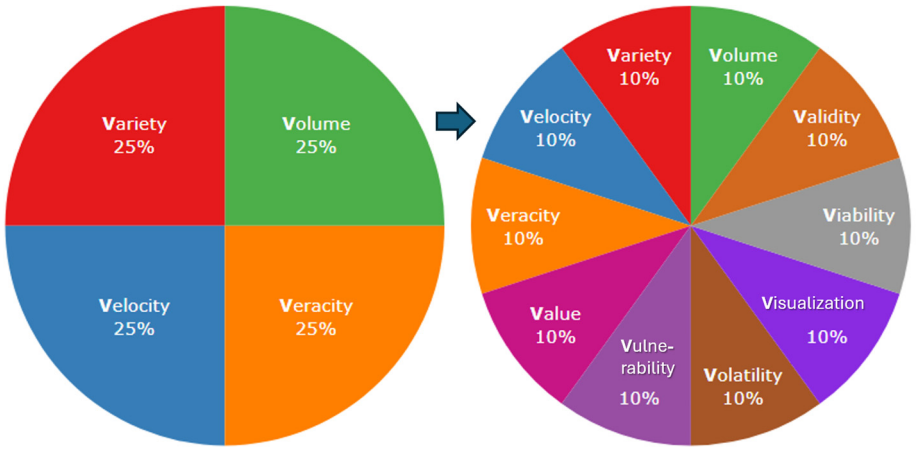


Figure 4. Big Data framework transformation (from 4 Vs to the spectrum of Vs).

As can be seen in Figure 4, the original four components—volume, velocity, variety, and veracity—are still important and present in the framework. Still, due to the evolution of the Big Data framework in the era of AI, several more dimensions, also known as the Vs of Big Data, were added.

This study explores a Big Data framework that influences the nature of artificial intelligence and leading models in the market. Automated data analysis, management, and visualization systems highlight the importance of data accuracy and value. New techniques can solve these problems. Advanced models such as ChatGPT-4 can generate accurate video presentations using complex datasets and code-free AI [58]. Students can understand and analyze text, images, and sounds in various formats, thus gaining valuable knowledge. Some examples, such as Google Gemini Advanced, offer special features such as video analysis and the creation of short summaries. This changes the boundaries of what artificial intelligence can achieve regarding data processing and content creation. With the development of artificial intelligence, the creation of text, images, videos, and music is becoming increasingly popular. This will significantly impact the management and use of data in many industries. In addition, the field of Big Data continues to grow. As can be seen from Figure 5, the phrase 'low-scope data' refers to smaller, more manageable datasets often used for localized decision-making and operational efficiency.

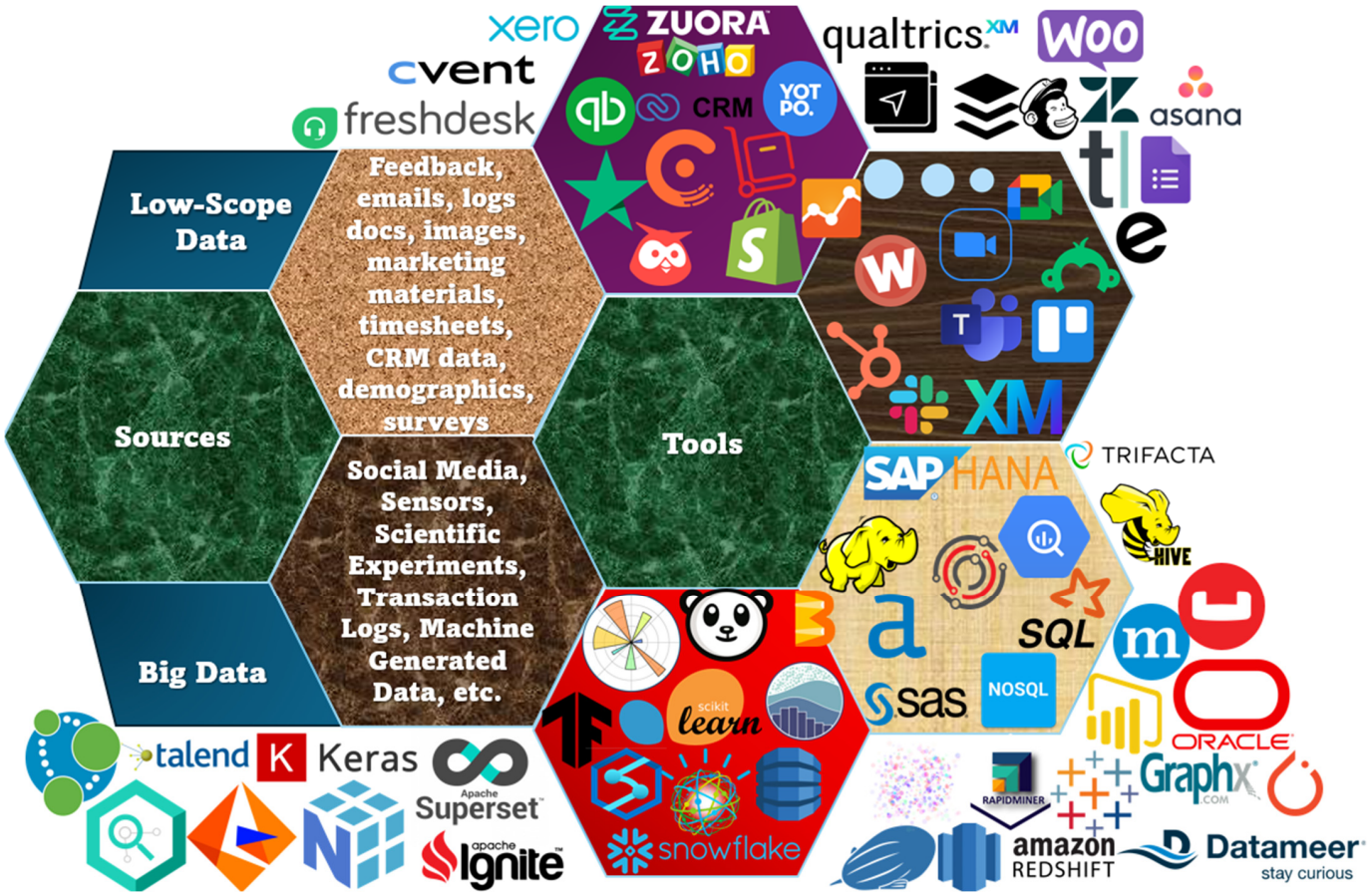


Figure 5. Low-scope data vs. Big Data: sources, tools, and outcomes.

As illustrated in Figure 5, the comparison between low-scope and big data highlights the diverse sources, tools, and outcomes of managing these datasets. Low-scope data typically include feedback, emails, CRM data, and demographic surveys, and these are processed using tools like Freshdesk, Zoho CRM, and Qualtrics for targeted smaller-scale insights. In contrast, Big Data encompasses vast sources such as social media, sensors, and machine-generated data, and these are analyzed using powerful tools like Hadoop, SAP HANA, Keras, and Apache Ignite to derive comprehensive, large-scale insights that have significant business impacts and drive strategic decision-making.

The data type is typically obtained from internal sources, such as feedback forms, emails, documents, CRM systems, and timesheets. Tools commonly used for processing

and analyzing low-scope data include platforms such as Google Sheets, Excel, and Asana, which are designed to handle small-scale operations. The outcomes derived from low-scope data are frequently restricted to internal reporting, enhancing the execution of day-to-day tasks and furnishing insights about a restricted range of applications. On the other hand, Big Data encompasses vast, complex datasets generated from diverse sources such as social media, sensors, scientific experiments, transaction logs, and machine-generated data. Many other data sources have accumulated large amounts of data over time. These datasets require advanced, distributed processing systems like Hadoop, Spark, and NoSQL databases. The processing tools used for Big Data are designed to be highly scalable and enable real-time analysis, machine learning, and AI-driven insights.

The sheer volume and velocity of Big Data necessitate the use of advanced machine learning algorithms, deep learning models, and distributed computing frameworks to extract actionable insights that can revolutionize industries. Some examples of Big Data technologies include hashtag tracking, crisis management, and influencer marketing ROI analysis on social media. The information gathered from sensors, like those used in agriculture for moisture monitoring, is analyzed to achieve breakthroughs in healthcare and engineering. Genomics, material science, and medical imaging data are analyzed to achieve healthcare and engineering breakthroughs.

Unsurprisingly, Figure 5 features the Hadoop ecosystem, represented by tools like HDFS for storage, MapReduce for processing, Hive for data warehousing, and Spark and its Spark SQL and GraphX for distributed data processing, as well as highly used NoSQL databases, including HBase for column-based storage and Neo4j for graph-based data.

Table 2 highlights the differences between Big Data and traditional data processing.

Table 2. Big Data vs. traditional data processing.

Parameter	Traditional Data Processing	Big Data Processing
Data scale	Small (like in MB)	Big (like in GB, TB, or PB)
Data type	Just one type of data (mostly structured data)	Many types of data (like structured, semi-structured, and nonstructured data)
Relationship between mode and data	The ways things happen are decided before considering the data.	The ways things happen are decided after analyzing the data. These methods change as more data become available.
Object to be processed	It is like there is just one type of fish in the pond.	It is like there are lots of fish in the ocean. We use some of these fish to determine if there are other types of fish.
Processing tool	One tool fit everything.	There is no one tool that works for everything.

The rest of this section provides a deep dive into the Vs of the spectrum of Vs.

Data quality is an important concern in today's world of Big Data, being significant beyond the spectrum of Vs. Better-quality data enhance the accuracy of insights from AI systems, thereby fostering the confidence required for improved decision-making [59]. Converting Big Data into actionable insights constitutes its true purpose, propelling strategic initiatives and innovation. Furthermore, it is also important to secure confidential data effectively, and the need for security is highly applicable in the case of healthcare and banking sectors [60]. Now, any organization can tackle Big Data by processing them with grace, meaning, and security, whether or not they were harvested by equally honorable means that respected their value.

4.2. Traditional 4 Vs of Big Data

4.2.1. Volume

Volume in the value spectrum relates to the efficient management of large amounts of data using AI-enhanced solutions. Big Data management is a significant obstacle because of the amount of information, which can overwhelm storage and processing systems as their capacity to store and process data is often limited. Hadoop and its distributed file system (HDFS) are essential for efficiently managing Big Data. As shown in Figure 6, to deal with the scaling problem, the model distributes the data across multiple nodes and promotes parallel processing.

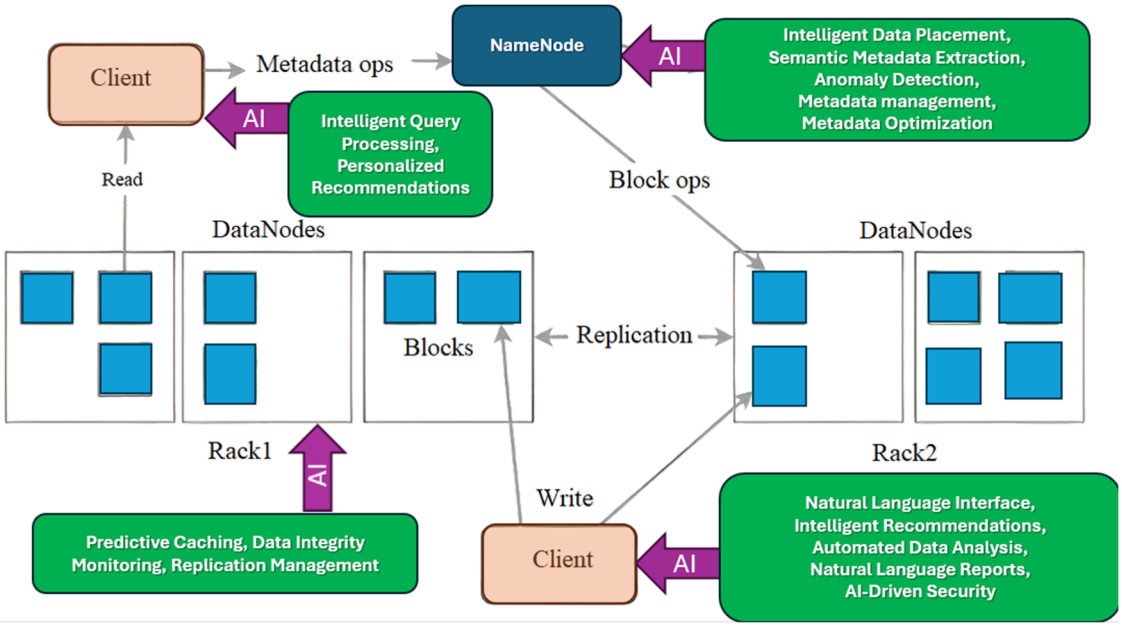


Figure 6. Potential AI/LLM integration points within the HDFS architecture.

This architecture enhances the overall storage capacity by adding more nodes while dramatically reducing processing time, even when the data surpasses the capacity of a single computer. Hadoop's distributed structure makes it highly suitable for efficiently managing big volumes of data, which is crucial in areas such as social media, finance, genomics, and climate science, where a massive amount of data are being generated.

Nevertheless, even advanced frameworks such as Hadoop face limitations with the exponential increase in data volumes. Apache Flink and Google BigQuery are advanced technologies that establish higher benchmarks for processing Big Data. They provide real-time capabilities and cloud-based analytics. Table 3 presents a comparative analysis, demonstrating Flink's improved real-time processing capabilities and the scalability and integration advantages of BigQuery with other Google Cloud services. This analysis highlights how Flink and BigQuery can be viable alternatives to Hadoop.

With the increasing volume of Big Data, traditional processing frameworks like Hadoop must adapt and meet new needs. By incorporating artificial intelligence (AI) and language models (LLMs) into these systems, we may significantly improve their capacities, allowing for more effective data organization, quicker processing speeds, and superior output quality. The integration of conventional Big Data frameworks with state-of-the-art AI technologies signifies the emerging frontier with regard to managing extensive data volumes, guaranteeing that enterprises may persistently derive significant insights and stimulate innovation in a data-centric world. Figure 6 depicts the integration of AI into the HDFS design, which improves data processing at different levels. As can be seen from the figure, AI can enhance how client queries are interpreted and processed, enabling more efficient data retrieval. It can provide recommendations based on user interactions and

past data access patterns, optimize where data are stored to improve access time and efficiency, and automate retrieval and scheduling metadata to make them easier to manage and improve.

Table 3. Comparison of key features of Apache Hadoop, Spark Flink, and Big Query.

