computers

MDPI

Article

Opportunities and Challenges in Big Data Analytics for Decision Making: An Integrated Framework

Wafa' Za'al Alma'aitah [1,*], Fatima N. AL-Aswadi [2], Addy Quraan [3], Nader Abdel Karim [1], Hussein Alahmer [4] and Mohamad Y. Mustafa [5,*]

1 Department of Intelligent Systems, Faculty of Artificial Intelligence, Al-Balqa Applied University, Al-Salt 19117, Jordan; nader.salameh@bau.edu.jo
2 Institute of Computer Science and Digital Innovation, UCSI University, Kuala Lumpur 56000, Malaysia; fatima.nadeem@ucsiuniversity.edu.my
3 Department of Basic Sciences, Faculty of Science, The Hashemite University, Zarqa 13133, Jordan; addy@hu.edu.jo
4 Department of Automated Systems, Faculty of Artificial Intelligence, Al-Balqa Applied University, Al-Salt 19117, Jordan; hussein@bau.edu.jo
5 Department of Building, Energy and Material Technology, UiT the Arctic University of Norway, 9037 Tromsø, Norway
* Correspondence: wafaa_maitah@bau.edu.jo (W.Z.A.); mohamad.y.mustafa@uit.no (M.Y.M.)

Abstract

Big Data Analytics (BDA) has evolved from a predominantly technical batch function into a socio-technical capability integrating cloud-native platforms, stream processing, Lakehouse architecture, machine learning operations (MLOps), visualization, governance, and managerial judgment. This paper proposes an integrated BDA decision-making framework developed through a structured conceptual synthesis of research on data platforms, analytical capabilities, decision processes, organizational readiness, technology adoption, governance, and responsible artificial intelligence. The framework comprises seven interconnected stages: data sources, ingestion and integration, storage and platform, processing, analytics and artificial intelligence, visualization and interpretation, and decision, action, and learning. Governance, human oversight, organizational readiness, task characteristics, and continuous feedback influence all stages. Key implementation requirements include data quality, interoperability, security, privacy, scalability, cost, explainability, bias, skills, and sustainability. The proposed configurable reference architecture links technical integration, task–analytics fit, governance assurance, human judgment, and organizational readiness with decision quality and organizational outcomes. Organizational size and maturity, sectoral risk, decision criticality, technological context, and regulatory environment are defined as boundary conditions for future empirical validation.

Keywords: big data analytics; decision making; conceptual framework; stream analytics; cloud analytics; data lakehouse; MLOps; data governance

Check for updates

Academic Editor: Teng Zhou

Received: 6 August 2026
Revised: 31 August 2026
Accepted: 2 September 2026
Published: 4 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Big Data Analytics (BDA) has become a central organizational capability for transforming heterogeneous, high-volume, and rapidly generated data into evidence for operational and strategic decisions. Foundational research connected BDA with data management, business intelligence, and decision support [1–3], while subsequent studies expanded the technological basis through cloud computing, stream processing, advanced analytics, and modern data platforms [4–6]. Capability-oriented research further shows that the value of

BDA depends on organizational context, analytical competence, strategic alignment, and risk management [7,8]. Accordingly, BDA should be examined not merely as a collection of technologies, but as a coordinated socio-technical system in which data, infrastructure, analytical methods, governance, and human judgment jointly shape decision outcomes.

BDA is used to identify issues, predict trends, find opportunities, mitigate risks, individualize services, track performance and aid strategic planning [9–11]. However, the use of analytical tools does not necessarily create decision value. Contemporary systems combine batch and real-time processing, cloud-native services, Lakehouse platforms, machine learning, MLOps, and governance [12–15]. Despite this, there is existing work that is scattered across technology-oriented architectures, capability models, model-lifecycle approaches, governance frameworks and research on decision-support. This means that there is little theoretical understanding of the joint influence of technical integration, governance control, human judgment, organizational readiness, task characteristics and institutional constraints on decision quality and organizational outcomes. The research gap addressed in this paper is, thus, not the lack of another data-to-decision pipeline, but the lack of an integrative socio-technical explanation of architecture, adoption conditions, accountability mechanisms and organizational learning in one configurable decision lifecycle.

This paper has five contributions. First, it establishes the conceptual relationship between BDA capability and organizational decision making by outlining mechanisms that can influence the analytical usefulness and decision quality when BDA capabilities are technically integrated. Second, it identifies four design dimensions (processing, platform, analytics and governance), and places human judgment, organizational readiness, task characteristics and institutional conditions as complementary determinants of effective use. Third, it differentiates implementation requirements and constraints such as data quality, interoperability, security, privacy, explainability, skills, cost and sustainability and explains how security and privacy function on related but different risk pathways. Fourth, it does not consider any single technology superior, and it explains the architectural transformation from batch centric systems to real-time, cloud-native, Lakehouse and MLOps. Fifth, it suggests a framework for BDA decision making that is configurable through cross-cutting governance, human oversight, organizational readiness and ongoing feedback, as well as design propositions, boundary conditions and evaluation domains for future empirical testing. Section 2 provides a conceptual background and approach to framework development, Section 3 defines the dimensions of the framework design, Section 4 discusses implementation requirements, Section 5 presents the architectural rationale and design propositions, Section 6 presents and differentiates the proposed framework and finally Section 7 concludes with future research directions.

2. Conceptual Background and Framework Development

2.1. BDA Capability and Decision Making

The relationship between BDA and organizational decision making has been investigated in several studies. Shmueli and Koppius [16] detailed the significance of predictive analytics in information systems research and Chen and Zhang [17] talked about data-intensive applications and large-scale analytics requirements. Gupta and George [18] viewed BDA capability as a company resource and Ghasemaghaei et al. [19] presented that the data quality and analytical competence affect decision quality. Recent research has related BDA to firm performance, start-up outcomes, and performance measurement effectiveness [8,11,20].

Algorithms are not the only source of decision-making value. Janssen et al. [21] identified governance, transparency and data quality as essential enablers for public-sector big data decisions and Müller et al. [22] observed the organizational factors that influence

the adoption of BDA. Ghasemaghaei [23] demonstrated that there is a relationship between analytical insight quality and decision-making benefits. According to Huynh et al. [7] the ability of BDA is conceptually fragmented in the literature and Rodepeter et al. [8] state that BDA adoption may lead to performance improvement as well as implementation risks.

Big data is often expressed in terms of four fundamental attributes: volume, velocity, variety, and, in some cases, veracity and value [6,24–26]. The 5Vs model has been used as the working model for the paper, as it helps address the decision-making, data quality, governance, and organizational value focus of the paper. Volume is the quantity of data being generated, velocity is the speed at which data is being generated and processed, variety is the different forms of structured, semi-structured and unstructured data, veracity is the trustworthiness and quality, and value is the usefulness of data for making decisions. There are other models derived from the 5Vs, for example models with variability, validity, volatility, visualization, or vulnerability; these models are considered domain-specific extensions of the 5Vs, not replacements.