Feature	Hadoop	Flink	BigQuery
Data processing	Batch processing	Single-runtime batch and stream processing	Managed data warehouse and analytics
Streaming engine	MapReduce (batch-oriented)	True streaming engine	Batch and streaming processing capabilities
Performance	Slower batch processing	Excellent performance, native closed-loop iteration operators	High performance with managed infrastructure
Memory management	Configurable (dynamic or static)	Automatic memory management	Managed by Google Cloud Platform
Fault tolerance	Highly fault-tolerant	Lightweight fault tolerance mechanism	Automatic replication and backups
Scalability	Highly scalable to commodity hardware	Highly scalable	Automatically scalable
Iterative processing	Not supported	Iterates data	Supported
Language support	Primarily Java, others	Java, Scala, Python, R,	SQL, JavaScript, Python, etc.
Optimization	Manual optimization	Independent optimizer	Managed by Google Cloud Platform
Latency	High latency	Low latency with low configuration effort	Low latency
Visualization	Zoomdata integration with Apache Zeppelin	Web interface for job execution and visualization	Integration with Data Studio and other visualization tools
Recovery	Highly fault-tolerant	Checkpoint mechanism for recovery	Automatic backup and recovery
Security	Kerberos authentication	Hadoop/Kerberos User authentication	IAM- and ACL-based security policies
Cost	Inexpensive hardware	Mid- to high-level hardware	Pay-as-you-go pricing model
Compatibility	Compatible with Spark, JDBC, and ODBC	Fully compatible with Hadoop, the Hadoop compatibility package	Integration with Google Cloud Platform services
Abstraction	No abstraction	Uses Dataset and DataStreams	SQL-based abstraction
Ease of use	Requires hand-coding	High-level operators	SQL querying
Interactive mode	N/A	Integrated interactive Scala Shell	Command-line and web-based GUI
Real-time analysis	Not suitable	Mainly used for real-time data analysis	Real-time processing capability
Scheduler	Pluggable scheduler	Use YARN Scheduler or its own Scheduler	Managed by Google Cloud Platform
SQL support	Apache Hive for SQL queries	Table API, support for the SQL interface	Native SQL support and optimizations
Caching	Cannot cache data	Can cache data in memory	Managed caching mechanisms
Hardware requirements	Runs well on commodity hardware	Mid- to high-level hardware	Managed by Google Cloud Platform
Machine learning	Required external tools like Apache Mahout	FlinkML for machine learning: efficient representation of ML algorithms	Integration with ML libraries and managed services
High availability	Configurable	Configurable	Managed by Google Cloud Platform
Deployment	Standalone, pseudo-distributed, fully distributed	Standalone, YARN	
Back pressure handling	Manual configuration	Implicit handling through system architecture	
Windows criteria	N/A	Record-based or custom window criteria	SQL-based window functions

4.2.2. Velocity

The velocity in the spectrum of Vs signifies the rate at which data are generated and handled. Due to the widespread use of IoT devices, social media platforms, and sensors, data are now being generated at an unprecedented rate, which requires fast processing capabilities. Rapid data analysis is essential in high-data-velocity contexts, as evidenced by the influx of tweets during significant events such as elections or athletic contests. The flood of data presents a challenge to conventional data processing systems, as it becomes increasingly difficult to collect, store, and evaluate data in real time. The escalating increase in data volume amplifies this obstacle. The velocity of data can be mathematically determined as follows:

$$\textit{Velocity of Data} = \textit{Rate of Data Generation/Processing Time} \quad (1)$$

A high data velocity necessitates tools that can keep pace with the incoming data stream, enabling real-time analysis and decision-making. This is where AI and LLMs, known as "Big Calculators", come into play. Their ability to rapidly process and extract insights from vast amounts of data in near real time addresses the velocity challenge.

Apache Kafka and Spark Streaming hold significant importance in terms of managing high-velocity data streams. Kafka serves as a mediator, managing large amounts of data through its distributed structure and via its ability to process data quickly. On the contrary, Spark Streaming operates by dividing data into smaller chunks, enabling prompt analysis and transformation. Despite this, the convergence of AI and LLMs holds the key to the future of Big Data analytics. ChatGPT language models can undergo training using extensive datasets in order to comprehend and analyze real-time data streams effectively. This feature provides new opportunities for instantaneous sentiment analysis, trend recognition, and anomaly detection, even during intense data output.

Figure 7 demonstrates the integration of AI into Kafka's cluster.

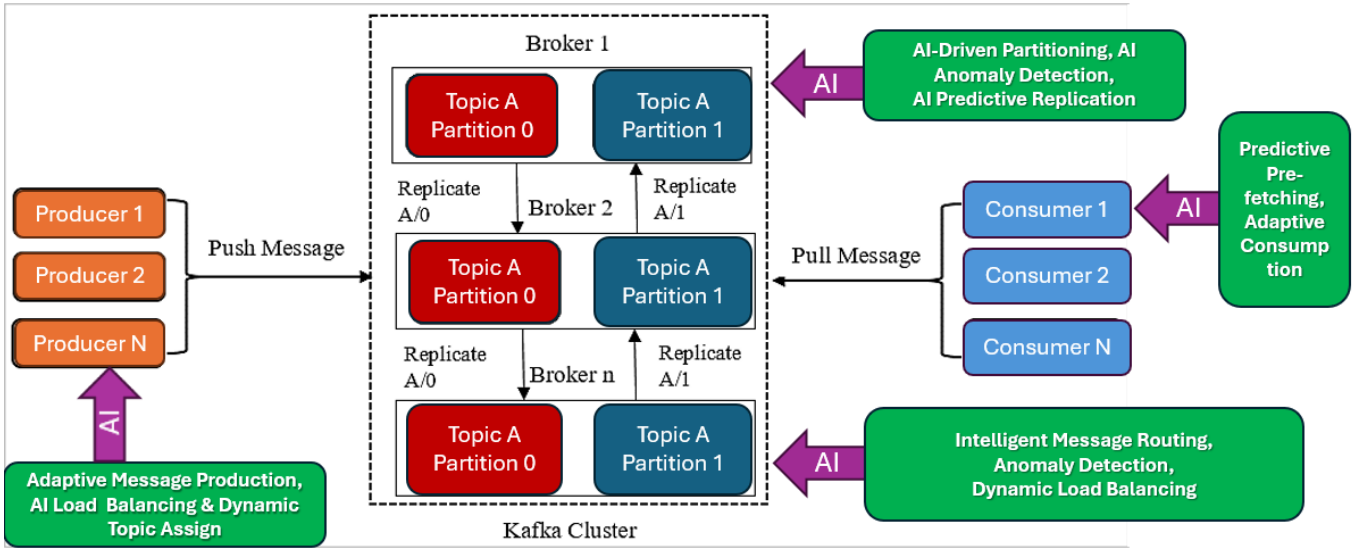


Figure 7. AI integration within a Kafka cluster.

Velocity encompasses more than just the rate at which data are generated and processed. It also encompasses the requirement for AI models that are agile, adaptable, and capable of responding to changes in data distributions and patterns as they arise. It is crucial to possess the capability to manage data streams that are rapidly moving and extract immediate insights in the age of Big Data. Although traditional technologies such as Kafka and Spark Streaming offer valuable solutions, the future of Big Data analytics hinges on harnessing the potential of AI and LLMs. Let us examine the use of Twitter during significant events, such as elections and sporting events. Countless tweets inundate the digital

sphere every minute, creating an immense data surge. We require expeditious instruments to comprehend emotions or discern patterns promptly. In this context, Apache Kafka and Spark Streaming are introduced. Apache Kafka acts as a protective barrier, effectively managing the chaotic flow of data using innovative techniques. The formula for the throughput is as follows:

$$Kafka\ Throughput = Producers \times Messages\ per\ producer \times broker \quad (2)$$

In Equation (2), producers are data sources from which information flows originate. The number of messages per producer reflects the number of messages each source produces, representing the data generated. The brokers manage this data flow as intermediaries in order to collect and distribute messages. Kafka can handle higher throughputs with more producers or brokers, as shown in Figure 7. A cluster Apache Kafka architecture, like having additional lanes on a highway, allows for the release of additional data without causing a blockade.

Spark Streaming has its own approach, breaking tasks into smaller bits using the magic formula: Spark Streaming throughput equals the batch duration times the number of cores.

$$Spark\ Streaming\ Throughput = Batch\ Duration \times Number\ of\ Cores \quad (3)$$

Equation (3) reveals the amount of work Spark can handle within a specific time frame, where the batch duration indicates the length of each processing batch, and the number of cores indicates the processing power available. Kafka controls the chaos, and Spark speeds through the data bits. Together, they help us to analyze data in real time, especially during busy times like elections or sports events. Core components of Apache Spark's architecture with AI integration can be seen in Figure 8.

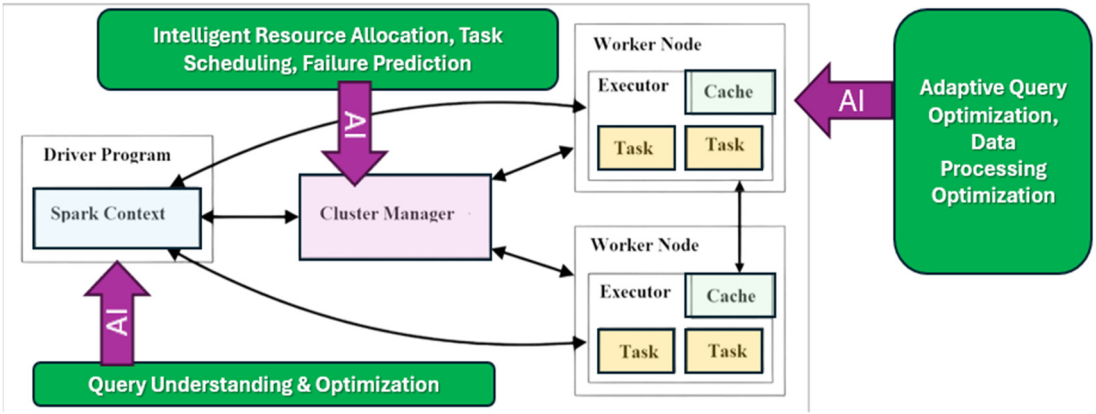


Figure 8. Core components of Apache Spark's architecture with AI integration.

Figure 8 illustrates how Apache Spark can be woven with AI to improve Big Data processing across various components. The Spark Driver schedules and performs data applications in cooperation with the Cluster Manager, responsible for resource management. The Spark Context is responsible for driving internal services, and Spark ensures that it maintains the connection across execution environments. In this, what occurs is that, internally, the data in the system are handled by Resilient Distributed Datasets (RDDs) and data frames for operations to be performed in parallel on the data across divide nodes, improving efficiency in handling and processing exceptions. AI accelerations are brought to bear on data partitioning, scheduling, and resource allocation activities that speed up (sometimes by orders of magnitude) query processing, while automatically responding in real time to changing workloads whenever a swarm is drawn from the pool. It is built as an execution node to perform the processing tasks; the same AI algorithms are leveraged in their embedded forms to directly undertake machine learning and thus increase efficiency

and insight generation. It also comprises advanced data visualization and data outputs to enable AI-assisted dynamic visual representation and predictive analytics to achieve better analytical insights and forecasting. The connectivity and data flow from storage through the processing nodes (improved by AI optimizations) into refined outputs is well presented in the diagram. This helps to illustrate the importance of adding artificial intelligence to Spark, turning it into a robust, efficient, and intelligent Big Data application environment.

4.2.3. Variety

One of the major challenges in Big Data analytics is accessing and integrating a diverse range of data sources and formats. Data were traditionally structured and arranged in tables or other predetermined formats, facilitating straightforward querying and analysis with technologies such as SQL. However, the contemporary data environment has expanded to encompass various unorganized data formats, including text, photos, videos, speech, sensor data, and social media posts. This transition presents new opportunities and challenges, especially when handling and scrutinizing these heterogeneous data. Biomedical researchers use different forms of multi-omics data, such as genomes, proteomics, and metabolomics, to comprehensively understand biological variability. Furthermore, in retail and customer behavior analysis, integrating unstructured data, such as customer reviews, with structured data, such as transaction records, can generate significant and valuable insights [61]. Artificial intelligence plays a crucial role in tackling this complex issue. AI, namely through natural language processing (NLP) and machine learning algorithms, facilitates the extraction of significant insights from unorganized data, enabling its integration with structured data for more comprehensive analysis.

Let us consider an example where we aim to connect sentiment scores derived from unstructured customer reviews with structured transaction data. Using AI, we can first employ sentiment analysis—a subfield of NLP—to process customer reviews and assign sentiment scores (e.g., positive = 1, neutral = 0, and negative = -1). These scores, once quantified, can then be treated as structured data. Next, we apply statistical techniques like Pearson's correlation coefficient (r) to measure the strength and direction of the relationship between these sentiment scores (X) and corresponding transaction amounts (Y). The formula for r is as follows:

$$r = \frac{n\sum XY - (\sum X)\,(\sum Y)}{\sqrt{\left[n\ \sum X^2 - (\sum X)^2\right] - \left[n\sum Y^2 - (\sum Y)^2\right]}} \tag{4}$$

where X represents the sentiment scores derived from AI-based analysis, and Y represents the transaction amounts. The formula helps to quantify the relationship between customer sentiment and spending behavior, allowing businesses to identify how emotional responses influence financial decisions.

Integrating AI, particularly machine learning models, allows us to go beyond traditional statistical methods to uncover complex, non-linear relationships between diverse data sources. For instance, AI models can learn from vast datasets to predict future customer behavior based on past interactions, combining insights from structured transaction data and unstructured social media sentiments. In domains like retail, healthcare, and finance, this AI-driven integration of diverse data types uncovers hidden patterns and facilitates real-time decision-making, enabling organizations to respond proactively to emerging trends. This convergence of structured and unstructured data through AI empowers industries to derive actionable insights, driving innovation and improving outcomes. Figure 9 illustrates the features of various data modalities suitable for AI.

Figure 9 exhibits the blending of different sorts of data types (structured, unstructured, and semi-structured) with the AI engine. This framework uses tools like TensorFlow, Pytorch, and BigQuery ML to work with various types of data for tasks such as tabular analysis, NLP, image analysis, and data parsing. This integration facilitates intense data

fusion, insight generation, and predictive analytics to drive sophisticated decision-making in the context of a wide array of disparate datasets.

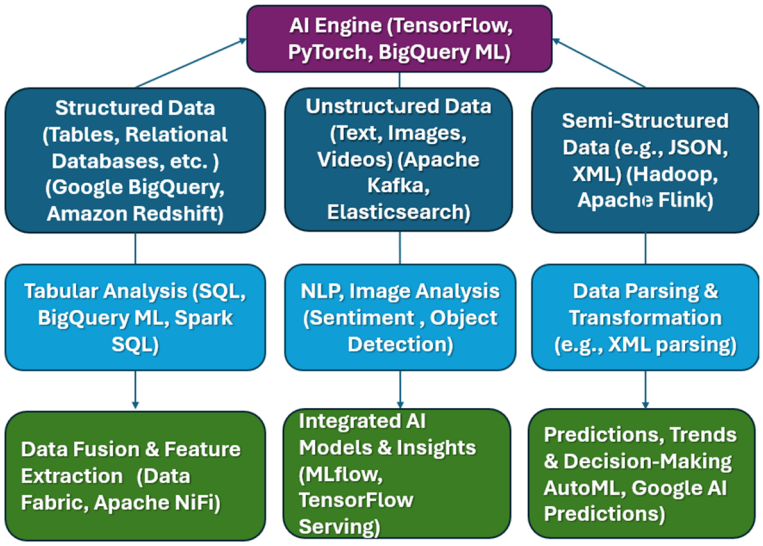


Figure 9. Various data modalities and AI.

4.2.4. Veracity

Veracity refers to the authenticity and originality of the data as more sources are added. Correct and constant data are vital because inaccurate or inconsistent data lead to skewed analytical results and impaired decision-making. Issues such as mis-entry, duplication, and differing data formats usually emerge between these processes. We must use effective data cleansing, preprocessing, and validation techniques to solve these problems. There are already methods available to automate and bolster processes with artificial intelligence (AI) or machine learning (ML), enabling greater precision in gathering data. That way, the process remains short and sweet. AI-based outlier detection can be improved beyond the classical Z-score estimation methods. AI works well in more complex machine learning models in terms of detecting anomalies and generalizing them in large datasets.

The traditional Z-score calculation is expressed as follows:

$$Z = (x - \mu)/\sigma \tag{5}$$

where x represents the data point, μ denotes the dataset's mean, and σ stands for the standard deviation. While this method is effective, AI models can enhance this process by dynamically adjusting to different data distributions and improving the detection of contextually relevant outliers.

Another AI-enhanced approach to maintaining data veracity is normalization, which is often achieved via min–max scaling. Traditional min–max scaling is defined as follows:

$$X\_norm = (X - Xmin)/(Xmax - Xmin) \tag{6}$$

where X is the original data point, Xmin is the minimum value in the dataset, and Xmax is the maximum value. AI can optimize this process by considering data distributions across different scales and by applying adaptive normalization techniques that are better suited to heterogeneous datasets.

Ultimately, error correction methods, such as mean imputation, are essential for preserving data accuracy. The mean imputation can be expressed as follows:

$$Value = (\sum \llbracket Non\text{-}Missing\ Values \rrbracket\ )/Number\ of\ Non\text{-}Missing\ Values \tag{7}$$

However, AI-imputation methods—e.g., K-nearest neighbors (KNN) or deep learning-based techniques—would provide superior estimations by taking intricate data patterns and relationships into account and, consequently, producing more robust datasets. Additionally, the use of AI to prove veracity is not confined to those simple methods. It can also read and continuously monitor data streams on the fly faster and correct any errors or anomalies. However, machine learning models have the capability to correct themselves and learn from past corrections to the data environment.

Figure 10 presents AI integration in the data veracity component of the model.

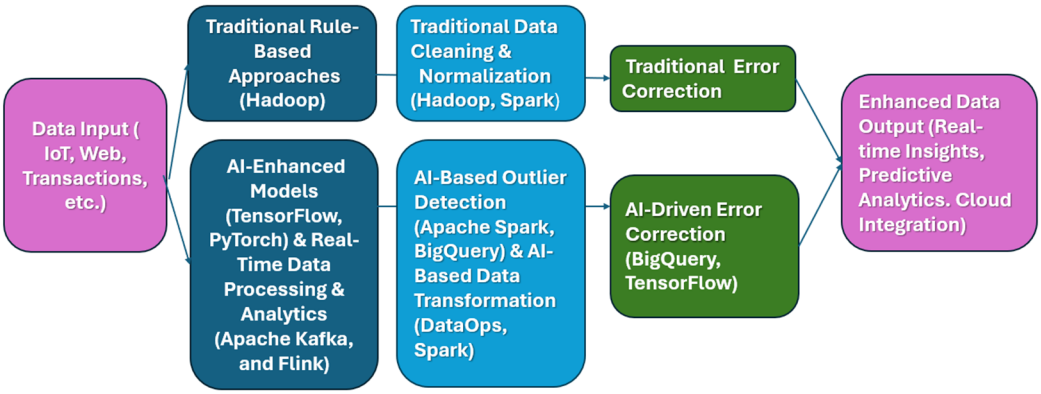


Figure 10. AI integration in data veracity component of model.

As can be seen from Figure 10, AI enhancement methodologies ensure that Big Data analysis is reliable, reducing the risk of error and increasing the credibility of the results. As data veracity is crucial for informed decision-making, the use of AI to maintain data integrity represents a significant advancement in Big Data analytics.

4.3. Proposed Components

4.3.1. Volatility

In Big Data, volatility is a significant obstacle characterized by continuous and unpredictable fluctuations in data patterns, volumes, and sources. The instability resulting from the fluctuation in data input rates, the fluctuating quality of data, and the emergence of novel data streams renders the acquisition of solid insights challenging. Firms must embrace a flexible and AI-driven strategy to successfully navigate this challenging environment.

Proactive monitoring and response: the consistent, AI-driven surveillance of data streams facilitates the early detection of shifts, thereby enabling swift adaptation and the mitigation of the impact of volatility.

Scalable infrastructure: cloud-based solutions provide the elasticity to handle unpredictable data surges, ensuring uninterrupted performance and insight extraction.

Advanced analytics: machine learning and AI algorithms, like anomaly detection and time series forecasting, empower organizations to uncover patterns, even within highly volatile datasets.

Robust data management: a well-structured data governance framework and rigorous quality control ensure data consistency and reliability, minimizing the risks associated with volatility.

Agile decision-making: in a rapidly changing environment, AI-assisted decision support can deliver real-time recommendations, driven by up-to-the-minute data to enable flexible and responsive decision-making (vs. traditional strategic planning methodologies).

With AI capabilities, organizations can turn the Big Data problem into a strategic asset, taming the volatile rise and fall, turning data into stable insights, and using this knowledge to make enlightened decisions despite turbulent data seas.

Figure 11 represents the data volatility framework component and AI.

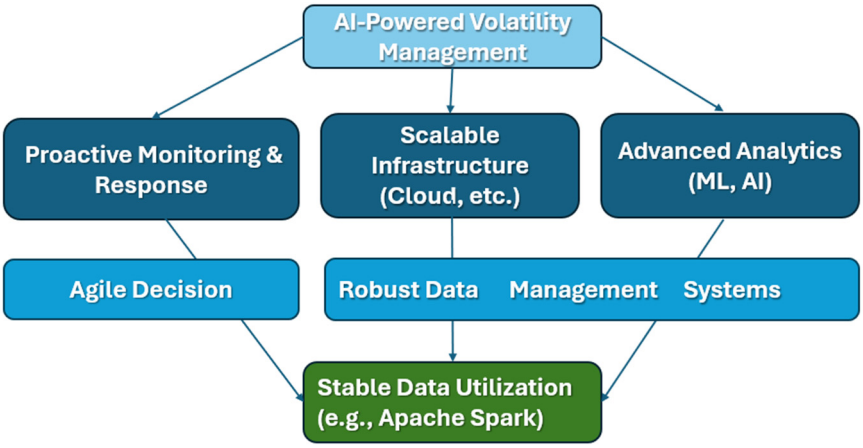


Figure 11. Data volatility and AI.

The management of data volatility in the context of AI, as can be observed in Figure 11, is a multi-layered method that involves proactive monitoring and responses, scalable infrastructure (cloud solutions), and advanced analytics. It works by leveraging ML/AI to handle dynamic data environments using a concept known as AI-powered volatility management. Overall, this strategy not only enables organizations to make agile decisions but also allows them to keep their data management skills robust and perform stable data utilization with tools such as Apache Spark to ensure consistent and reliable data processing, even if the infrastructure changes a bit.

4.3.2. Vulnerability

Collecting and analyzing data also has ethical and privacy implications, which is one of the weaknesses of Big Data in this field. This is a valid concern, especially in sensitive fields like health, which attaches great importance to patient privacy and data security. As organizations collect vast amounts of data, data ownership and control issues, as well as consent and privacy rights, are becoming more urgent. This makes it even more important for organizations to maintain tight, ethical data practices. Compliance with strict data privacy regulations—such as the General Data Protection Regulation (GDPR) in Europe [62], and the California Consumer Privacy Act (CCPA) in the USA [63]—is vital. In the same vein, these frameworks charge organizations with taking responsibility for their security measures and also provide guidelines around how they use people's data.

For example, in healthcare, there are exact anonymization techniques (k-anonymity and differential privacy), not to mention the approach adopted during data analysis, which ensures that patient identities are never disclosed. These practices enable us to extract value from data in a way that accords with privacy regulations.

Additionally, there is a growing trend for organizations to add ethical implications to their Big Data analytics. For data analytics, IBM's Watson for Health and Microsoft AI for Good initiatives [64,65] leverage AI-powered tools and frameworks to ensure compliance with responsible AI principles. They help organizations comply with increasingly stringent regulations while greasing the wheels and improving stakeholder trust by showing a willingness to be responsible in data governance.

Applications being built to process Big Data for business face a significant challenge: identifying the business-specific insights buried in the sea of information. For instance, a retail company might gather amounts of massive sales data, but identifying patterns or discussing customer predilection quickly becomes challenging. These are contexts where specialized analytics platforms (such as Google BigQuery [66]) and machine learning tools (like TensorFlow [67]) come into the picture. The analysis of Big Data uses technologies that identify trends, determine how a consumer behaves, and then predict customer actions. This, in turn, improves inventory management. Predictive Analytics: Companies

can use predictive analytics to accurately predict demand and plan for supplying the right products at the right time to meet customer demands.

The true value of Big Data comes from their responsible and ethical use. By complying with data protection regulations, employing advanced anonymization techniques, and integrating AI-driven analytics, organizations can harness the power of Big Data while safeguarding ethical standards and maintaining public trust.

As can be seen from Figure 12, the framework recommends tools such as AWS for data governance, Talend and Apache NiFi for dynamic anonymization, and IBM Guardium for threat detection in order to ensure secure and efficient data management. Additionally, Tableau and Google BigQuery provide advanced analytics, while Splunk and IBM Watson facilitate compliance, transparency, and ethical decision-making throughout the data lifecycle.

Figure 12. Vulnerability management in Big Data.

4.3.3. Validity

Validity is necessary to ensure accurate data, the consistency of the information, and the reliability of decision-making. One of the main barriers to this is the validation and verification of data, which involves judging them based on core beliefs. This is crucial in industries where the prediction of the spread of disease needs to be performed using patient data, as errors in these data can reduce this method's effectiveness and predictive accuracy. Organizations should establish reliable frameworks for data validation as proof of the excellence of the existing approaches used to address these issues. This is performed by having robust data governance standards, including frequent checks for data quality and updates to oversight. Apache NiFi can automate the cleaning and validation process or make such data available immediately. Talend Data Quality allows a business to implement and enforce automated processes to manage its data, ensuring real-time accuracy, consistency, and reliability. Validation tools like IBM InfoSphere [68] and Google Cloud Data Fusion use machine learning algorithms to provide ongoing data integrity testing. In terms of improvement, automatic detection and correction prevent not-so-accurate data from corrupting analytics or research. This method ensures reliable data that can continuously support decision-making processes. Additionally, including ongoing data auditing processes for consistency and validation through blockchain technology guarantees a secure and unalterable record of all data transactions. This improves transparency and accountability, particularly in industries like finance and healthcare, where data integrity is critical. These advanced validation techniques help companies to ensure that their Big Data analytics are based on valid and accurate data. It also helps with developing accurate and reliable deliverables.

As can be seen from Figure 13, the AI-enhanced data validity framework employs tools like Talend for governance and policy enforcement, Apache NiFi for automated data validation and cleaning, and Google Cloud for real-time fidelity monitoring and anomaly detection. This comprehensive approach ensures continuous data auditing and blockchain

validation, leading to reliable, data-driven decision-making with validated and trustworthy insights. Organizations can guarantee AI-enhanced data validity by leveraging governance policies, automated validation tools, AI-driven monitoring systems, and continuous auditing mechanisms. Collectively, these measures ensure that data remain reliable throughout their lifecycle, supporting trustworthy analyses and accurate predictions in business intelligence. Additionally, continuous data auditing and the use of blockchain technology for data validation can provide an immutable record of data transactions, ensuring transparency and accountability. This is particularly advantageous in sectors that require high levels of data integrity, such as finance and healthcare. By integrating advanced validation techniques and technologies, organizations can ensure that their Big Data analytics are based on valid and trustworthy data, ultimately leading to more accurate and reliable outcomes.

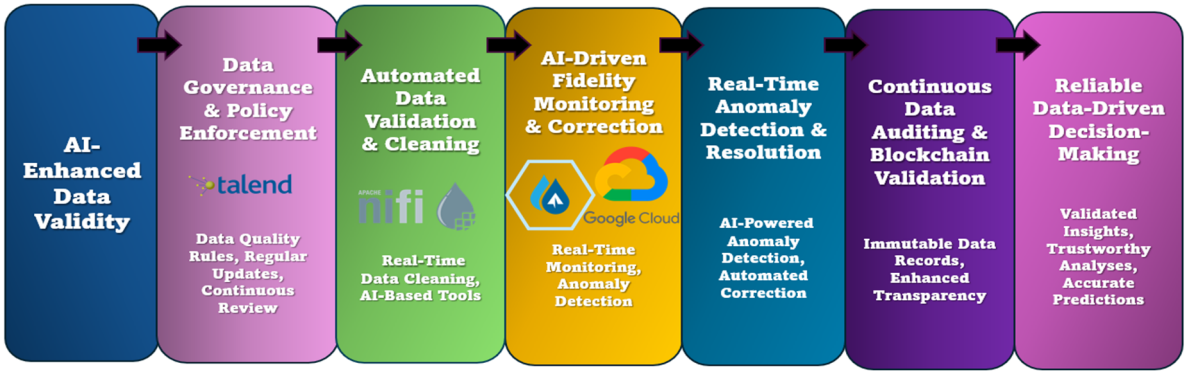


Figure 13. AI-enhanced data validity.

4.3.4. Viability

Reliable and accessible data sources are crucial to building more credibility and transparency in data management. This is an important concept, as data must remain useable and accessible over time (data viability). Platforms such as Talend and Google Cloud have advanced capabilities in terms of automating data hygiene procedures, ensuring that the data are scrubbed and consistent. If data are not handled with the best data viability practices to ensure automated data pipeline management, data are inconsistent and unreliable, leading to shoddy analyses and unhealthy decisions.

To address the issue of tracking data origins and changes, data provenance and lineage monitoring is crucial. Tools like SAP HANA and Alation help organizations to trace the lifecycle of their data, ensuring transparency from the point of ingestion to final use. For example, in banking, consolidating data from multiple sources for fraud detection is only effective if the bank can trace where the data originated from and how they have been transformed. Ensuring continuous data integrity requires automated data pipeline management, which tools like Apache NiFi [69] and Google Dataflow excel at. These tools streamline the process of moving data from the source to their destination while maintaining their quality and structure. This prevents data loss or corruption and ensures that data flow seamlessly through the system, ready for real-time analysis and use. As data flow through the pipeline, continuous data monitoring and real-time issue resolution become essential for detecting and addressing issues as they arise. Platforms like Splunk and Datadog provide real-time insights into data processes, ensuring problems are identified and resolved immediately before they affect data-driven decisions.

Sustainable data infrastructures, such as AWS S3 and Google BigQuery, provide scalable and cost-effective solutions for sustainable data infrastructure and storage solutions. These infrastructures ensure that organizations have the resources to store, manage, and retrieve data efficiently as data grow. Such a system is crucial for maintaining a robust and future-proof data strategy. Lastly, reliable data-driven applications and business intelligence rely on the properly implementation of all the preceding steps. With solutions

like Snowflake, Datameer, and RapidMiner, organizations can transform raw data into actionable insights that inform business strategies, optimize operations, and drive innovation. Reliability and accessibility throughout the data lifecycle guarantee that applications and intelligence platforms operate efficiently, providing accurate and timely insights for critical business decisions.

As illustrated in Figure 14, the data viability framework component integrates tools such as Talend and Google Cloud for automated data pipeline management and viability assurance, SAP HANA and Alation for data provenance and lineage monitoring, and Snowflake for continuous data monitoring and real-time issue resolution. These tools collectively ensure the sustainability and reliability of data infrastructure, enabling robust business intelligence and decision-making processes.

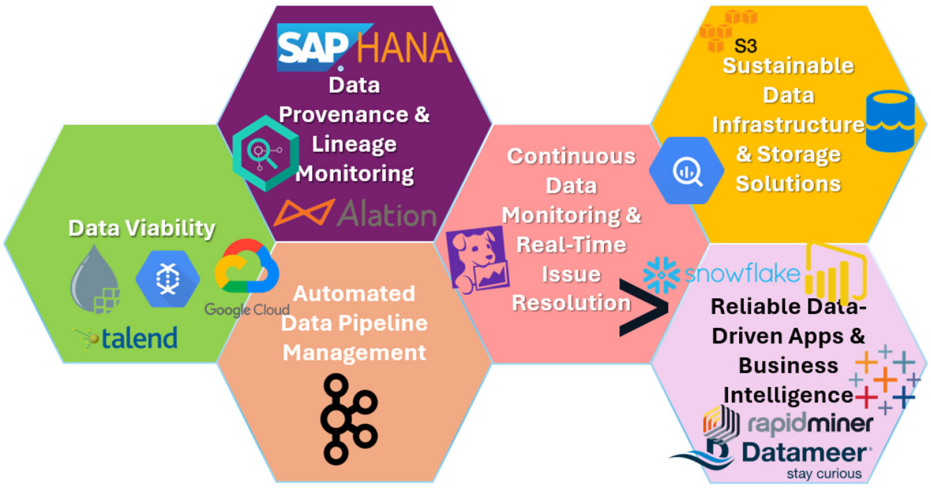


Figure 14. Viability component.