The different types of BDA include descriptive, diagnostic, predictive, and prescriptive analytics [13,25,27]. Descriptive Analytics is used to understand what has transpired, Diagnostic Analytics is used to understand why, Predictive Analytics is used to predict what might occur next, and Prescriptive Analytics is used to recommend actions. Predictive Analytics is closely related to Data Science, as they all involve models making data-driven decisions, as reported by Provost and Fawcett [28]. Chatterjee et al. [11] also highlighted BDA's ability to aid in forecasting and performance improvement, and Wang et al. [29] demonstrated its potential in a healthcare organization. BDA can be useful in decision making, when the analytical outputs are timely, relevant, interpretable and attuned with the organizational goals [9,14,30]. Phan and Baird also demonstrate that there is a relationship between BDA use and the effectiveness of the performance management system [20]. Thus, there needs to be a link between technical ability, managerial application, and organizational learning and decision value. Table 1 shows the main decision-making phases and the types of BDA support that are relevant to each.

Table 1. Mapping of decision-making phases and BDA support.

Decision-Making Phase	Main Analytical Need	Suitable Big Data Support	Illustrative Scholarly Support
Intelligence	Recognizing problems, opportunities, anomalies and environmental signals.	Data collection, integration, dashboards, descriptive analytics and early-warning indicators.	Chen et al. [1]; Chatterjee et al. [11]
Design	Formulating alternatives, scenarios and possible courses of action.	Predictive analytics, simulation, data mining and exploratory modeling.	Shmueli and Koppius [16]; Jamarani et al. [5]
Choice	Selecting the most suitable alternative under uncertainty and constraints.	Prescriptive analytics, optimization, ranking models and evidence-based comparison.	Elgendy and Elragal [10]; Cao [31]
Implementation	Deploying decisions, monitoring outcomes and adjusting actions.	Streaming analytics, operational dashboards, MLOps and performance monitoring.	Ghasemaghaei [23]; Amou Najafabadi et al. [32]
Learning and control	Evaluating decision outcomes and improving future decision rules.	Governance, audit trails, feedback loops, model monitoring and accountability mechanisms.	Abraham et al. [33]; Xia et al. [34]

2.2. Conceptual Synthesis and Framework Development

This study adopts a structured conceptual synthesis design, not a systematic-review or meta-analytic design. It aims to integrate existing theories from BDA capability, data architecture, decision support, data governance, MLOps, responsible AI and technology-adoption research into a unified explanatory model. The synthesis does not attempt to be comprehensive with respect to all BDA publications. Rather, literature was selected purposively based on conceptual relevance, methodological influence and recency in the following six streams: (1) BDA capability and organizational value, (2) data architecture and distributed processing, (3) analytics and decision support, (4) governance, security, privacy and responsible AI, (5) MLOps/DataOps and post-deployment monitoring, and (6) technology adoption, organizational readiness and human judgment. Foundational studies were included if they were used to define constructs or architectural principles that are still relevant in the latest studies, and recent peer-reviewed studies were given priority to reflect the latest advancements in cloud-native analytics, Lakehouse architecture, stream processing, AI governance, and organizational readiness.

Four analytical steps were used in the synthesis. To begin, studies were filtered to include only those that were directly relevant to the mechanisms of data, analytics, governance, and organizational decision making; studies that focused only on individual algorithms and did not have implications for architecture, adoption, governance, or decision use were not considered to be core conceptual evidence. Second, the literature selected was coded using the constructs and relationships of integration, latency, analytical fit, data quality, explainability, accountability, human oversight, organizational capability, readiness, and feedback. Third, those findings that converged were clustered into design dimensions and implementation requirements; those that disagreed or were dependent on context were left as boundary conditions instead of being resolved by generalized claims. Fourth, the resulting constructs were synthesized into testable design propositions and then mapped to the proposed lifecycle framework. This procedure makes the logic from literature to constructs, from constructs to propositions and from propositions to framework design transparent and recognizes that empirical validation is needed before causal effects can be established.

2.3. Socio-Technical Mechanisms and Contextual Conditions

The framework sees the use of BDA as a socio-technical process, where technological capability is not enough. The interpretation and challenge of analytical outputs, and their translation into action, are influenced by human judgment, while organizational conditions influence the availability of data, skills, leadership support, financial resources, and cross-functional routines and processes, and institutional conditions influence acceptable use of data, accountability, and regulatory compliance. These are not external to architecture but are in interaction with it. For instance, a technically correct real-time model might not be of much use to users who do not have the domain knowledge, who do not have the authority to make the decision, or who need explanations that the model does not provide. On the other hand, if the data quality is not good or the architecture is not suitable for the latency and reliability needs of the decision, then the organizational readiness is not enough. The proposed framework thus posits that decision value is a result of fit between technical architecture, analytical capability, human/organizational readiness, task demands, and governance conditions, not simply a result of technology deployment [7,18,21,23,30,35,36].

3. Design Dimensions of Big Data Analytics

There are multiple technical and organizational approaches that have been discussed for classifying BDA. Hashem et al. [37] demonstrated the potential of cloud computing in the storage and processing of big data, by using elastic resources, and Assuncao et al. [38] examined the role of cloud computing in big data processing. Kune et al. [39] explored big data computing systems anatomy and Yaqoob et al. [40] discussed how big data evolved from the initial groundwork to prospects. Subsequent work related BDA to IoT environments and modern stream-processing systems [41,42].

3.1. Processing Dimension

Alternatively, processing-oriented approaches can be seen as a transition from MapReduce to memory-based and stream-based systems. Dean and Ghemawat [43] presented the model of MapReduce for simple processing of large data sets and Sakr et al. [44] surveyed the methods for managing large data sets in cloud environments. This discussion is continued in more recent works by Henning and Hasselbring [45], Almeida et al. [46] and Marcu and Bouvry [42] that explore stream-processing frameworks, cloud deployment, time-series big data, and real-time analytics. Zaharia et al. [47] and Carbone et al. [48] also proposed Spark as a single engine for batch, streaming and interactive workloads and Flink as a stream and batch processing engine, respectively.

Processing-oriented approaches are about processing large-scale data. When data are gathered and analyzed periodically, e.g., daily reporting, historical analysis, and offline model training [24,26,49], batch processing is still useful. For instance, to process large datasets, Hadoop and MapReduce were popularly used to distribute work over clusters. Gandomi and Haider [25] stated that the kind of analytics method should be determined by the nature of the data and the decision problem. Thus, batch processing is suitable if timeliness is not an important requirement, but completeness and accuracy and repeatable analysis are.

When decisions are made based on continuous or near real-time data, stream processing is needed. These are fraud detection, cyber security monitoring, traffic management, logistics tracking, financial trading and online recommendation systems. Event-driven data processing has been demonstrated to enable low-latency analytics in scholarly works about Flink, Spark, and modern stream-processing frameworks [42,45–48]. Velocity is an important attribute of big data, and stream processing is important in decision contexts where the value of the decision diminishes with delay. Secure access, secure storage and privacy-preserving processing are also important challenges for real-time cloud analytics, as mentioned by Rajan and Vetriselvi [14].