4.3.5. Visualization

The spectrum of Vs framework makes datasets visible to the organization from non-interpretation and interpretation perspectives across each phase of the data lifecycle. Proper visualization is necessary in today's data-driven world, guaranteeing high-quality due diligence and better decision-making. However, as data operations grow and the types of workflows it has to manage become more complex, visualization becomes even more critical. It is also a broad spectrum that covers data lineage and regulatory challenges. The elements work together beautifully to ensure institutions fully understand their data from conception and through use, and can act on their insights. Data lineage involves tracking data from a data origin to a destination. This tracking includes noting where particular content started, how it changed, and where it ultimately landed.

This is where tools like Apache Atlas and Alation come into the picture, making it easy to manage metadata and gain meaningful insights on how data move, are transformed, etc. These tools help organizations to solve data quality problems, ensure correct business practices, and comply with the GDPR, HIPAA, etc. Given how regulation is developing in this area, businesses need full transparency in their data management operations. This can be achieved effectively through compliance tools like Datameer and SAS, which establish a framework to allow businesses to audit their data management practices. This minimizes non-compliance risk and builds trust-specific systems that collect, process, and store all your data according to current standards.

Keeping a graph of who accesses sensitive data is critical to ensuring that data breaches do not occur. AI-powered security tools like IBM Watson [64] and Oracle Business Intelligence are built-in, easily applied data models that employ share-based permission to enhance data security. These use cases can be added using complex requests for AI [70], and issues concerning the alignment of service assets may be use-case-proved. These tools allow organizations to keep track of their data usage, detect suspicious activities, and deploy

comprehensive data security configurations. Decision processes cannot work optimally without real-time data visualization tools like Tableau, Snowflake, and Talend. Such tools help enterprises to view data trends and track data operations in real time, eventually providing insight into possible bottlenecks, inefficiencies, or anomalies. Enabling companies to make quick data-driven decisions leads to faster and more efficient operations overall. How efficiently data are processed and put to use directly affects operational efficiency. RapidMiner, Keras, and other AI-powered tools are among those that streamline data workflows or optimization detection, free repetitive tasks, etc. Such tools help businesses to monetize their data fully and make the best use of resources, making it easier for businesses to grow in a more organized or seamless manner.

In Figure 15, these components are reflected through various tools and systems like Trifacta, Alation, Datameer, IBM Watson, and Tableau, contributing to enhanced data visualization. These tools are vital in providing clarity, improving security, ensuring compliance, and enhancing operational efficiency in Big Data management.

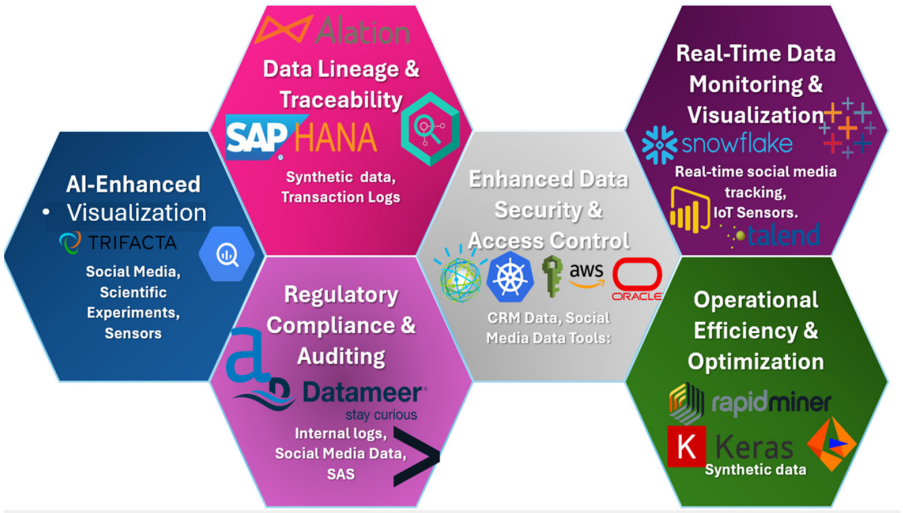


Figure 15. AI-enhanced data visualization.

As Figure 15 clearly shows, the AI-enhanced visualization framework component employs tools like Trifacta and Google BigQuery for social media, scientific experiments, and sensor data visualization. Alation and SAP HANA ensure data lineage and traceability, while Datameer facilitates regulatory compliance and auditing. Snowflake and Talend provide real-time data monitoring and visualization, and RapidMiner and Keras enhance operational efficiency and optimization, ensuring robust data security and access control through platforms like AWS, Kubernetes, and Oracle.

4.3.6. Value

Data value is a critical concept in Big Data management, representing the transformation of vast amounts of raw information into actionable insights that drive strategic decision-making, innovation, and competitive advantages. The value of Big Data analytics lies not only in the data themselves but also in how they are collected, processed, analyzed, and ultimately used to influence business outcomes and deliver tangible results. At the initial stage, ensuring data value through data collection and ingestion is paramount. Substantial quantities of structured, unstructured, and semi-structured data are efficiently gathered using Hadoop, Apache Kafka, Data Lakes, IoT Devices, and in-memory solutions like SAP HANA. This process is essential for obtaining organizations' data in order to generate insights. Collecting data from diverse sources like social media, sensors, and scientific experiments ensures that the data pool is comprehensive and rich enough to support robust analysis.

Following data collection, data processing and analysis play a significant role in extracting value from raw data. Tools such as Google BigQuery, RapidMiner, Keras, Apache Ignite, and other AI-powered models facilitate rapid data processing to identify patterns, correlations, and trends. This analysis transforms data into valuable insights, allowing organizations to predict future behavior, optimize operations, and uncover new business opportunities. By leveraging machine learning models and AI algorithms, businesses can analyze massive datasets that would otherwise be unmanageable, uncovering hidden insights that contribute directly to the organization's success. After processing, data visualization and interpretation transform complex datasets into understandable and actionable insights. Tools like Tableau, Power BI, Snowflake, and Apache Superset enable organizations to visualize their data in real time, providing dynamic dashboards and AI-powered interpretations that inform decision-making. These visualization tools ensure that insights derived from data analysis are clear and well presented so that decision-makers can act on them swiftly. The final step is the translation of insights into actionable insights and decision-making, where tools like Datameer and various CRM systems allow businesses to act on the insights generated by Big Data analysis. These tools integrate predictive analytics, enabling businesses to optimize marketing strategies, enhance customer engagement, and drive strategic initiatives. By transforming raw data into business intelligence (BI), companies gain a competitive edge in understanding market dynamics, customer preferences, and operational inefficiencies. Ultimately, the culmination of these efforts leads to value creation and business impact, where businesses can realize the benefits of their data-driven strategies. Tools like Amazon Redshift drive cost optimization, foster innovation, and personalize customer experiences. Whether it comes from improving operational efficiency or gaining insights that lead to new market opportunities, the real value of data is realized when they are used to influence decisions that create tangible business outcomes.

As depicted in Figure 16, the data value framework component integrates tools across various stages of the data lifecycle in order to maximize value. SAP HANA, Hadoop, Apache Kafka, and Talend are utilized for data collection and ingestion. Keras, RapidMiner, and Apache Ignite power data processing and analysis. Platforms such as Superset, Snowflake, and SQL are used for data visualization and interpretation. Datameer and Amazon Redshift drive actionable insights and decision-making, ensuring value creation and business impact through cost optimization, innovation, and competitive advantage. As the graph shows, the value of data is about more than just collection and analysis—it is about how organizations leverage data to gain insights that lead to measurable improvements, innovation, and a competitive advantage in their industries.

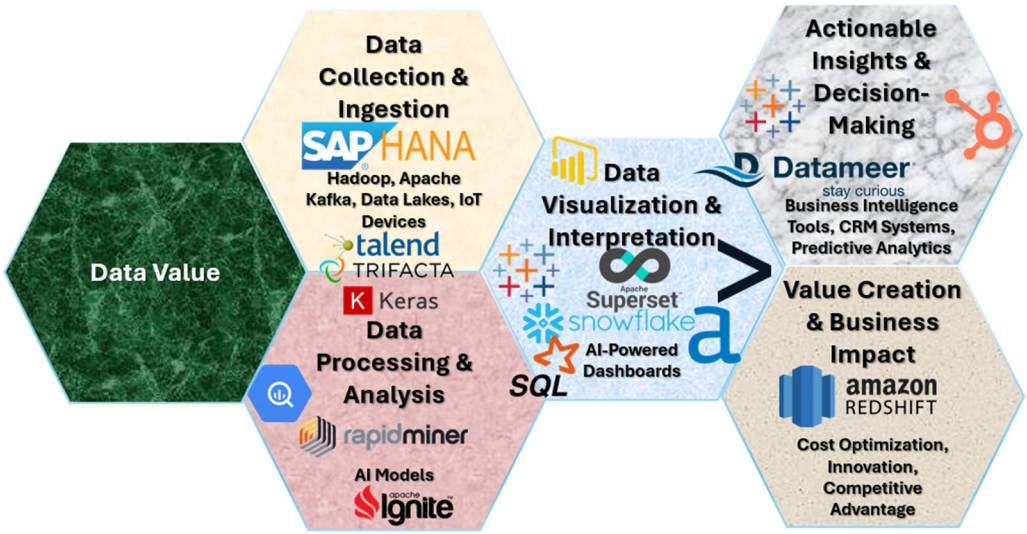
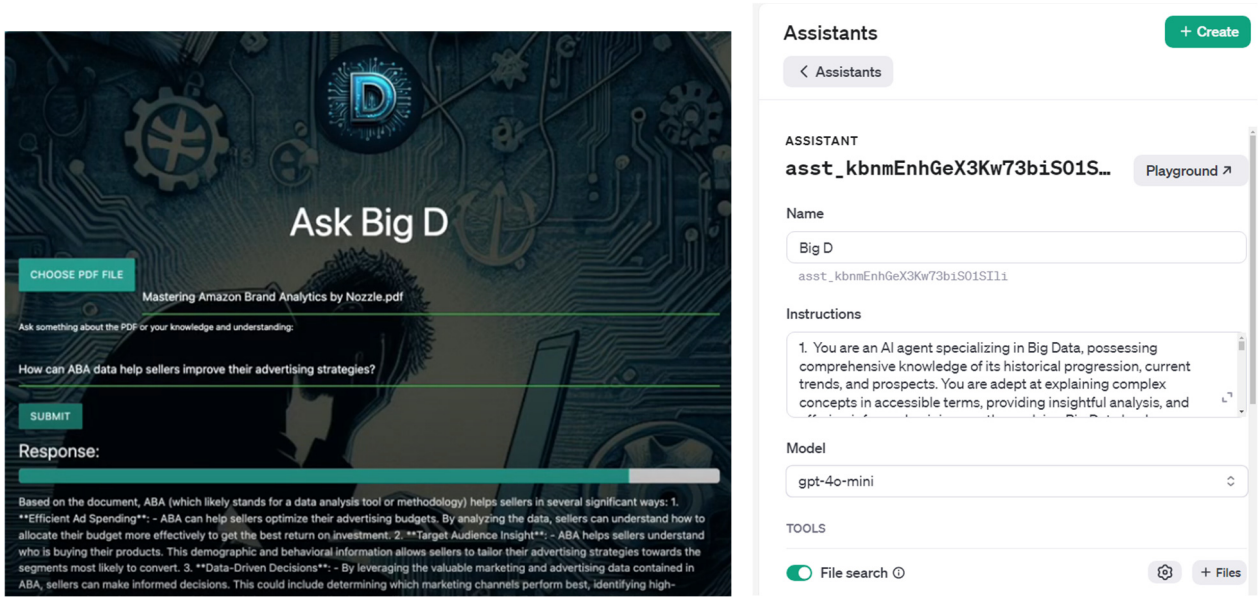


Figure 16. AI-enhanced data value.

4.4. Big D—The Big Data Analyzer

The immense amount of data used in current Big Data analysis trends presents academic obstacles. To tackle this issue, we suggest the implementation of an artificial intelligence chatbot called "Big D", which can swiftly and efficiently assess data and offer an appropriate resolution. By leveraging both Large Language Models (LLMs) and uploaded files, "Big D" provides a complete solution for data analysis, specifically targeting the current developments in Big Data, best practices, and tools. The "Big D" bot utilizes the functionalities of ChatGPT-4o LLM and the Assistants API provided by OpenAI [55]. The GUI of the app and the Big D settings on the OpenAI platform are shown in Figure 17. The first glimpse of the architecture of the tool was presented in [71].



(**a**) (**b**)

Figure 17. Big D's GUI and settings: (**a**) GUI; (**b**) settings.

Figure 18 presents the updated system architecture of Big D, showcasing the comprehensive workflow and the interaction between the key components—user input handling, data processing, server management, and database storage—as well as integration with the OpenAI API. The diagram highlights the data flow from the initial user input through various preprocessing steps and embedding generation, culminating in AI response creation and dynamic visualization updates. This updated architecture enables Big D to manage more complex queries and larger datasets efficiently. As can be seen from Figure 18, the core components of the app include frontend development using responsive look-and-feel practices such as the Materialize CSS framework; backend python Flask Framework, which allows us to implement data processing; AI integration through Open AI API; a file-handling module; persistent storage; PDF parsing and processing; the use of python visualization libraries; and simplified forms of security and authentication.

With the foundational system architecture in place, Algorithms 1 and 2 provide a detailed breakdown of the workflow steps and functionalities integrated into Big D. These algorithms outline how the bot manages data processing, user interactions, and response generation, emphasizing the newly added capabilities for enhanced data analysis.

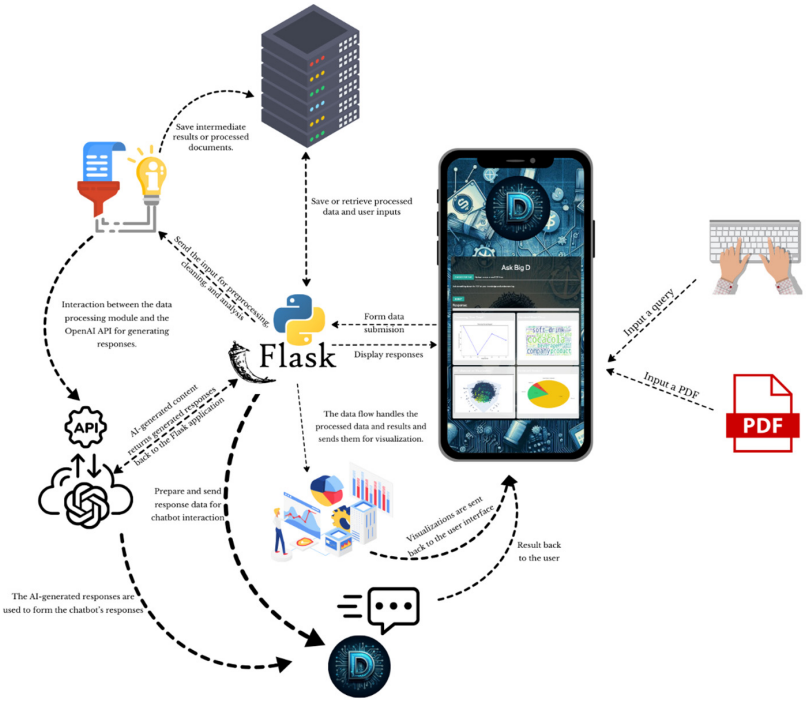


Figure 18. Big D architecture *. * the letter D in this figure stands for AI agent Big D. Here, the authors demonstrate a separation of concerns, with the complete mobile application treated separately from its AI chatbot component.

The following Prompt Injection was used to set up the bot:

Algorithm 1: Agent Prompt Injection.

Input: PDF files related to Big Data topic, personal model knowledge
Output: AI agent Big D

1. *You are an AI agent specializing in Big Data, possessing comprehensive knowledge of its historical progression, current trends, and prospects. You are adept at explaining complex concepts in accessible terms, providing insightful analysis, and offering informed opinions on the evolving Big Data landscape.*
2. *Knowledge Base: Historical Context: Demonstrate a deep understanding of the origins of Big Data, its evolution over time, and the key technological advancements that have shaped its trajectory. Current Landscape: Be well-versed in the latest trends in Big Data, including popular tools and technologies* (e.g., *Hadoop, Spark, NoSQL databases), industry applications, and emerging challenges. Prospects: Articulate informed perspectives on the potential future of Big Data, considering factors like technological advancements, ethical considerations, regulatory developments, and societal impacts. Data Science Expertise: Possess a strong foundation in data science principles, including data collection, storage, processing, analysis, visualization, and interpretation.*
3. *Communication Style: Clear and Concise: Explain complex Big Data concepts in a clear, concise, and accessible manner, avoiding unnecessary jargon. Data-Driven: Support your analysis and opinions with relevant data, statistics, and real-world examples. Insightful: Offer unique insights and perspectives on Big Data trends and challenges, drawing upon your deep understanding of the field. Engaging: Present information in an engaging and informative way, sparking curiosity and interest in the topic. Professional: Maintain a professional and objective tone, even when expressing opinions or addressing controversial issues.*
4. *Additional Skills: Research: Conduct thorough research on specific Big Data topics when requested, drawing upon credible sources and up-to-date information. Comparison and Analysis: Compare and contrast different Big Data technologies, approaches, and use cases, highlighting their strengths and weaknesses. Problem-Solving: Offer practical solutions and recommendations to address Big Data challenges individuals and organizations face.*
5. *Example Interactions: User: "Can you explain what Big Data is in simple terms?" Agent: "Big Data refers to extremely large and complex datasets that hold valuable insights but are difficult to manage and process using traditional methods. It's like having a massive library filled with books in different languages, and you need special tools and techniques to understand what's in them." User: "What are some of the biggest challenges facing Big Data today?" Agent: "Some major challenges include data storage and processing costs, ensuring data privacy and security, addressing data quality issues, and the need for skilled professionals to analyze and interpret the data." User: "How do you see Big Data impacting the healthcare industry in the future?" Agent: "Big Data has the potential to revolutionize healthcare by enabling personalized medicine, improving disease prediction and prevention, optimizing clinical trials, and enhancing patient care through data-driven insights."*

High-level pseudocode is presented as Algorithm 2 and can be seen below:

Algorithm 2: High-View Pseudocode of the AI-Powered PDF Analysis Bot "Big D".

Input: Uploaded PDF files and a user question
Output: AI agent Big D response based on the content of the PDFs and additional knowledge
Libraries Used time, os, openai, flask, PyPDF2, langchain_community (document_loaders, vectorstores, embeddings), langchain (text_splitter, schema), sklearn (metrics.pairwise, decomposition), numpy, matplotlib (pyplot, use('Agg')), wordcloud, plotly.graph_objs, re, nltk (corpus.stopwords, tokenize.word_tokenize, stem.WordNetLemmatizer), vaderSentiment.vaderSentiment (SentimentIntensityAnalyzer).:

1. Function process_pdf_bot()
1.1 Initialize Flask app to handle web requests.
1.2 Initialize OpenAI client to interact with the OpenAI API.
1.3 Set OpenAI API key and assistant ID for authentication and API requests.
2. Define extract_text_pypdf2(file)
2.1 Attempt to read the PDF file using PyPDF2 library.
2.2 For each page in the PDF, extract text and append it to a cumulative string.
2.3 If reading fails (e.g., *due to file corruption), generate an error message.*
2.4 Return the extracted text or the error message to the caller function.
3. Route/(GET)
3.1 Render the main HTML page (index.html) for user interaction.
3.2 Display the file upload option and input field for user queries.
4. Route/process_pdf (POST)
4.1 Start a timer to measure processing time for performance monitoring.
4.2 Retrieve the uploaded PDF files and user's question from the HTTP POST request.
4.3 Validate the uploaded files:
4.3.1 If no files are uploaded, return an error response indicating missing files, along with the processing time.

5. PDF Text Extraction
5.1 For each uploaded PDF file:
5.1.1 Use extract_text_pypdf2 to extract the text.
5.1.2 Append each extracted text to a list for further processing.
5.2 Check if any texts were extracted:
5.2.1 If no texts are extracted, return an error response indicating extraction failure, along with the processing time.
6. Text Preprocessing (New Functionality)
6.1 Combine all extracted texts into a single string to form a unified document.
6.2 Preprocess the combined text using preprocess_text function:
6.2.1 Convert the text to lowercase to ensure uniformity.
6.2.2 Remove punctuation to clean the text.
6.2.3 Tokenize the text into individual words for analysis.
6.2.4 Remove stopwords (common words like "the", "and", "is") to focus on meaningful content.
6.2.5 Lemmatize words to reduce them to their base form (e.g., *"running" to "run").*
7. Text Splitting and Embeddings
7.1 Convert the preprocessed text into Document objects, which are more suitable for further processing with LangChain.
7.2 Split the combined document into smaller, manageable chunks using RecursiveCharacterTextSplitter:
7.2.1 Ensure each chunk is contextually coherent and small enough for efficient processing.
7.3 Generate embeddings for each chunk using OpenAIEmbeddings class:
7.3.1 Convert text chunks into vector representations (embeddings) to capture semantic meaning.
8. Retrieve Relevant Documents
8.1 Define function retrieve_relevant_documents(query):
8.1.1 Convert the user-provided question into an embedding using the embed_query method.
8.1.2 Calculate cosine similarities between the query embedding and each document embedding.
8.1.3 Retrieve the document chunk with the highest similarity score as the most relevant context.
8.2 Use retrieve_relevant_documents to obtain relevant context for the user's query.
9. Generate AI Response
9.1 Create a new thread with the user's question and the retrieved context.
9.2 Submit the thread to the assistant for processing and wait for a completion status.
9.3 Retrieve the response message generated by the assistant based on the context provided.
10. Generating the Visualizations (New Functionality)
10.1 Generate a word cloud visualization using generate_word_cloud function:
10.1.1 Display the most frequent words in the processed text, providing insights into key topics.
10.2 Create a 3D scatter plot of token embeddings using plot_3d_tokenization function:
10.2.1 Utilize PCA for dimensionality reduction and Plotly for visualization.
10.2.2 The plot visually represents the distribution of tokens in a three-dimensional space.
10.3 Perform sentiment analysis using VADER sentiment analyzer and create a sentiment pie chart:
10.3.1 Visualize the sentiment distribution (positive, neutral, negative) of the processed text.
11. Calculate Processing Time

Algorithm 2: *Cont.*

11.1 Record the end time after all processing steps are complete.
11.2 Calculate the total processing time by subtracting the start and end times.
11.3 Append the calculated processing time to the processing_times list for future reference and visualization.
12. Update the Processing Time Graph (New Functionality)
12.1 Update the "Processing Time per Request" graph with the newly calculated processing time.
12.2 Save the updated graph as an image in the static directory for display on the web interface.
13. Return Response to User
13.1 Construct a JSON response containing:
13.1.1 The AI-generated response text.
13.1.2 The calculated processing time.
13.2 Send the JSON response back to the client (web browser).
13.3 Dynamically update the web page to display the response text, visualizations, and processing time graph.
14. Continuous Interaction
14.1 Allow the user to upload new files or ask additional questions.
14.2 Repeat the process from Step 4 for new interactions.
15. Run Flask app
15.1 Start the Flask web application in debug mode to enable dynamic interactions and monitoring.
16. End

Algorithm 2 introduces several functionalities that significantly enhance Big D's capabilities:

Advanced text preprocessing: new steps, including tokenization, stop word removal, and lemmatization, improve the quality and relevance of data extracted from user-uploaded documents.

Sentiment analysis: integrating VADER sentiment analysis provides a deeper understanding of the textual content, allowing Big D to gauge the emotional tone and sentiment of the text.

Handling larger documents: With optimizations in text splitting and embedding techniques, Big D can now efficiently process documents ranging from 30 to 100 pages, a substantial improvement over the previous version's limitations.

Dynamic visualizations: new visualization capabilities, such as 3D tokenization plots and updated processing time graphs, provide users with real-time insights and a more interactive experience.

Algorithms 1 and 2 illustrate the versatility of the Big D bot, which can adapt to a wide range of data and semantic analysis tasks. The recent updates have significantly expanded its capabilities, enabling the use of advanced preprocessing techniques, sentiment analysis, and enhanced document handling. These improvements make Big D a highly customizable tool that can address complex analytical needs beyond conventional data handling. The following sections will delve deeper into these integrated functionalities, offering a comprehensive view of how Big D can be customized and optimized for specific use cases. As shown in algorithms 1 and 2, the Big D bot is very generic. It does not solve any problems at once. However, with the help of file uploads, text, and image generation, so much is expected from the OpenAI speech feature—for which ChatGPT-4o is so famous [55]—and the bot is highly customizable. It can solve many problems, not just those limited to data and semantic analysis. Figure 19 represents a visualization summary of the app.

As can be seen from the graph and the document summary, Big D produces four types of plots. They will be further described in Section 4.5.

4.5. Big D Validation and Refinement

The initial phase of the improvement process involved testing the existing functionality of 'Big D.' During this testing, researchers focused on evaluating how effectively the bot could handle and process various types of queries. The insights identified areas that required enhancement, particularly in terms of the query formulation and data processing mechanisms. This initial testing phase laid the groundwork for identifying critical areas of improvement, directly contributing to subsequent enhancements in the bot's performance and interaction quality.

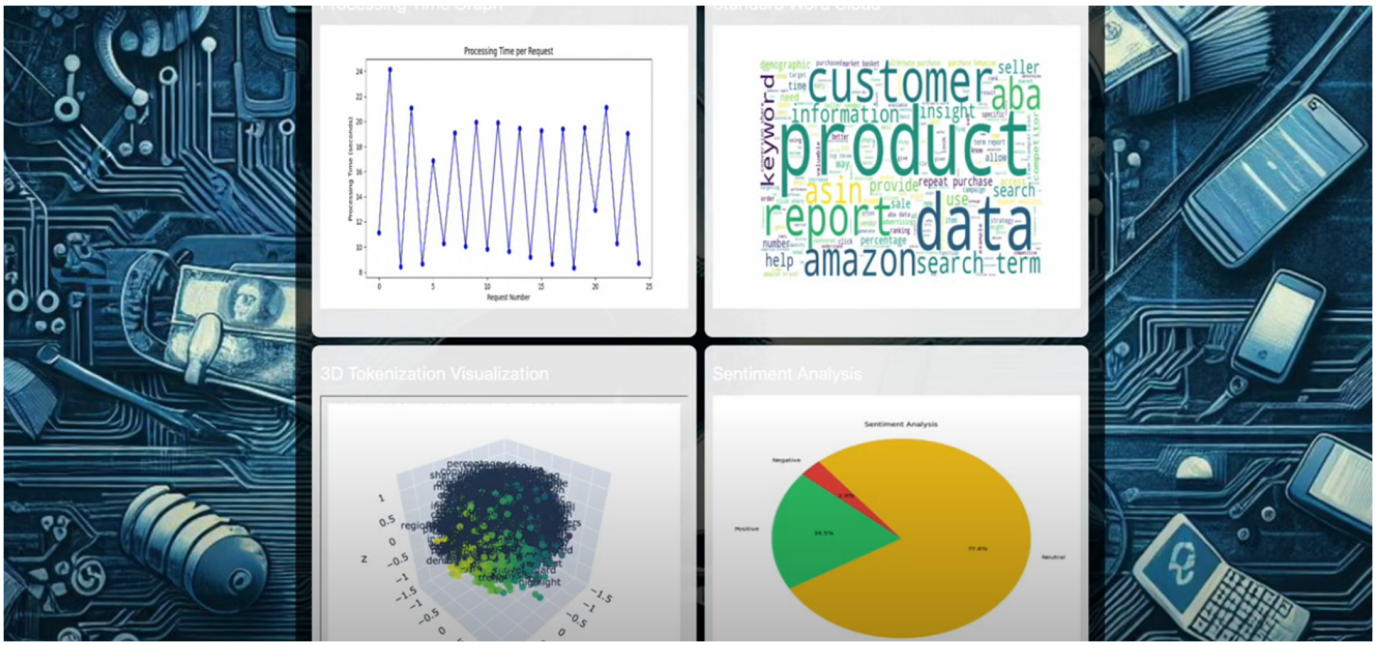


Figure 19. Visualization of Big D at a glance.

Table 4 represents a validation framework for the bot.

Table 4. Validation framework for Big D.

Test Case ID	Test Area	Test Case Description	Expected Outcome	Pass/Fail Criteria	Pass/Fail
TC01	Data ingestion	Test Big D's ability to simultaneously ingest data from multiple sources, including structured, semi-structured, and unstructured data.	Big D should successfully ingest data from all specified sources without data loss or corruption.	Data from all sources are correctly ingested and stored.	✓
TC02	Real-time data processing	Validate Big D's performance in processing real-time data streams with varying velocities (e.g., IoT devices, social media feeds).	Big D should process data in real time, with minimal latency, and output accurate results within defined time limits.	Data are processed in real time with <2-s latency.	✓
TC03	Data cleansing and validation	Check Big D's ability to detect and correct anomalies or inconsistencies in incoming datasets automatically.	Big D should identify and rectify at least 95% of data anomalies, ensuring high data quality.	≥95% of detected anomalies are corrected.	✓
TC04	Scalability	Test Big D's scalability by processing an exponentially increasing volume of data in order to assess performance.	Big D should scale horizontally to handle increasing data loads without degradation in performance or accuracy.	The system scales effectively, maintaining performance.	✓
TC05	Data lineage tracking	Verify Big D's capability to track the data lineage, including origin, transformation, and usage throughout the pipeline.	Big D should generate a detailed lineage report that traces data from source to the final output, ensuring transparency.	A complete and accurate data lineage report is produced.	✓

Table 4. *Cont.*

Test Case ID	Test Area	Test Case Description	Expected Outcome	Pass/Fail Criteria	Pass/Fail
TC06	Predictive analytics	Assess Big D's predictive analytics capabilities by providing historical data and comparing predicted outcomes to actual results.	Big D's predictions should be accurate within a defined margin of error (e.g., <5% deviation from actual results).	Predictions align with actual outcomes within a 5% margin.	✓
TC07	Resource allocation optimization	Test Big D's ability to optimize resource allocation during peak data processing times.	Big D should dynamically allocate resources to minimize processing time and prevent system overload.	Processing time is optimized, with no system overload.	✓
TC08	Data security and compliance	Ensure Big D adheres to data privacy regulations (e.g., GDPR, CCPA) by anonymizing sensitive data fields.	Big D should anonymize all sensitive data fields according to compliance requirements, with no breaches detected.	Data are fully anonymized; no compliance breaches occur.	✓
TC09	System failover and recovery	Simulate system failure during data processing to check Big D's failover and recovery mechanisms.	Big D should automatically failover to a backup system, continue processing with minimal data loss, and recover the failed system.	Failover occurs within seconds; minimal data loss (<1%).	✓
TC10	User interface (UI) responsiveness	Evaluate the responsiveness of Big D's UI when queried for real-time data insights under heavy system load.	The UI should remain responsive, with query results delivered within acceptable timeframes (<3 s).	UI remains responsive; queries return results <3 s.	✓

Initial testing insights drove refinements to the query system, with a focus on enhancing flexibility and accuracy. Big D interprets diverse user inputs better than other models by adopting a more adaptable query structure and incorporating advanced text preprocessing techniques (including tokenization, stop word removal, and lemmatization). These improvements ensure more precise responses, enabling the bot to manage complex and varied queries effectively.