3.2. Platform Dimension

The platform-based paradigm has evolved from distributed file systems and data warehouses to cloud platforms, data lakes, and Lakehouse architecture. Grolinger et al. [50] discussed Data Management in Cloud environments and Nargesian et al. [51] pointed out Data Lake Management challenges with respect to Discovery, Metadata, and governance. The Lakehouse concept was introduced as a new generation of open platforms for analytics and AI by Armbrust et al. [52]. More recent studies continue this discussion by comparing data lake, data warehouse and Lakehouse systems and reviewing Lakehouse architectures for time-series data [13,53].

Platform-oriented approaches are based on the platform that stores, manages and analyzes data. Traditional on-premises clusters offer control and come with a heavy price tag in hardware, maintenance and technical skills [26,54]. Cloud-native platforms, on the other hand, offer elastic resources, managed services, and analytics and machine learning

integration [14,55]. Akter et al. [30] noted that technology cannot be the only solution here as organizations need to be in sync with the business strategy and analytics capability.

Data Lakehouse architecture is a new one that tries to merge the best of both data warehouses and data lakes [13]. Data lakes offer flexibility to store all kinds of raw data, and data warehouses offer a structured, reliable, and optimized environment for analysis. In this context the Lakehouse systems are designed to help facilitate large-scale analytics, while providing better data management, reliability and query performance [13]. This architecture is crucial for decision making, as they must have trusted analytical outputs and flexible data storage [9,11].

3.3. Analytics Dimension

Approaches that leverage analytics include statistical learning, machine learning, deep learning, and explainable modeling. The study by Wu et al. [56] reviewed the data mining with big data and Qiu et al. [57] surveyed the machine learning for big data processing. Najafabadi et al. [58] conducted a related study on deep learning applications and challenges in BDA. Jordan and Mitchell [59] introduced machine learning as a discipline that makes predictions out of data, while Witten et al. [60] gave an account of the methodology of data mining and practical machine learning. This issue has recently been confirmed to still be relevant [61].

Analytics approaches are geared towards deriving knowledge from data. Descriptive and diagnostic analytics are used for reporting, monitoring and explaining [25,46]. Predictive analytics involves the use of models using statistical and machine learning approaches to predict future events, and prescriptive analytics involves making recommendations [11,28]. Different analytics can be associated with different decision needs, as analyzed by Elgendy and Elragal [10], and mapped to different decision-making stages. Later, Wang et al. [29] demonstrated the value of analytics capabilities in healthcare organizations and how they can be leveraged for better data utilization and healthcare services.

AI is increasingly integrated into BDA. Automated machine learning can minimize the work invested to build models and MLOps can help deploy, monitor and maintain models after building. Big data analytics and AI-based modeling and enterprise decision optimization are coupled, as highlighted in [31] by Cao. With AI-powered analytics however comes a greater requirement for explainability, data quality monitoring, accountability, and governance.

3.4. Governance Dimension

Governance-oriented approaches are necessary since BDA is about data ownership, metadata, access control, accountability and compliance. The need for decision rights and accountability mechanisms in data governance was pointed out by Khatri and Brown [62] and Otto [63] investigated data governance as a corporate function. Tallon et al. [64] connected information governance with business value and Abraham et al. [33] did a review on data governance in the information systems field. More recent reviews are extended to big data governance, data platforms, data spaces, data mesh and data fabric [61,65,66].

Governance-oriented approaches are aimed at reliable, secure, ethical and organizationally responsible data and analytics. BDA needs data quality management, metadata management, access control, privacy protection, security monitoring and compliance processes [14,26]. As stated by Mikalef et al., governance can be considered as part of the capability of the organization, as it is necessary to make the data resources trusted and useful for decision making. Huynh et al. [7] pointed out the importance of more clearly conceptualizing BDA capabilities.

AI governance is also a critical extension of big data governance, as many analytics systems are now incorporating machine learning and AI components [15]. In recent

empirical evidence, Xia et al. [34] shows that responsible AI governance mechanisms can enhance corporate performance. This strengthens the argument for considering governance as a value creation capability and not just a compliance obligation. The design dimensions are presented from three complementary angles: Table 2 shows the major tools and the different platforms categories, Table 3 shows the strengths and limitations of the approaches and Table 4 shows the relation between the different layers of the analytics lifecycle, the different techniques and the output of the decisions.

Table 2. Tools and platform categories in BDA.

Category	Representative Tools/Platforms	Main Use	Decision-Making Value
Distributed processing	Apache Spark, Hadoop MapReduce	Large-scale batch and distributed data processing	Supports analysis of large historical datasets
Stream analytics	Apache Kafka, Apache Flink, Spark Structured Streaming	Real-time data pipelines and continuous analytics	Supports fast operational and risk-related decisions
Cloud analytics	BigQuery, Redshift, Azure Synapse, Microsoft Fabric	Elastic storage, integration, and analytics services	Reduces infrastructure barriers and improves scalability
Lakehouse systems	Databricks, Delta Lake, Apache Iceberg, Apache Hudi	Unified data lake and warehouse capabilities	Supports flexible analytics and machine learning
Visualization and BI	Power BI, Tableau, Looker, MicroStrategy	Dashboards, reports, and visual exploration	Improves interpretability and communication
Machine learning and MLOps	H2O.ai, MLflow, Kubeflow, Azure ML	Model development, deployment, and monitoring	Supports predictive and prescriptive decisions
Governance tools	Microsoft Purview, Collibra, Apache Atlas	Data cataloging, lineage, quality, and access management	Improves trust, compliance, and accountability

Table 3. Strengths and limitations of BDA approaches.

Taxonomy Approach	Main Advantages	Main Disadvantages	Representative Scholarly Support
Processing-oriented approaches	They support large-scale data transformation, parallel execution and analysis of historical data. They are useful when organizations need to process massive datasets that cannot be handled by single-machine systems.	They may require complex cluster configuration, specialized technical skills and high resource consumption. Batch-oriented processing can also delay decisions when real-time response is required.	MapReduce introduced distributed batch processing [43], while recent stream-processing reviews emphasize the move toward continuous analytics [42].
Platform-oriented approaches	They provide scalable infrastructure for storage, computation, analytics services and deployment. Cloud platforms also reduce the need for organizations to own all hardware resources internally.	They can create dependence on providers, unpredictable costs, data sovereignty concerns and migration difficulties. Performance and governance also depend on platform configuration.	Cloud elasticity was framed by Armbrust et al. [67], while recent cloud-native Lakehouse work emphasizes portable, resilient platforms [68].

Table 3. *Cont.*

Taxonomy Approach	Main Advantages	Main Disadvantages	Representative Scholarly Support
Analytics-oriented approaches	They convert data into descriptive, predictive and prescriptive insights. They are valuable for classification, forecasting, anomaly detection, recommendation and decision support.	Their effectiveness depends on data quality, model selection, interpretability and domain expertise. Complex models may produce accurate output but weak explanations for decision makers.	Machine learning became central to analytics [59], and recent reviews highlight predictive analytics applications in big data environments [5].
Governance-oriented approaches	They improve trust, accountability, data quality, privacy protection and regulatory readiness. They also clarify ownership and responsibility across the analytics lifecycle.	They can slow implementation when governance procedures are heavy, fragmented or not aligned with organizational culture. Governance requires continuous coordination between technical and managerial teams.	Data governance clarifies decision rights and accountability [62], while recent work extends governance to data catalogs and responsible AI [34,69].