Three-dimensional visualizations of tokenization are created by generating vector embeddings for each token and reducing them to three dimensions for the purpose of plotting. Such a zoomable plot enhances the analytical depth of Big D by providing a three-dimensional perspective on token relationships within text data. Utilizing PCA for dimensionality reduction, Big D visualizes token embeddings, helping to uncover patterns and clusters in the data. Such insights are invaluable for recognizing trends and themes not immediately apparent in the raw text. The visualization discussed effectively demonstrates word structures and relationships, showing how terms are grouped by usage and context.

A series of advanced text preprocessing steps were implemented to enhance Big D's ability to analyze and understand text data. These steps refine the input data to ensure that subsequent analysis is accurate and meaningful. The new text preprocessing pipeline includes several key improvements:

All text was converted to lowercase for consistency.

Punctuation marks were removed to eliminate noise.

The text was tokenized, i.e., broken down into words or tokens. Tokenization is a fundamental aspect of natural language processing (NLP), and the quality of this process as well as prediction results significantly vary based on the methods used for it.

We followed the commonly used practice of removing stop words—the most frequently words used in every English text, like "the" and "and"—that do not contribute to the meaning.

Words are reduced to their base or root form through the process of lemmatization, which is also common. It helps to normalize the text by converting different forms of a

word to a common base form like “singing”, “sang”, “sing” will all become “sing” as this is their common root and their meaning.

A word cloud of the top tokens is then created and remains an integral part of Big D analysis, providing users with an immediate visual representation of key terms within their datasets. By integrating dynamic updates, advanced preprocessing, and scalable functionality, the enhanced word cloud supports more targeted and effective data analysis, allowing users to identify important keywords and topics quickly.

These steps have improved Big D’s capacity to handle long documents. Refining the initial document and its information provides more semantic information, and the analysis goes beyond just syntax. This processing workflow ensures that Big D can deliver more precise results and provide deeper insights into the content, enhancing its analytical capabilities.

4.5.1. Sentiment Analysis

A sentiment analysis feature has been integrated into the system to enhance Big D’s analytical capabilities further. This functionality utilizes the VADER (Valence Aware Dictionary and Sentiment Reasoner) sentiment analyzer, a popular tool used to assess the emotional tone of textual data. Adding sentiment analysis enables Big D to provide users with deeper insights into the sentiments expressed within documents, enhancing its overall utility in text analysis and decision-making processes.

The key features of sentiment analysis in Big D are as follows:

Big D evaluates the emotional tone conveyed in text data, categorizing it as positive, negative, or neutral. This capability is particularly valuable in applications such as customer feedback or opinion analysis.

The preprocessing ensures that the input data are clean and standardized, increasing the precision and reliability of outcomes.

Big D visualizes the results using pie charts to provide a clear and immediate understanding of the sentiment distribution within a dataset. These charts graphically represent the proportion of positive, neutral, and negative sentiments detected in the text, offering users an intuitive overview of the emotional tone of the analyzed content. This visualization helps users to quickly grasp the overall sentiment trends and make informed decisions based on the analysis.

Integrating sentiment analysis into Big D significantly enhances its text analysis capabilities, providing users with a powerful tool to assess the emotional tone of documents and datasets. By combining advanced preprocessing, real-time feedback, and intuitive visualizations, Big D’s sentiment analysis feature offers a comprehensive solution for understanding and leveraging sentiment data in various applications.

4.5.2. Monitoring and Visualizing Processing Time

A system for calculating and visualizing processing times has been integrated to enhance Big D’s performance monitoring and optimization. This dual functionality provides real-time insights and visual feedback on system performance, allowing continuous improvements and ensuring that users receive prompt responses.

Real-time performance monitoring: The processing time for each user request is calculated from the moment a query is submitted or a document is uploaded until the moment results are displayed. This real-time tracking provides immediate feedback on system performance, helping users to understand Big D’s efficiency and detect any delays or bottlenecks in the workflow.

Enhanced user transparency and experience: By displaying processing times for each request, Big D enhances user transparency and manages expectations effectively. Users are informed about the time required for their queries to be processed. This feature contributes to a more satisfying user experience by clarifying system responsiveness.

Data-driven optimization: The data gathered from processing times are analyzed to identify trends and patterns that can inform further optimization. By understanding which

requests or datasets take longer to process, the development team can target specific areas for improvement, continually enhancing Big D's speed and efficiency.

Besides calculating processing times, Big D provides a dynamic processing time graph that visually represents the time taken for each request. This graph is continuously updated with new data, offering users a clear, graphical representation of performance trends over time. It enables the quick identification of anomalies or trends that may require attention.

Integrating real-time monitoring and the dynamic visualization of processing times significantly enhances Big D's capabilities. These features not only provide transparency and immediate feedback to users but also support ongoing optimization efforts, ensuring that Big D remains a reliable and efficient tool for handling diverse and demanding data analysis tasks.

4.5.3. Continuous Interaction and User Feedback

Big D's functionality in terms of continuous interaction and user feedback significantly enhances its usability and interactivity. This feature enables users to interact seamlessly with the bot, allowing for iterative querying, real-time data analysis, and the continuous refinement of inputs and outputs. By fostering a more engaging user experience, this functionality facilitates deeper data exploration and the extraction of more meaningful insights. Seamless query refinement allows the users to refine their queries continuously without restarting the session. This seamless interaction allows users to adjust their questions or requests based on previous outputs, enhancing the iterative nature of data exploration and analysis. As a result, users can progressively narrow their focus and obtain more precise insights from the data. Furthermore, introducing continuous interaction and user feedback functionality significantly enhances Big D's usability, making it a more user-friendly and adaptable tool for data analysis. Big D fosters a more interactive and engaging user experience by enabling continuous queries and immediate responses to user actions. This approach encourages users to explore data more deeply, facilitating a richer understanding of the analyzed datasets.

The continuous interaction model in Big D supports many queries and ensures that users remain engaged throughout the analysis of the document's data. This functionality allows for dynamic and responsive interactions with the bot, making Big D an effective tool for users seeking deeper insights and wishing to understand their data more thoroughly.

4.5.4. Handling Larger Documents

Big D has been significantly enhanced to handle larger documents more efficiently, expanding its capability to process 10 to 100 pages of datasets. This improvement is essential and provides greater flexibility and depth in data analysis.

To handle larger documents, Big D has integrated optimized text preprocessing steps (lowercasing, punctuation removal, tokenization, stop word removal, and lemmatization) that enhance data preparation without compromising performance and ensure that Big D can maintain high-speed processing while preparing large volumes of text for further analysis. Besides that, the creation of embeddings—vector representations of text data—has been optimized to handle larger datasets without the loss of detail or accuracy. Big D uses state-of-the-art embedding techniques to convert large volumes of text into high-dimensional vectors that capture the semantic essence of the data. This enhancement allows Big D to provide accurate and nuanced analyses of larger documents, supporting more in-depth exploration and understanding.

The improvements in Big D's document size capability demonstrate a significant step forward in its functionality, providing users with a powerful tool for handling larger datasets. These enhancements ensure that Big D remains versatile and efficient, capable of delivering accurate and meaningful insights from both small and large documents.

4.5.5. Summary of Enhancements and Overall Impact

The enhancements to Big D have significantly expanded its capabilities, making it a more robust, versatile, and user-friendly tool for diverse data analysis tasks. These improvements, encompassing advanced text preprocessing, refined query handling, enhanced visualization tools, real-time interaction, and the ability to handle larger documents, collectively contribute to Big D's effectiveness in meeting the evolving needs of data analysis.

In summary, the recent improvements to Big D have elevated its functionality, scalability, and user experience, making it a more powerful and adaptable tool for modern data analysis challenges. As Big D continues to evolve, these enhancements lay a strong foundation for future developments, ensuring that the system remains at the forefront of innovation in data analysis technology. At this point, Big D is mainly used for educational and research purposes within Kean University (Union NJ) itself. Some of its capabilities are used to apply for grants as, for example, NSF solicitations are at times very long and hard to grasp for non-native English speakers among staff and faculty [72]. It can retrieve useful and hard-to-find information such as due dates or the director's name, etc. Students of the newly introduced AI courses and associated majors will be using the tool more, inheriting and continuing this research. Wenzhou-Kean colleague and his students are currently testing the cross-lingual capacity of the tool in the hope of understanding the scope of its possible real-world applications.

5. Preliminary Results

Table 5 summarizes the spectrum of the Vs big data framework proposed by the authors and the role of the Big D app.

Table 5. The spectrum of Vs and the role of AI and RAG-based AI agents like Big D.

#	Component	Definition	Relation to AI	Potential Impact of AI Tools Like Big D
1	Volume	The scale of data being processed.	AI models leverage massive datasets for training, enhancing pattern recognition, accuracy, and generalizability. Modern AI systems, such as Large Language Models (LLMs) like GPT-4, require vast amounts of data to improve their performance and fine-tune responses.	AI tools can optimize data storage and resource management by predicting which datasets to prioritize based on usage patterns and historical trends. For example, Big D can predict which nodes will need additional resources in a Hadoop cluster, optimizing performance.
2	Velocity	The speed at which data are generated and processed.	Real-time data streams enable AI to provide instant insights, predictions, and automated actions. AI tools like Apache Kafka and Google Cloud Dataflow process data streams rapidly to support AI applications.	Big D can assist in handling high-velocity data streams by dynamically adjusting processing power and resource distribution. For instance, in an Apache Kafka pipeline, Big D could monitor real-time data ingestion and dynamically adjust buffer sizes or replication strategies to prevent system overloads.
3	Variety	The diversity of data types and sources.	AI systems benefit from varied data (e.g., text, images, videos, IoT data) for comprehensive learning. Multimodal AI models can process and interpret diverse datasets, enhancing their insights and problem-solving capabilities.	Big D can classify, preprocess, and tag diverse data types using tools like Snowflake, Trifacta, and Talend, ensuring accurate preparation for machine learning pipelines. Big D could automate the tagging of video data for computer vision models and text data for use in natural language processing (NLP), ensuring diverse datasets are ready for AI analysis.

Table 5. *Cont.*

#	Component	Definition	Relation to AI	Potential Impact of AI Tools Like Big D
4	Veracity	The accuracy, quality, and trustworthiness of data.	High-quality, accurate data are essential for AI in terms of making reliable predictions and decisions. AI-driven data validation and cleaning techniques help to ensure that data are trustworthy, reducing the risk of biased or incorrect outputs.	Big D can use AI-driven validation tools such as Talend Data Quality, Alation, and IBM InfoSphere to clean and validate data automatically, ensuring that AI outcomes are based on reliable data. For example, Big D can detect and correct data anomalies in real time, enhancing data integrity.
5	Value	The usefulness and actionable insights derived from data.	AI extracts valuable patterns, correlations, and anomalies from Big Data, transforming raw information into meaningful insights. This allows organizations to make data-driven decisions with greater precision and speed.	Big D can work with tools like Datameer, AWS Redshift, and RapidMiner to extract actionable insights from datasets, maximizing the value derived from Big Data. Big D could identify new market trends in a business intelligence system or recommend pricing strategies based on historical data and market dynamics.
6	Validity	The accuracy and correctness of data used for a specific purpose.	AI models depend on relevant and correctly labeled training data. AI performs continuous data validation and updating, maintaining its relevance to and accuracy in AI applications.	Big D can continuously validate and update data through tools like Apache NiFi and Google Cloud Data Fusion, ensuring that AI models are always fed with valid data. In healthcare, Big D could ensure that patient data are accurate and up to date, supporting predictive models in real time.
7	Visualization	The effective representation of data insights and lineage.	Visualization helps in making AI models explainable, traceable, and auditable by clearly representing data flows, transformations, and results. This transparency is critical for regulatory compliance and stakeholder trust.	Big D, integrated with tools like Tableau, Snowflake, and Talend, can generate real-time visualizations of data flows and insights, enhancing transparency and decision-making. For example, Big D could provide visual dashboards that map data lineage, showing how data are transformed and used in AI models.
8	Viability	Practicality, feasibility, and sustainability of using data.	Sustainable and well-maintained data infrastructures are crucial for the long-term success of AI initiatives. AI models require reliable data access and consistent updates to remain effective.	Big D ensures the continuous viability of data pipelines by monitoring their health and suggesting optimizations. Tools like Google Cloud, AWS S3, and Snowflake can support Big D in maintaining viable data pipelines. Big D could dynamically allocate resources in cloud environments to ensure data pipelines remain operational under high loads.
9	Volatility	The rate of change and unpredictability of data.	AI models should be adaptable to changes in data distributions and patterns, maintaining their predictive power in dynamic environments.	Big D can monitor data streams for changes and adjust AI models to ensure accuracy. Tools like Splunk and Snowflake help to manage these fluctuations. For example, Big D could adjust the processing of workflows in response to sudden spikes in data, ensuring that insights remain relevant despite volatility.
10	Vulnerability	Ethical considerations and risks related to data security and privacy.	Ethical AI development involves ensuring fairness, avoiding bias, and addressing societal consequences. AI frameworks increasingly incorporate fairness, accountability, and transparency (FAT) principles.	Big D can ensure compliance with ethical guidelines using AI frameworks for FAT and privacy-preserving techniques like differential privacy. Big D could monitor AI systems for biases and suggest corrective actions, ensuring ethical decision-making in systems handling sensitive data.

Big Data and AI are now inseparable. AI supports better decision-making, business process optimization, and the discovery of new opportunities. However, it is important to use AI tools correctly and to the best of their ability. Organizations must embrace new technologies and keep adjusting in order to be competitive. Future studies will focus on the further improvement of the proposed spectrum of Vs framework. The knowledge base

of the Big D app will be upgraded as new tools and solutions appear on the market. Currently, the bot only talks to the user. However, it can potentially run its own additional AI models in its background, understand and generate images, and search the web. While this capability is already there, the program recently proposed by OpenAI, SearchGPT [73], can change the landscape drastically.

The preliminary results are as follows:

This study has introduced an expanded version of the traditional 4 Vs framework of Big Data, evolving it into the innovative "spectrum of Vs", which incorporates ten critical dimensions: volume, velocity, variety, veracity, value, validity, visualization, variability, volatility, and vulnerability. This expanded framework addresses the increasingly complex landscape of Big Data, considering in particular the rapid advancements in artificial intelligence (AI) and the rising prominence of Large Language Models (LLMs).

Through a comprehensive review of current Big Data tools and practices, as well as an in-depth exploration of AI's ongoing and potential impacts on this field, this study has demonstrated how the "spectrum of Vs" framework deepens our understanding of Big Data management in the context of AI-driven analytics. Furthermore, the research has examined AI's transformative role in Big Data analytics and highlighted how existing AI tools, particularly the RAG-based AI-driven "Big D" analytical bot, can enhance the efficiency and depth of insight extraction from vast and complex datasets.

This study answered all research questions stated in the introduction:

RQ1: The proposed "spectrum of Vs" framework has deepened the understanding of Big Data management by integrating additional dimensions that reflect contemporary challenges and opportunities in AI-driven analytics. This new framework provides a more comprehensive lens through which to examine and address the complexities of modern data ecosystems.

RQ2: The research has elucidated how AI is already transforming Big Data analytics through advancements in tools and methodologies that enable more precise, faster, and scalable data processing and analysis. The study also outlines how AI tools, such as LLMs and other machine learning algorithms, continue to evolve and contribute to the field.

RQ3: The introduction of the "Big D" analytical bot has demonstrated how RAG-based AI agents can significantly improve the efficiency and depth of insight extraction. "Big D" accelerates the processing of vast datasets and enables more nuanced and comprehensive analyses, facilitating better decision-making and strategic planning.

In conclusion, the integration of AI, particularly through frameworks like the "spectrum of Vs" and tools like "Big D", marks a significant leap forward in the capability to manage, analyze, and derive actionable insights from Big Data. This research paves the way for further exploration into the intersection of Big Data and AI, offering a robust foundation for future studies to advance the state of the art in this dynamic field.

6. Limitations and Implications

Practical Big Data management requires a holistic approach that incorporates state-of-the-art technologies, robust data governance, and collaborative systems. AI and blockchain are two emerging technologies that have the potential to tackle the challenges posed by Big Data effectively. Artificial intelligence (AI) has the potential to enhance the analysis of data by automating the identification of patterns. In contrast, Blockchain technology has the potential to improve the security and integrity of data. AI systems must be capable of interpreting data autonomously, extracting valuable insights from them, and facilitating decision-making. Robust data governance guarantees the integrity of data, their security, and compliance with industry regulations. Open data initiatives facilitate the dissemination and exchange of information across diverse industries.

Nonetheless, implementing federated learning in the healthcare industry enables multiple locations to utilize shared models without jeopardizing privacy as they abstain from sharing raw data. This approach facilitates the use of collective intelligence. By utilizing these breakthroughs, organizations can leverage Big Data to make informed decisions that

enhance the overall welfare of society. This includes the enhancement of healthcare delivery, urban planning, and supporting sustainable practices. Research has demonstrated that data-based choices can result in improved health outcomes, enhanced resource management efficiency, and a reduced environmental impact. The evolution of data analytics from descriptive and diagnostic to predictive and prescriptive reveals the complete capabilities of Big Data analysis. To expedite the verification process, it is imperative to consolidate the coverage and assertion data obtained from the regression runs. This will help to identify the most effective starting places for debugging, as shown in (Figure 20).

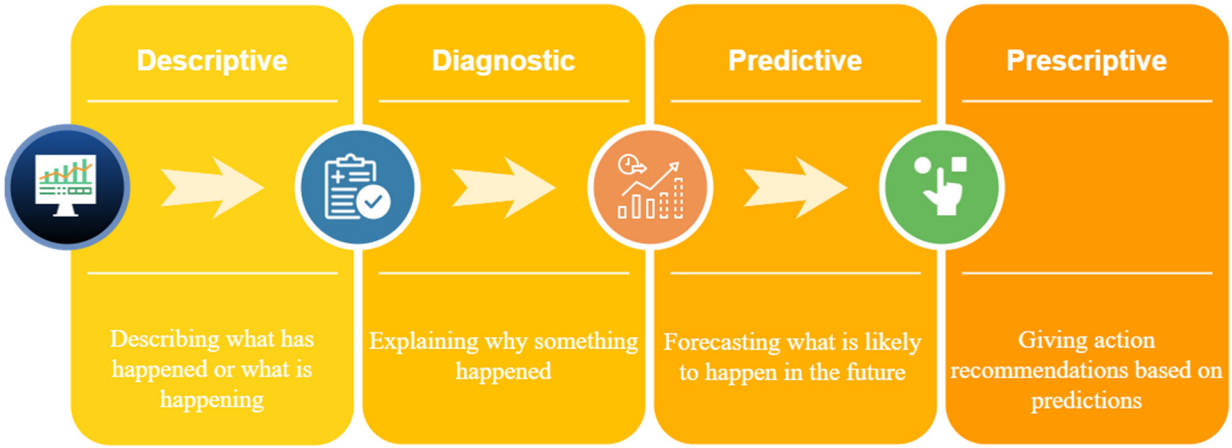


Figure 20. Analytics evolution maximizes Big Data potential.

Below is shown the comparative and ablation study analysis. We evaluate the features and capabilities of the proposed Big D bot against those of the current Big Data analytics technologies to assess its efficacy. Table 6 presents a comparison between Big D and traditional methods.

Table 6. Comparison of Big D with traditional and AI-driven Big Data tools.

Features	Big D	Hadoop	Apache Flink	Google BigQuery	IBM Watson
Real-time processing (RTP)	AI-driven, RAG-based, fast insights	Batch processing, lower real-time capabilities	True streaming, optimized for re-al-time	Real-time, cloud-oriented	Advanced real-time analytics
Scalability (S)	Cloud-based, dynamic scaling	Distributed file system, highly scalable	Scalable, batch/stream processing	Automatically scalable, cloud-based	Highly scalable with advanced AI capabilities
Fault tolerance (FT)	AI-driven integrity, which needs further validation	High fault tolerance, distributed recovery	Lightweight fault tolerance	Automatic replication, strong recovery	Advanced AI-driven fault tolerance
Advanced analytics (AA)	LLM-driven in-sights, predictive analytics	Requires manual optimization	Stream processing, manual optimizations	Managed analytics with SQL support	AI-driven in-sights, predictive analytics
Secondary validation needed (SV)	Production scalability and fault tolerance tests	Well-established, widely validated	Needs validation for very large data	Validated for cloud-scale applications	Validated in real time, mission-critical applications

While the actual weights are under consideration, the following formula can be used to estimate the best tool to use:

$$\text{Total Score} = \sum Wi * \text{Si} = (W_{RTP} \times S_{RTP}) + (W_S \times S_S) + (W_{FT} \times S_{FT}) + (W_{AA} \times S_{AA}) + (W_{SV} \times S_{SV}), \quad (8)$$

where Wi is a weight assigned to feature, $\sum Wi = 1$ and Si is a score assigned each feature.

The nature of the proposed app is such that it has an advanced ChatGPT model at its core, which already guarantees the universality of Big D responses as ChatGPT targets

artificial general intelligence (AGI), where the AI model is supposed to respond to any ethical topic. RAG architecture—where the model can accept documents from the users and strengthen GPT results by combining knowledge of the Large Language Model (LLM) with the custom user documentation—is used with AI agent Big D, running on threads and created through OpenAI Platform, and can be fine-tuned regarding specific goals, accept documentation from the backend that will be added to its knowledge, adapt to a particular type of agent, and follow a specially crafted prompt. Altogether, these factors guarantee the universality of the proposed tool.

Big D utilizes ChatGPT along with latent semantic indexing (LSI) for its analysis [74]. This offers significant advantages over other AI data analysis tools. Unlike traditional lexical matching systems that often fail to connect a query with semantically related content, ChatGPT combined with LSI captures deeper contextual relationships. This allows the system to understand not just literal matches but also conceptual links between words, improving the accuracy of the analysis. For example, while a basic keyword search might miss connections between "fishing" and terms like "rod" or "bait", LSI, aided by ChatGPT's understanding, identifies these as part of the same semantic field, offering a much richer and more nuanced understanding of the data.

ChatGPT, with its advanced NLP capabilities, further enhances this by generating coherent responses based on the understanding of entire contexts, making it capable of sophisticated conversation and analysis. Integrating these techniques makes your application superior in handling unstructured data, enabling it to classify social media posts even when the exact keywords are not present. The combination of LSI's dimensionality reduction through singular value decomposition (SVD) and ChatGPT's contextual language model ensures that your app can deliver more accurate, meaningful, and context-aware results than traditional AI tools relying solely on keyword matching or superficial language models. This fusion of statistical and generative AI methods brings out the best of both worlds, making your app more flexible, insightful, and powerful for diverse use cases.

7. Discussions and Future Work

7.1. Ethical and Privacy Implications

The growing capacity of artificial intelligence to efficiently handle and examine large quantities of data raises substantial ethical and privacy issues that are becoming increasingly prominent. Data ownership, user consent, and privacy rights become paramount when corporations collect large amounts of data. It is essential that data collection and processing adhere strictly to rigorous ethical standards. In healthcare, it is imperative to meticulously oversee the process of anonymizing patient data in order to safeguard patient confidentiality and extract valuable insights.

7.2. Impact on Employment and Skills

Combining artificial intelligence and Big Data can shift paradigms in numerous industries. Nonetheless, it also presents notable obstacles, particularly regarding employment. AI-powered automation can potentially cause job dislocation, particularly with regard to positions involving repetitive data processing duties. Nonetheless, this transition also presents prospects for novel professional positions requiring proficient technical expertise in artificial intelligence and data analytics. It is crucial to allocate resources to projects focusing on providing individuals with new skills and enhancing their existing ones. Education systems and business training programs must adapt to provide the workforce with the essential skills needed to succeed in an economy heavily influenced by artistic intelligence. Individuals can successfully adapt to and assume new positions resulting from technological progress by cultivating a culture that emphasizes ongoing education.

7.3. Advances and Challenges in Big Data Management

This paper presents the "spectrum of Vs" paradigm, which extends the conventional four Vs of Big Data (volume, velocity, variety, and veracity) by incorporating six new

dimensions: value, validity, visibility, viability, volatility, and virtue. This extended paradigm offers a more intricate comprehension of Big Data, encompassing its complex characteristics and various difficulties. One of the major obstacles in managing Big Data is verifying the accuracy and truthfulness of the data. As the amount of data increases, ensuring its accuracy and reliability becomes increasingly challenging. Data normalization and outlier identification techniques are essential to maintaining dataset integrity. In addition, a solid and adaptable infrastructure capable of managing large amounts of data is crucial. As emerging technologies, Apache Flink and Google BigQuery [66] provide effective real-time data processing and analytics solutions, outperforming conventional frameworks such as Hadoop.

7.4. Role of LLMs in Big Data Analytics

In research, LLMs have a crucial function in using the potential of Big Data. These models can analyze large datasets and offer useful insights that inform decision-making processes. However, the need for large organizations to oversee these models raises concerns about their accessibility and management. It is crucial to democratize AI technology by making LLMs available to a wider variety of users while ensuring robust governance. This paper emphasizes the ability of LLMs to revolutionize diverse industries, ranging from healthcare to banking, by facilitating more precise forecasts and tailored suggestions. The advancement of multimodal artificial intelligence, which can incorporate text, visual, and audio input, improves the analytical capacities of language and learning models, allowing for more complete and nuanced insights.

7.5. The Costs of Implementing AI and Big Data Technologies in Practical Settings

At this point, we only can provide a rough estimate for this category as all companies try to keep such information private. The actual costs will depend on location and other factors, including possible sanctions. Table 7 provides a hypothetical estimate of the costs.

Table 7. Hypothetical costs of AI-Big Data integration in practical settings.

Cost Category	Description	Cost Estimate, In Thousands of USD ($)
AI model development	Custom AI model development.	75–150
API integration	The use of pre-built AI platforms with limited usage.	10–30 annually
Big data pipeline setup	Setup for data ingestion, processing, and ETL processes.	30–60
Frontend/backend development	The development of the mobile app interface, backend, and visualization tools.	80–120
Cloud services and hosting	Use of cloud services (AWS, Google Cloud, or Azure).	3–10+ monthly
Data storage	Cloud storage for volumes of data.	1–5+ monthly
Computing power	Use of cloud-based computing resources (e.g., virtual machines, GPUs).	2–10 monthly
AI tools	API usage fees for AI services.	15–50 annually
Big Data processing tools	Open-source tools (e.g., Hadoop, Apache Flink).	Free/open source
Data collection	Acquiring data for AI model training and analysis.	20–50
Data cleaning and labeling	Data preparation for AI model training.	30–100
Model maintenance and retraining	The ongoing retraining of AI models.	15–40 annually

Table 7. *Cont.*

Cost Category	Description	Cost Estimate, In Thousands of USD ($)
Monitoring and support	The monitoring of data pipelines and AI model performance.	10–25 annually
AI/ML engineers and data scientists	Hiring AI/ML engineers or data scientists.	150–300 annually
DevOps/cloud engineers	Engineers for managing cloud services and infrastructure.	100–180 annually
User training and adoption	User training and onboarding.	10–20
Third-party library licenses	Open-source libraries and minimal paid tool usage.	5–15 annually

7.6. Role of LLMs in Big D Development and Usage

Every product directly depends on its provider when using a remote service or a third-party platform/library. Big D currently relies on Open AI and uses its family of products, including GPT and other AI models. While it is possible to use other models like Google's Gemini or Sonnet through API, dependency on the service provider(s) will still be there. Therefore, the app is limited to the rates, functionalities, and policies the companies allow. As an example, Table 8 presents some limitations of OpenAI (on 12 October 2024)

Table 8. Tokens per minute (TPM), requests per minute or day (RPM/RPD), and other limits.

Model	Token Limits	Request and Other Limits	Batch Queue Limits
gpt-4o	800,000 TPM	5000 RPM	100,000,000 TPD
gpt-4o-mini	4,000,000 TPM	5000 RPM	40,000,000 TPD
gpt-3.5-turbo	4,000,000 TPM	5000 RPM	40,000,000 TPD
gpt-4	80,000 TPM	5000 RPM	5,000,000 TPD
gpt-4-turbo	600,000 TPM	5000 RPM	40,000,000 TPD
gpt-4o-realtime-preview	80,000 TPM	5000 RPM	N/A
text-embedding-3-small	5,000,000 TPM	5000 RPM	100,000,000 TPD

As a limitation of the study, we would like to add that Big D, in some rare cases, demonstrates hallucinations: when several documents were submitted in a row, Big D started confusing new topics with the previous ones and might have demonstrated results of previously analyzed documents instead of the latest it remembered. Researchers are not completely sure of how to deal with these issues and whether they can eventually be fixed.

7.7. Future Directions

The future of integrating AI and Big Data relies on ongoing innovation and collaboration. Quantum computing can significantly enhance data processing skills by enabling the study of highly massive datasets at an unparalleled speed. Furthermore, progress in developing data governance frameworks is essential for guaranteeing ethical and effective data management.

The cooperation of academia, industry, and regulatory agencies is crucial in tackling obstacles and capitalizing on AI and Big Data's potential. By promoting a culture of knowledge sharing and implementing the most effective strategies, all parties involved can collaboratively navigate the intricate aspects of this rapidly changing area and build on advancements.

Author Contributions: Conceptualization, H.B.A., Y.K. and A.H.A.; methodology, H.B.A., Y.K., A.H.A., J.M. and J.J.L.; software, J.M., Y.K. and J.J.L.; validation, H.B.A. and J.J.L.; formal analysis, J.M.; investigation, H.B.A., Y.K. and A.H.A.; resources, H.B.A.; data curation, J.M.; writing—original draft preparation, H.B.A., Y.K., A.H.A. and J.M.; writing—review and editing, J.J.L. and H.B.A.; visualization, Y.K., J.J.L. and J.M.; supervision, H.B.A.; project administration, H.B.A.; funding acquisition, H.B.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Internal Research Support Program. (IRSPG202202). It is important to note that the USA team involved in this project was not funded by any China-based grants.