Table 4. BDA lifecycle: layers, techniques and decision outputs.

Lifecycle Layer	Typical Techniques or Functions	Representative Tools/Platforms	Decision Output
Data ingestion	Batch ingestion, event capture, sensor data capture and log collection.	Kafka, ETL/ELT pipelines, APIs and data connectors.	Timely access to operational and external signals [42,70].
Storage and organization	Distributed file storage, cloud object storage, data lakes and Lakehouse organization.	HDFS, cloud storage, Delta Lake, Apache Iceberg and Lakehouse platforms.	Reliable historical and Lakehouse-based data foundation for analysis and reporting [13,71].
Processing	Batch processing, iterative processing, in-memory computation and stream processing.	MapReduce, Spark, Flink and cloud processing services.	Faster transformation of raw data into analytical variables [45,47].
Analytics and modeling	Descriptive, diagnostic, predictive, prescriptive, machine learning and deep learning analytics.	R, Python ecosystems, H_2O.ai, ML platforms and MLOps pipelines.	Patterns, predictions, classifications and recommended actions [31,58].
Visualization and communication	Dashboards, scorecards, visual exploration and self-service analytics.	Power BI, Tableau, Looker and business intelligence platforms.	Understandable insights for managers and domain experts [20,72].
Governance and control	Metadata management, lineage, access control, data quality monitoring and compliance.	Data catalogs, governance platforms and policy-based access tools.	Trustworthy and accountable decision support [15,62].

4. Implementation Requirements and Constraints

Effective BDA design must address a set of interdependent implementation requirements. As shown in the work by Sivarajah et al. [73], challenges can be grouped into data, process, management, and infrastructure dimensions. A complementary study by Dem-

chenko et al. [74] discussed scientific data infrastructure challenges, while Katal et al. [75] summarized big data issues, technologies, and methods. The studies by Khan et al. [76], Elgendy and Elragal [27], and Oussous et al. [26] also show that big data challenges are not limited to storage and processing but include quality, analytics, visualization, and security. This earlier perspective has been complemented by recent big data research [52].

The implementation requirements in this section are grouped by data quality, scalability and cost, security, privacy, interoperability, explainability and bias, skills, responsible AI and sustainability. Security and privacy are considered as separate aspects: security relates to the protection of systems, data and services from unauthorized access, manipulation, disruption and loss; privacy relates to legitimate collection, linkage, inference, sharing and use of information about individuals or sensitive entities. This distinction is important because the controls, stakeholder perceptions, adoption effects and regulatory implications of the two dimensions are not the same. Table 5 offers a roadmap to connect each challenge category with the decision-making effect and the response explored in the subsections that follow.

Table 5. Section 4 taxonomy of BDA challenges and responses.

Taxonomy Challenge	Nature of the Challenge	Effect on Decision Making	Recommended Response
Data quality	Incomplete, noisy, inconsistent, duplicated, outdated, or biased data from multiple sources	Weakens forecasting, risk detection, reporting accuracy, and confidence in analytical outputs	Apply data profiling, cleaning, validation, lineage, metadata management, and continuous quality monitoring
Scalability and cost	Growing storage, compute, network, and real-time processing demand	Delays insights or increases operational cost when analytics workloads are not optimized	Use elastic architecture, workload optimization, lifecycle policies, cost monitoring, and appropriate platform selection
Security	Unauthorized access, manipulation, disruption, insecure interfaces, and attacks across distributed data and analytics infrastructure	Threatens confidentiality, integrity, availability, reliability, and willingness to depend on analytical systems	Use identity and access management, encryption, secure pipelines, monitoring, incident response, audit trails, and resilience controls
Privacy	Excessive collection, linkage, inference, re-identification, secondary use, or inappropriate sharing of sensitive data	Reduces stakeholder trust, constrains legitimate data use, and creates ethical and regulatory exposure	Apply purpose limitation, data minimization, privacy assessment, anonymization/pseudonymization, differential privacy, retention rules, and governed data sharing
Interoperability	Different systems, schemas, formats, tools, and repositories may not exchange data effectively	Produces fragmented evidence and weakens cross-functional decision making	Use APIs, metadata standards, semantic integration, data catalogs, common models, and lakehouse architecture
Explainability and bias	Complex models may be difficult to interpret and may reproduce biased patterns from data	Reduces trust, accountability, fairness, and acceptance of analytics-supported decisions	Use explainable AI, bias testing, model documentation, validation, human review, and fairness monitoring

Table 5. *Cont.*

Taxonomy Challenge	Nature of the Challenge	Effect on Decision Making	Recommended Response
Skills gap	Organizations may lack the technical, analytical, domain, and managerial skills required for BDA	Limits the ability to interpret results and transform analytics into action	Develop data literacy, interdisciplinary teams, training programs, and communication between technical and business users
Responsible AI and sustainability	AI-driven analytics raises ethical, social, accountability, and environmental concerns	May create harmful, opaque, costly, or unsustainable decision systems	Adopt responsible AI governance, sustainability metrics, workload optimization, human oversight, and transparent accountability mechanisms

4.1. Data Quality

Data quality issues are completeness, consistency, duplication, noise, timeliness and contextual inaccuracy. Batini et al. [77] surveyed data quality assessment and enhancement methodologies and Cai and Zhu [78] investigated data quality issues in the big data era. Based on this discussion, Merino et al. [79] introduced a data-quality-in-use model for big data, Taleb et al. [80] conducted a survey of big data quality and Ji et al. [54] synthesized quality-assurance technologies for big data applications. These studies prove that Data Quality problems can impact the accuracy of large-scale analytics and diminish the output of decisions.

One of the most enduring problems in BDA is data quality ([24,25,81]). The data is typically generated by various internal and external data sources, including sensors, transactions systems, logs, social media, cloud services, etc., and is often heterogeneous. The data can be incomplete, inconsistent, duplicated, noisy, or hard to integrate, as Oussous et al. [26] discussed. According to Wixom et al. [9] the value of analytics is dependent on the capacity to deliver useful and trusted business insight. Thus, bad data quality can lead to less reliable dashboards, models, forecasts and recommendations.

The impact of poor data quality is particularly important in decision-making contexts, as decision makers can use analytical results as an objective fact [9,10,55]. Analytical systems can provide false conclusions if data are biased, incomplete, or out-of-date. Chatterjee et al. [11] states that while analytics can help with decision making and forecasting, this can only happen if the underlying data and analytical process is of good quality. According to Mikalef et al. [35], BDA capability needs data resources that can be converted to organizational value.

4.2. Scalability and Cost

Scalability issues have been discussed with respect to distributed systems, cloud computing and workload management. For instance, Dinh et al. [82] discussed issues in mobile cloud computing and Ranjan [83] streamed big data processing in the cloud. Buyya et al. [84] outlined the directions for the next generation of cloud computing and Toosi et al. [85] presented the issues of interoperability and resource management in multi-cloud. Buyya et al. [55] recently relate scalability with energy efficiency and sustainable cloud resource management.

Scalability is the capacity of a BDA system to process growing amounts of data, speeds, and types of data and to perform analytics without compromising performance [24,26,86]. Traditional systems can become problematic if data volumes are growing quickly or if

multiple users and applications are utilizing analytics services simultaneously. According to Zaharia et al. [47], modern processing engines are designed to handle large-scale data engineering, machine learning and streaming workloads. Carbone et al. [48] highlighted stateful computations over bounded and unbounded streams, which is essential for real-time scalability.

Cost is closely related to scalability. Although cloud-native analytics can significantly lower the up-front investment in infrastructure, there can be unpredictable operational costs if organizations are not monitoring storage, computation, data movement, and model training [55]. According to the survey conducted by Rajan and Vetriselvi [14], cloud BDA also needs secure storage and access methods, which can lead to complexity in the system. Cost management is thus not just a financial concern, but also a design concern related to architecture, workload scheduling, data lifecycle policies, and governance [26,55].

4.3. Security and Privacy as Distinct Risk Dimensions

Security and privacy are similar but are conceptually different requirements in BDA. Security is about maintaining confidentiality, integrity, availability, authentication, authorization, and operational resilience of distributed data and analytics infrastructures. Privacy is concerned with the collection, combination, inference, retention and use of personal and sensitive information in a way that is "legitimate, proportionate, transparent, and compatible with the expectations of the stakeholders and regulation" [14,87–92]. Failure to differentiate between the two can make it confusing about how each works and can result in incomplete governance responses.

Security risks include unauthorized access, credential compromise, malicious insiders, data tampering, model or pipeline attacks, insecure interfaces and service disruption. These risks can have a direct impact on the reliability of the system and on the confidence of the organization in relying on the analytical results. Examples of appropriate responses are identity and access management, encryption, segmentation, secure software and pipeline practices, continuous monitoring, incident response, audit trails, and resilience planning [14,74,87–89]. From a decision perspective, security is important to ensure the integrity and availability of evidence and analytical services on which organizational action is based.

Even if a system is technically secure, there are privacy risks involved. Sensitive attributes may be revealed in the process of over-collection, data linkage, re-identification, secondary use or model inference. This means that privacy needs to be ensured through purpose limitation, data minimization, lawful and transparent processing, anonymization or pseudonymization where appropriate, differential privacy, federated approaches, retention controls, and governance of data sharing [14,90–92]. Unlike a traditional security failure, the impact of the decision making is a loss of stakeholder trust, restricted access to data, regulatory risk and loss of legitimacy of the analytical decisions, even if they are accurate. The framework thus conceptualizes security and privacy as distinct governance processes which can have varying impacts on adoption, trust and organizational outcomes.

4.4. Interoperability

Big data environments bring together a mix of data models, storage systems, APIs, cloud services and analytics platforms, which is a challenge to interoperability. When there are many datasets in an organization, Goods-style data discovery is required, as described by Halevy et al. [93]. Furthermore, Dong and Srivastava [94] surveyed big data integration and Lenzerini [95] gave foundations for data integration from different sources. The work of Stonebraker et al. [96] and Doan et al. [97] also demonstrates that integration, schema

matching and metadata management are still key challenges in data-intensive systems. This line of discussion has been further developed in recent studies [45].

Interoperability is the capability of various systems, tools, platforms, data formats, schemas and organizational units to exchange and use data effectively [26,91]. BDA environments typically include databases, data lakes, data warehouses, data streaming, visualization systems and machine learning platforms. Moreover, Harby and Zulkernine [13] talked about the Lakehouse architecture in the context of the fragmentation of data lakes and data warehouses. Interoperability is also related to governance, as metadata, catalogs, lineage, and access policies are necessary to make data understandable and reusable [14,35].

The interoperability challenge has a direct impact on decision making, as the lack of interoperability results in incomplete views of the problem [10,12]. For instance, a business decision might need customer, sales, logistics, financial and social media data; these data might be in various formats and systems. Wixom et al. [9] demonstrated that the value of analytics is dependent on linking data to valuable insights. Furthermore, Chatterjee et al. [11] demonstrated that analytics can impact forecasting and performance, which means that data needs to be integrated and readily available throughout the organization.

4.5. Explainability and Bias

When BDA is used to help make decisions that impact people, services, or allocation of resources, there is a special concern for explainability and bias. The methods to explain black-box models were surveyed by Guidotti et al. [98] and a more general taxonomy of explainable artificial intelligence was presented by Barredo Arrieta et al. [99]. Rudin [100] supported the use of interpretable models in high-stakes decisions and Mehrabi et al. [101] provided an overview of bias and fairness in machine learning. Other studies have also revealed that explanations, audits, and fairness practices should be incorporated into the analytics development process, and not just as an afterthought once deployed [102–104].

Explainability is becoming more important, since many BDA systems rely on models based on machine learning and deep learning, which can be hard to explain to users [12,15]. Users may reject a model or use it inappropriately if they do not understand the rationale behind the recommendation. According to Provost & Fawcett [28], data-driven decision making is not only about interpreting data to make decisions but also involves critical understanding of the analytical outputs. In the same way, Elgendy and Elragal [10] pointed out that analytics must be used in the decision phases, but not in lieu of managerial judgment.

Bias is another big issue. BDA can be affected by bias introduced in the data, sampling bias, missing data, proxy variables, data measurement error, or bias in the model design [15]. Rajan and Vetriselvi [14] have also talked about privacy-preserving learning and secure analytics, which also concerns the responsible handling of sensitive data. The impact of bias on decision making is that an apparently accurate model can still have a negative impact on groups or perpetuate past inequities. Hence, fairness assessment, auditing of the models, representative data and human oversight are required [12,15].

4.6. Skills Gap

There is a technical and managerial skills gap. Davenport and Patil [105] highlighted the need for data science skills, and Chiang et al. [106] related data science education with the development of data analytics capability. According to Debortoli et al. [107], skills needed for analytics work include text mining, data preparation, data interpretation

and domain understanding. Other research suggests that these five elements (talent, culture, infrastructure, managerial support, and MLOps capability) collectively influence the capacity to effectively leverage analytics [12,18,36,108].

The skills gap is an organizational challenge which can lead BDA to failure in providing decision-making value [7,30,35]. Technical skills are needed for big data projects, including data engineering, cloud computing, distributed processing, machine learning, visualization, cybersecurity and data governance. Akter et al. [30] stated that it is important to match BDA capability with business strategy to improve performance. In the study by Mikalef et al. [35], the idea of combining tangible, human and intangible resources was put forward for analytics capabilities.

One of the most frequently encountered issues is the lack of human capability to use the tools and infrastructure that have been acquired [30,35,92]. Data literacy is also important for decision makers to understand the outputs of analytics and ask the right questions and to not rely completely on automation. Huynh et al. [7] notes that the literature on BDA capabilities is still in the conceptual stage and needs to be further developed, such as the understanding of human and organizational capabilities. Wixom et al. [9] found that the value of analytics lies in the use of the insights in practice.