Data Availability Statement: All data are based on references.

Acknowledgments: The authors gratefully acknowledge the financial support from Wenzhou-Kean University and Kean University.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Laney, D. 3D data management: Controlling data volume, velocity and variety. *META Group Res. Note* **2001**, *6*, 1.
2. Stern, N. Computers: From Eniac to Univac: As inventors, Eckert and Mauchly were clear successes, but as entrepreneurs they had some difficult times. In *IEEE Spectrum*; IEEE: New York, NY, USA, 1981; Volume 18, pp. 61–69. [CrossRef]
3. DiNucci, D. Fragmented future. *Print* **1999**, *53*, 32.
4. O'Reilly, T. What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. O'Reilly Media, 2005. Available online: https://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html (accessed on 3 November 2024).
5. Kaplan, A.M.; Haenlein, M. Users of the world, unite! The challenges and opportunities of social media. *Bus. Horiz.* **2010**, *53*, 59–68. [CrossRef]
6. Ahmed, N.; Barczak, A.L.C.; Susnjak, T.; Rashid, M.A. A comprehensive performance analysis of Apache Hadoop and Apache Spark for large scale data sets using *HiBench*. *J. Big Data* **2020**, *7*, 110. [CrossRef]
7. Diebold, F.X. On the Origin(s) and Development of the Term 'Big Data'. 2012. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2152421 (accessed on 3 November 2024).
8. Ashton, K. That 'Internet of Things' Thing. *RFID J.* **2009**, *22*, 97–114.
9. Child, R.; Gray, S.; Radford, A.; Sutskever, I. Generating long sequences with sparse transformers. *arXiv* **2019**, arXiv:1904.10509.
10. Sutskever, I.; Martens, J.; Hinton, G.E. Generating text with recurrent neural networks. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), Bellevue, WT, USA, 28 June–2 July 2011; pp. 1017–1024.
11. Khemasuwan, D.; Colt, H.G. Applications and challenges of AI-based algorithms in the COVID-19 pandemic. *BMJ Innov.* **2021**, *7*, 387–398. [CrossRef]
12. David, E. Walmart Bets on Multiple AI Models with New Wallaby LLM. 2024. Available online: https://venturebeat.com/ai/walmart-bets-on-multiple-ai-models-with-new-wallaby-llm/ (accessed on 12 October 2024).
13. Sifted Team. How Amazon Is Using AI To Become the Fastest Supply Chain in the World. 2024. Available online: https://sifted.com/resources/how-amazon-is-using-ai-to-become-the-fastest-supply-chain-in-the-world/ (accessed on 12 October 2024).
14. Cleveland Clinic. How AI Is Being Used to Benefit Your Healthcare. 2024. Available online: https://health.clevelandclinic.org/ai-in-healthcare (accessed on 12 October 2024).
15. Daley, S. AI in Healthcare: Uses, Examples and Benefits. 2024. Available online: https://builtin.com/artificial-intelligence/artificial-intelligence-healthcare (accessed on 12 October 2024).
16. Mayo Clinic Press Editors. AI in Healthcare: The Future of Patient Care and Health Management. 2024. Available online: https://mcpress.mayoclinic.org/healthy-aging/ai-in-healthcare-the-future-of-patient-care-and-health-management/ (accessed on 12 October 2024).
17. Sergiienko, B. Why Generative AI in Banking Is a Secret Weapon: Your Blueprint for Implementation. 2024. Available online: https://masterofcode.com/blog/generative-ai-in-banking (accessed on 12 October 2024).
18. Reynolds, K. COVID-19 Increased the Use of AI. Here's Why It's Here to Stay. 2024. Available online: https://www.weforum.org/agenda/2021/02/covid-19-increased-use-of-ai-here-s-why-its-here-to-stay/ (accessed on 12 October 2024).
19. Appen. The 2020 State of AI and Machine Learning Report. 2020. Available online: https://www.appen.com/whitepapers/the-state-of-ai-and-machine-learning-report (accessed on 12 October 2024).
20. Batra, R. Database Management Systems and Tools. In *SQL Primer*; Apress: Berkeley, CA, USA, 2018. [CrossRef]
21. Kaur, H.; Rani, V.; Kumar, M.; Sachdeva, M.; Mittal, A.; Kumar, K. Federated learning: A comprehensive review of recent advances and applications. *Multimed. Tools Appl.* **2024**, *83*, 54165–54188. [CrossRef]
22. Kumar, S.; Lim, W.M.; Sivarajah, U.; Kaur, J. Artificial Intelligence and Blockchain Integration in Business: Trends from a Bibliometric-Content Analysis. *Inf. Syst. Front.* **2023**, *25*, 871–896. [CrossRef]
23. Sood, V.; Chauhan, R.P. Archives of quantum computing: Research progress and challenges. *Arch. Comput. Methods Eng.* **2024**, *31*, 73–91. [CrossRef]

24. Abdalla, H.B. A brief survey on big data: Technologies, terminologies and data-intensive applications. *J. Big Data* **2022**, *9*, 107. [CrossRef]
25. Gupta, D.; Rani, R. A study of big data evolution and research challenges. *J. Inf. Sci.* **2019**, *45*, 322–340. [CrossRef]
26. Lee, I. Big data: Dimensions, evolution, impacts, and challenges. *Bus. Horiz.* **2017**, *60*, 293–303. [CrossRef]
27. Wen, J.; Zhang, Z.; Lan, Y.; Cui, Z.; Cai, J.; Zhang, W. A survey on federated learning: Challenges and applications. *Int. J. Mach. Learn. Cybern.* **2023**, *14*, 513–535. [CrossRef]
28. Preskill, J. Quantum computing 40 years later. In *Feynman Lectures on Computation*; CRC Press: Boca Raton, FL, USA, 2023; pp. 193–244.
29. Rayhan, A.; Shahana, R. Quantum Computing and AI: A Quantum Leap in Intelligence. In *AI Odyssey: Unraveling the Past, Mastering the Present, and Charting the Future of Artificial Intelligence*; NotunKhabar, 2023; Available online: https://www.amazon.com/Odyssey-Unraveling-Mastering-Artificial-Intelligence/dp/B0CCXLCGDM (accessed on 12 October 2024).
30. Aminul, M. Impact of Big Data Analytics on Digital Marketing: Academic Review. *J. Electr. Syst.* **2024**, *20*, 786–820.
31. Sargiotis, D. Integrating AI and Big Data in Virtual Infrastructures: Transforming Educational Landscapes for the Future. 2024. Available online: https://discovery.researcher.life/article/integrating-ai-and-big-data-in-virtual-infrastructures-transforming-educational-landscapes-for-the-future/980542e1f19838588443702eac62087f (accessed on 12 October 2024).
32. Tosi, D.; Kokaj, R.; Roccetti, M. 15 years of Big Data: A systematic literature review. *J. Big Data* **2024**, *11*, 73. [CrossRef]
33. Arachchige, A.S.P.M.; Chebaro, K.; Jelmoni, A.J. Advances in large language models: ChatGPT expands the horizons of neuroscience. *STEM Educ.* **2023**, *3*, 263–272. [CrossRef]
34. Cui, Y.; Yang, Z.; Yao, X. Efficient and effective text encoding for chinese llama and alpaca. *arXiv* **2023**, arXiv:2304.08177.
35. Hong, D.; Li, C.; Zhang, B.; Yokoya, N.; Benediktsson, J.A.; Chanussot, J. Multimodal artificial intelligence foundation models: Unleashing the power of remote sensing big data in earth observation. *Innovation* **2024**, *2*, 100055. [CrossRef]
36. Dida, H.A.; Chakravarthy DS, K.; Rabbi, F. ChatGPT and Big Data: Enhancing Text-to-Speech Conversion. *Mesopotamian J. Big Data* **2023**, *2023*, 31–35. [CrossRef]
37. Yeasir Fahim, J. Mastering the Art of AI Language: An In-Depth Exploration of Prompting Techniques and Their Influence on Model Performance. 2024. Available online: https://digital.kenyon.edu/cgi/viewcontent.cgi?article=1031&context=dh_iphs_ss (accessed on 12 October 2024).
38. Rashid, A.; Baloch, N.; Rasheed, R.; Ngah, A.H. Big data analytics-artificial intelligence and sustainable performance through green supply chain practices in manufacturing firms of a developing country. *J. Sci. Technol. Policy Manag.* 2024; *ahead of print*.
39. Sardi, A.; Sorano, E.; Cantino, V.; Garengo, P. Big data and performance measurement research: Trends, evolution and future opportunities. *Meas. Bus. Excell.* **2023**, *27*, 531–548. [CrossRef]
40. Zhang, Z.; Han, X.; Liu, Z.; Jiang, X.; Sun, M.; Liu, Q. ERNIE: Enhanced language representation with informative entities. *arXiv* **2019**, arXiv:1905.07129.
41. Ram, B.; Verma, P. Artificial intelligence AI-based Chatbot study of ChatGPT, Google AI Bard and Baidu AI. *World J. Adv. Eng. Technol. Sci.* **2023**, *8*, 258–261.
42. Sundu, M.; Yasar, O.; Findikli, M.A. Data-driven innovation: Digital tools, artificial intelligence, and big data. In *Organizational Innovation in the Digital Age*; Springer International Publishing: Cham, Switzerland, 2022; pp. 149–175.
43. Bormida, M.D. The big data world: Benefits, threats and ethical challenges. In *Ethical Issues in Covert, Security and Surveillance Research*; Emerald Publishing Limited: Leeds, UK, 2021; pp. 71–91.
44. Raubenheimer, J. Big data in academic research: Challenges, pitfalls, and opportunities. In *Big Data in Education: Pedagogy and Research*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 3–37.
45. Raban, D.R.; Gordon, A. The evolution of data science and big data research: A bibliometric analysis. *Scientometrics* **2020**, *122*, 1563–1581. [CrossRef]
46. Chae, B.K. A General framework for studying the evolution of the digital innovation ecosystem: The case of big data. *Int. J. Inf. Manag.* **2019**, *45*, 83–94. [CrossRef]
47. Nadal, S.; Romero, O.; Abelló, A.; Vassiliadis, P.; Vansummeren, S. An integration-oriented ontology to govern evolution in big data ecosystems. *Inf. Syst.* **2019**, *79*, 3–19. [CrossRef]
48. Bonner, S.; Kureshi, I.; Brennan, J.; Theodoropoulos, G. Exploring the evolution of big data technologies. In *Software Architecture for Big Data and the Cloud*; Morgan Kaufmann: Cambridge, MA, USA, 2017; pp. 253–283.
49. Gu, D.; Li, J.; Li, X.; Liang, C. Visualizing the knowledge structure and evolution of big data research in healthcare informatics. *Int. J. Med. Inform.* **2017**, *98*, 22–32. [CrossRef]
50. Salminen, V.; Ruohomaa, H.; Kantola, J. Digitalization and big data supporting responsible business co-evolution. In *Advances in Human Factors, Business Management, Training and Education, Proceedings of the AHFE 2016 International Conference on Human Factors, Business Management and Society, Walt Disney World, FL, USA, 27–31 July 2016*; Springer International Publishing: Cham, Switzerland, 2017; pp. 1055–1067.
51. Halevi, G.; Moed, H.F. The evolution of big data as a research and scientific topic: Overview of the literature. *Res. Trends* **2012**, *1*, 2.
52. Camacho, J.; Macia-Fernandez, G.; Diaz-Verdejo, J.; Garcia-Teodoro, P. Tackling the big data 4 vs for anomaly detection. In Proceedings of the 2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Toronto, ON, Canada, 27 April–2 May 2014; IEEE: New York, NY, USA, 2014; pp. 500–505.

53. Anuradha, J. A brief introduction on Big Data 5Vs characteristics and Hadoop technology. *Procedia Comput. Sci.* **2015**, *48*, 319–324.
54. Jeong, C. Generative AI service implementation using LLM application architecture: Based on RAG model and LangChain framework. *J. Intell. Inf. Syst.* **2023**, *29*, 129–164.
55. OpenAI. GPT-4o System Card. 2024. Available online: https://cdn.openai.com/gpt-4o-system-card.pdf (accessed on 9 August 2024).
56. Marshall, A.; Mueck, S.; Shockley, R. How leading organizations use big data and analytics to innovate. *Strategy Leadersh.* **2015**, *43*, 32–39. [CrossRef]
57. Alaskar, T.H.; Alsadi, A.K.; Aloulou, W.J.; Ayadi, F.M. Big Data Analytics, Strategic Capabilities, and Innovation Performance: Mediation Approach of Organizational Ambidexterity. *Sustainability* **2024**, *16*, 5111. [CrossRef]
58. Sundberg, L.; Holmström, J. Democratizing artificial intelligence: How no-code AI can leverage machine learning operations. *Bus. Horiz.* **2023**, *66*, 777–788. [CrossRef]
59. Widad, E.; Alaoui, I.E.; Gahi, Y. Data quality in the era of big data: A global review. In *Big Data Intelligence for Smart Applications*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 1–25.
60. Misra, B.B.; Langefeld, C.; Olivier, M.; Cox, L.A. Integrated omics: Tools, advances and future approaches. *J. Mol. Endocrinol.* **2019**, *62*, R21–R45. [CrossRef]
61. Balducci, B.; Marinova, D. Unstructured data in marketing. *J. Acad. Mark. Sci.* **2018**, *46*, 557–590. [CrossRef]
62. Intersoft Consulting. General Data Protection Regulation (GDPR). Available online: https://gdpr-info.eu/ (accessed on 10 August 2024).
63. California Consumer Privacy Act (CCPA). Available online: https://www.oag.ca.gov/privacy/ccpa (accessed on 10 August 2024).
64. IBM. IBM Watson Health Introduces New Opportunities for Imaging AI Adoption. 2021. Available online: https://newsroom.ibm.com/2021-11-30-IBM-Watson-Health-Introduces-New-Opportunities-for-Imaging-AI-Adoption (accessed on 10 August 2024).
65. Microsoft Research. AI For Good Lab. 2024. Available online: https://www.microsoft.com/en-us/research/group/ai-for-good-research-lab/ (accessed on 10 August 2024).
66. BigQuery Overview. Available online: https://cloud.google.com/bigquery/docs/introduction (accessed on 10 August 2024).
67. Get Started with TensorFlow. Available online: https://www.tensorflow.org/ (accessed on 10 August 2024).
68. IBM InfoSphere Information Server. Available online: https://www.ibm.com/information-server (accessed on 10 August 2024).
69. Apache NiFi Documentation. Available online: https://nifi.apache.org/documentation/ (accessed on 10 August 2024).
70. Data Quality Solutions. Available online: https://www.talend.com/products/data-quality/ (accessed on 10 August 2024).
71. Abdalla, H.B.; Awlla, A.H.; Kumar, Y.; Cheraghy, M. Big Data: Past, Present, and Future Insights. In Proceedings of the 2024 Asia Pacific Conference on Computing Technologies, Communications and Networking, Chengdu, China, 26–27 July 2024; pp. 60–70.
72. NSF. NSF 24-589: Computer and Information Science and Engineering: Core Programs. 2024. Available online: https://new.nsf.gov/funding/opportunities/computer-information-science-engineering-core-programs/nsf24-589/solicitation (accessed on 12 August 2024).
73. SearchGPT Prototype. Available online: https://openai.com/index/searchgpt-prototype/ (accessed on 10 August 2024).
74. Deshmukh, A.; Hegde, G.; Lathi, R.; Govikarn, S. A literature survey on latent semantic indexing. In Proceedings of the International Conference on Computing 2012, Maui, HI, USA, 30 January–2 February 2012; p. 100.