4.7. Responsible AI and Sustainability

In today's world, AI and sustainability are key considerations in analytics systems. Jobin et al. [109] analyzed the AI ethics guidelines of the world and Floridi and Cowls [110] put forward principles for responsible AI. The literature on responsible AI governance is further enriched by more recent work by Papagiannidis et al. [15], who categorize responsible AI governance into three aspects: structural, relational and procedural practices. A similar study by Mittelstadt [111] cautioned that principles without mechanisms of implementation are not enough. Schwartz et al. [112] also discussed sustainability with the introduction of Green AI, and Buyya et al. [55] associated cloud resource management with energy efficiency and sustainability.

Responsible AI has become an important challenge due to the increased use of machine learning, automated decision systems, and AI-assisted analytics in BDA systems [12,15]. Responsible AI must be transparent, explainable, fair, private, secure, accountable and human-centric. In line with the study by Papagiannidis et al. [15], responsible governance of AI can be thought of as involving structural, relational, and procedural practices. Rajan and Vetriselvi [14] further stressed the need for security and privacy-preserving mechanisms to be incorporated in BDA in the cloud.

Sustainability is also gaining in importance in BDA. Large-scale storage, regular data transfers, constant model training and real-time processing can be very resource-intensive and energy-consuming [55]. Zaharia et al. [47], Almeida et al. [46] and Carbone et al. [48] demonstrate the capabilities of modern analytics ecosystems for supporting large-scale and streaming workloads, which, however, need to be managed efficiently. Thus, workload optimization, resource scheduling, storage lifecycle management and energy-aware architecture are essential to sustainability [26,55]. As described above, the utility, cost and complexity of the challenge responses vary. The key responses are therefore assessed in Table 6 in terms of their pros and cons, prior to the implementation priorities section.

After evaluating possible responses, the remaining issue is the order in which organizations should address them. Table 7 presents a priority matrix that connects each major challenge with its primary risk, decision-making consequence, and recommended priority response.

Table 6. Evaluation of response strategies for BDA challenges.

Challenge Area	Possible Response	Advantages	Limitations
Data quality	Data profiling, cleansing, validation, metadata management and lineage tracking.	Improves the reliability of dashboards, models and forecasts. It also reduces misleading results caused by missing, duplicate or inconsistent data [49,77].	Quality improvement can be costly and require continuous effort because big data sources change frequently and may be generated outside organizational control.
Scalability and cost	Elastic cloud resources, workload monitoring and cost-aware architecture design.	Allows organizations to scale storage and computation according to demand rather than fixed capacity, especially when cloud resources are managed efficiently [55].	Elasticity may make costs unpredictable when workloads, queries or data movement are not controlled.
Security	Encryption, identity and access control, secure pipelines, monitoring, incident response, and resilience controls.	Protects the confidentiality, integrity, availability, and reliability of data and analytical services [14,74,87–89].	Strong controls can add cost and operational complexity and must be calibrated to risk and criticality.
Privacy	Data minimization, purpose limitation, anonymization/pseudonymization, differential privacy, federated approaches, and governed sharing.	Supports legitimate use of sensitive data and can strengthen stakeholder trust and regulatory acceptability [14,90–92].	Privacy protection can reduce analytical granularity or data availability and may require context-specific trade-offs.
Interoperability	Schema mapping, data integration tools, common data models and semantic alignment.	Supports integration across heterogeneous systems and reduces fragmentation between operational and analytical data [69,94].	Integration remains difficult when sources have different formats, meanings, update cycles and ownership structures.
Explainability and bias	Explainable AI methods, fairness assessment, model documentation and human review.	Improves transparency and allows decision makers to understand and challenge model outputs [34,99].	Explanations may be incomplete, and bias mitigation requires social, organizational, and technical assessments.
Skills gap	Training, interdisciplinary teams and collaboration between domain experts, data engineers and analysts.	Improves the ability to translate analytical results into practical decisions. Big data and BI skills require both technical and business knowledge [20,107].	Training requires time and resources, and organizations may still face shortages in advanced data engineering and analytics expertise.
Responsible AI and sustainability	AI governance, model audit, energy-aware design and responsible data practices.	Supports accountable analytics and reduces environmental and ethical risks. Responsible AI and Green AI emphasize governance, computational cost and sustainability [15,112].	Responsible AI and sustainability metrics are still developing, and organizations may struggle to balance accuracy, speed, cost and ethical obligations.

Table 7. Priority matrix for implementing BDA challenge responses.

Challenge Area	Primary Risk to Analytics	Decision-Making Consequence	Priority Response
Data quality	Incomplete, inconsistent, duplicated or outdated data.	Incorrect forecasts, misleading dashboards and weak confidence in analytical results.	Establish quality rules, profiling, validation and continuous monitoring [49,77].
Security	Unauthorized access, manipulation, service disruption, weak authentication, and compromised analytical pipelines.	Unreliable or unavailable evidence, operational interruption, and lower confidence in analytics-dependent decisions.	Prioritize access control, encryption, monitoring, secure development, incident response, and resilience according to system criticality [14,74,87–89].
Privacy	Over-collection, re-identification, secondary use, inappropriate inference, or sharing of sensitive data.	Reduced stakeholder trust, restrictions on data use, legal exposure, and lower legitimacy of analytics-supported decisions.	Apply privacy-by-design, minimization, purpose limitation, privacy-preserving analytics, retention controls, and governed data sharing [14,90–92].
Interoperability	Fragmented systems, inconsistent schemas and semantic mismatch between data sources.	Partial decision views and difficulty combining operational, customer and external data.	Use common metadata, APIs, semantic integration and catalog-supported governance [69,97].
Explainability and bias	Opaque models, biased training data and difficulty justifying automated recommendations.	Unfair, untrusted or non-actionable decisions in high-impact domains.	Apply model documentation, explainable methods, bias testing and human review [34,98].
Scalability and cost	Processing delays, uncontrolled cloud spending and performance bottlenecks.	Slow decisions, delayed reporting and poor operational responsiveness.	Adopt elastic architecture, workload monitoring and cost-aware platform design [45,113].
Sustainability	High energy consumption caused by large-scale storage, training and processing.	Conflict between analytics value and environmental responsibility.	Apply green AI principles, efficient modeling and workload optimization [55,112].

4.8. Organizational, Task, and Institutional Conditions

Implementation constraints are not only technical. Organizational readiness determines whether leadership support, financial resources, data literacy, specialist skills, cross-functional coordination, and change-management capacity are sufficient to absorb BDA into routine decision processes. Task characteristics also matter; decisions differ in time sensitivity, uncertainty, reversibility, explainability requirements, and the consequences of error. Institutional conditions—including sector-specific regulation, professional standards, public accountability, data-sovereignty rules, and stakeholder expectations—can further limit which data, models, automation levels, and deployment arrangements are appropriate. These contextual factors help explain why the same BDA architecture may produce different adoption and decision outcomes across organizations and sectors [7,21,30,35,36]. They are therefore treated in the proposed framework as boundary conditions that configure, rather than replace, the seven technical and decision stages.

5. Architectural Rationale: From Traditional to Modern BDA

The differences between traditional and modern BDA are also evident in the shift from warehouse-centric architectures to data lakes, Lakehouse platforms, and cloud-native analytics. For instance, Inmon [114] gave guidelines for the design of a data warehouse, and Kimball and Ross [115] explained dimensional modeling for analytical systems. The data lake concepts and architectures were reviewed by Hai et al. [86], and Begoli et al. [81] talked about the transition from data lakes to more managed analytics platforms. Recent research by Harby and Zulkernine [13] and Pohl et al. [53] indicates that Lakehouse is an emerging and significant trend for analytical data management.

This section provides an explanation of the architectural reasons for the shift from conventional to contemporary BDA environments. The comparison is categorized into two parts, namely the traditional approach and the modern approach. On-premises clusters, batch processing, Hadoop/MapReduce and limited governance are traditional approaches [24,26,101]. Some modern approaches include cloud-native analytics, real-time analytics processing, Lakehouse architecture, MLOps/DataOps, and integrated governance [12–14].

5.1. Traditional Architectural Logic

Relational databases, data warehouse, extract/transform/load pipelines, and scheduled reporting all influenced traditional approaches. The work of Chaudhuri and Dayal [116] surveyed data warehousing and OLAP technology and Stonebraker et al. [117] presented the MapReduce and parallel database controversy. Pavlo et al. [118] made a comparison between MapReduce and parallel DBMS systems, and Dean and Ghemawat [43] demonstrated how MapReduce made large-scale data processing simpler. These studies clarify the reasons behind the traditional architecture being effective but inflexible, with poor latency and governance integration. This discussion has been expanded in more recent work in the context of modern analytics platforms [36].

Traditional BDA solutions have been designed for on-premises clusters, distributed file systems, batch-processing, and dedicated technical teams [24,26,53]. These methods were crucial, as they allowed organizations to handle large volumes of data that were beyond the capacity of traditional databases. As Gandomi and Haider [25] discussed, early methods of big data were oriented towards the challenges of dealing with high-volume, high-velocity and high-variety data. But the traditional methods involved complex infrastructure management and did not provide much flexibility for real-time analytics [47,48,104].

5.1.1. On-Premises Clusters

On-premises clusters give you control over infrastructure, but they come with capital investment, capacity planning and special administration. Cloud computing altered this model with its elastic resources and service delivery, as reported by Armbrust et al. [67]. Buyya et al. [55] and Rajan and Vetriselvi [14] further discuss sustainable cloud resource management and secure BDA in the cloud, respectively, while Henning and Hasselbring [45] present scalable deployments of stream-processing. Zhang et al. [119] surveyed the issues of cloud computing research and Assuncao et al. [38] described how cloud infrastructures can be used to process workloads of big data processing.

On-premises clusters are on-premises computing infrastructure that is managed by the organization. In this model, the servers, storage, networking, software installation, maintenance, backup and security are all handled in-house. Such infrastructure supported the first generation of BDA, as shown in earlier work by Chen et al. [24] and Oussous et al. [26]. Similarly, Rajan and Vetriselvi [14] pointed out that cloud environments raise

new security and privacy issues, while on-premises systems also have disadvantages, including high capital expenses, limited elasticity, and maintenance complexity.

5.1.2. Batch Processing

Batch processing remains important for historical analytics and large-scale periodic computation. The work by Dean and Ghemawat [43] introduced MapReduce as a batch-oriented processing model, while Sakr et al. [44] surveyed MapReduce-family systems. For example, Kune et al. [39] explained how batch processing fits within broader big data computing architectures. Recent studies by Marcu and Bouvry [42] and Almeida et al. [46] extend this older batch-processing discussion by comparing batch, stream, and hybrid processing models in modern big data environments.

Batch processing is one of the earliest and most common processing models in BDA [24,26,120]. It processes large volumes of stored data at scheduled intervals and is useful for historical reporting, periodic analysis, and offline model development. As demonstrated by Elgendy and Elragal [27], analytics methods are often selected according to the type of data and analytical objective. However, batch processing may be insufficient for decisions that require immediate response, such as fraud detection, intrusion detection, or real-time personalization [46,48,70].

5.1.3. Hadoop/MapReduce

Hadoop and MapReduce enabled scalable data storage and processing, but later studies identified limitations in latency and iterative workloads. For example, Shvachko et al. [71] described the Hadoop Distributed File System, while Zaharia et al. [121] introduced Spark to support iterative jobs through resilient distributed datasets. Bu et al. [122] proposed HaLoop to improve iterative MapReduce applications. Recent scholarship also confirms the continuing relevance of this issue [6].

Hadoop and MapReduce represent a major stage in the development of BDA [4,24,26]. Hadoop provides distributed storage and processing capabilities, while MapReduce divides large jobs into smaller tasks that can be processed across clusters. Gandomi and Haider [25] discussed MapReduce as one of the foundational technologies associated with BDA. Although Hadoop remains historically important, modern workloads increasingly require faster, more flexible, and more interactive processing engines, as demonstrated by later work on Spark and contemporary analytics platforms [46,47].

5.1.4. Limited Governance

Traditional big data environments often treated governance as a separate administrative task rather than an integrated design requirement. As illustrated by Weber et al. [123], data governance requires clear roles and decision rights. Within this context, Khatri and Brown [62] emphasized accountability in data governance, while Abraham et al. [33] synthesized governance mechanisms in information systems research. Recent studies by Bližnák et al. [65], Gieß and Hutterer [66], and Castro et al. [61] extend this literature by connecting data governance with modern distributed data platforms, data spaces, data mesh, data fabric, and ontology-based governance for big data.

Traditional big data environments often suffer from limited or fragmented governance [14,26]. Data may be distributed across multiple clusters, files, repositories, and teams without consistent metadata, access control, quality monitoring, or lineage tracking. Mikalef et al. [35] showed that governance and organizational capabilities are necessary for transforming data resources into business value. Related research by Papagiannidis et al. [15] showed that governance becomes even more important when analytics systems include AI-based decision support.

5.2. Modern Architectural Logic

Modern approaches combine cloud services, real-time processing, AI-assisted analytics, governance automation, and lifecycle management. The study by Kreps et al. [70] introduced Kafka for high-volume log processing, while Carbone et al. [48] described Flink for stream and batch processing. Subsequent work by Zaharia et al. [47] presented Spark as a unified engine. Recent scholarly studies by Almeida et al. [46], Henning and Hasselbring [45], and Marcu and Bouvry [42] update these older system-oriented contributions by reviewing data stream frameworks, benchmarking stream processing in cloud environments, and synthesizing big data stream-processing models and applications.

Modern BDA approaches extend the traditional ecosystem by emphasizing elasticity, real-time processing, integrated data management, machine learning operations, automation, and governance. Cloud-native analytics support flexible infrastructure, stream processing supports real-time decisions, and lakehouse architecture supports unified data management [13,55]. A later study by Kreuzberger et al. [12] showed that MLOps provides operational practices for deploying and monitoring machine learning models. Rajan and Vetriselvi [14] also emphasized secure and privacy-preserving analytics in cloud environments.

5.2.1. Cloud-Native Analytics

Cloud-native analytics relies on elastic infrastructure, managed services, and scalable storage–compute separation. The work by Hashem et al. [37] reviewed big data on cloud computing, while Grolinger et al. [50] examined data management in cloud environments. Buyya et al. [84] discussed future-generation cloud computing, and Toosi et al. [85] reviewed multi-cloud interoperability and resource management. Recent studies by Buyya et al. [55], Rajan and Vetriselvi [14], and Chatterjee et al. [11] extend this discussion by linking cloud analytics with sustainability, security, privacy, decision making, forecasting, and performance.

Cloud-native analytics uses cloud infrastructure and managed services to store, process, analyze, and visualize large-scale data [14,55]. Instead of purchasing and maintaining all infrastructure internally, organizations can use scalable computing resources, storage services, managed databases, and analytics platforms. Buyya et al. [55] also emphasized that cloud computing must address resource management, energy efficiency, and sustainability. Rajan and Vetriselvi [14] explained that secure access control, secure storage, and privacy-preserving processing are essential for cloud-based BDA.

5.2.2. Real-Time Processing

Real-time processing is central to modern decision support in fraud detection, cybersecurity, smart cities, logistics, and online services. As reviewed by Kolajo et al. [124], stream analysis requires methods that can process continuous high-velocity data. In this direction, Henning and Hasselbring [45] benchmarked stream processing frameworks in cloud environments, while Marcu and Bouvry [42] reviewed big data stream processing for modern data-intensive systems.

Real-time processing allows organizations to analyze data as it is generated or shortly after it is received. Stream-processing systems, including Kafka, Flink, and Spark Structured Streaming, support event-driven data pipelines and low-latency analytics in modern big data environments [46–48]. Chen et al. [24] emphasized velocity as a major characteristic of big data, while Elgendy and Elragal [10] showed that analytics can support different phases of decision making. Therefore, real-time processing is particularly important for decisions that lose value when delayed.

5.2.3. Lakehouse Architecture

Lakehouse architecture addresses the limitations of both data lakes and data warehouses. The work by Nargesian et al. [51] identified data lake management challenges, while Armbrust et al. [52] proposed the lakehouse as a unified architecture for analytics and AI. Building on this discussion, Harby and Zulkernine [13] experimentally compared data lake, warehouse, and lakehouse systems, and Pohl et al. [53] reviewed lakehouse architectures for time-series data.

Lakehouse architecture combines features of data lakes and data warehouses [13]. Data lakes are useful for storing large volumes of structured, semi-structured, and unstructured data, while warehouses are optimized for reliable querying and business intelligence. As noted by Harby and Zulkernine [13], the lakehouse approach seeks to reduce fragmentation by supporting storage flexibility, schema management, and analytical reliability. A review by Wixom et al. [9] showed that analytics value depends on usable and trusted information, which lakehouse architecture attempts to support.

5.2.4. MLOps and DataOps

MLOps and DataOps support reproducibility, monitoring, deployment, and maintenance of analytical models. As shown in the work by Kreuzberger et al. [12], MLOps integrates machine learning, software engineering, and data engineering practices. Sculley et al. [125] discussed technical debt in machine learning systems, while Amershi et al. [126] examined software engineering practices for machine learning. These studies show why modern BDA must manage models after deployment, not only during model development.

MLOps and DataOps extend BDA beyond data processing by focusing on the operational management of data pipelines and machine learning models [12]. DataOps emphasizes reliable data flow, automation, testing, and collaboration, while MLOps focuses on model deployment, monitoring, versioning, and lifecycle management. Papagiannidis et al. [15] discussed responsible AI governance, which complements MLOps by adding accountability, fairness, transparency, and human oversight. Rajan and Vetriselvi [14] also emphasized that secure and privacy-preserving mechanisms should be embedded into analytics workflows.

5.2.5. Integrated Governance

Integrated governance combines data quality, metadata, lineage, access control, model accountability, and responsible AI practices. The work by Tallon et al. [64] connected information governance with organizational value, while Abraham et al. [33] reviewed data governance frameworks. Raji et al. [104] proposed internal algorithmic auditing, and Papagiannidis et al. [15] presented a responsible AI governance framework for organizations.

Integrated governance is a key feature of modern BDA. It involves metadata management, data catalogs, lineage tracking, role-based access control, privacy protection, quality monitoring, model governance, and compliance. Mikalef et al. [35] argued that governance supports the transformation of analytics resources into organizational capability. Huynh et al. [7] identified the need for stronger theoretical integration in BDA capability research, while Papagiannidis et al. [15] and Xia et al. [34] show that responsible AI governance has become central to trustworthy and value-oriented analytics.

Overall, the movement from traditional to modern BDA represents a shift from infrastructure-centered processing toward capability-centered decision support [7,35]. Traditional systems focused mainly on storing and processing large datasets, whereas modern systems emphasize real-time insight, cloud elasticity, integrated governance, AI operations, and responsible decision making. Chatterjee et al. [11] showed that BDA can influence decision making, forecasting, and performance, which confirms the importance of moving

beyond infrastructure toward organizational value. Table 8 presents a direct comparison between traditional and modern approaches, Table 9 examines their strengths and limitations, and Table 10 traces the architectural evolution behind this transition.

Table 8. Comparison of traditional and modern BDA approaches.

Taxonomy Item	Traditional Approach	Modern Approach	Decision-Making Implication
Infrastructure	On-premises clusters managed internally	Cloud-native or hybrid infrastructure with elastic resources	Modern systems improve scalability and deployment speed; traditional systems offer direct control.
Processing model	Batch processing at scheduled intervals	Batch, stream, and hybrid real-time processing	Modern systems reduce decision latency and support faster operational response.
Core technology	Hadoop/MapReduce and distributed file systems	Spark, Kafka, Flink, cloud analytics services, and Lakehouse platforms	Modern tools support interactive, iterative, and continuous analytics more effectively.
Data architecture	Separate storage, processing, and reporting layers	Integrated Lakehouse architecture linking storage, analytics, ML, and BI	Integrated architecture improves data reuse and reduces fragmentation.
Governance	Limited or fragmented governance across systems	Integrated catalogs, lineage, access control, quality rules, and model governance	Stronger governance increases trust, compliance, and accountability.
Analytics capability	Mostly historical reporting and offline analysis	Predictive, prescriptive, AI-assisted, and real-time analytics	Modern analytics supports proactive and adaptive decision making.
Operational practices	Manual pipeline and model management	DataOps and MLOps for automation, monitoring, and versioning	Operational discipline improves reliability and reproducibility of analytical outputs.
Main limitation	High maintenance effort, slower scaling, and delayed insights	Cloud cost, vendor dependency, governance complexity, and skills requirements	Tool selection should match data sensitivity, budget, speed, and governance maturity.

Table 9. Strengths and limitations of traditional and modern analytics environments.

Environment or Approach	Advantages	Disadvantages	Suitable Use
On-premises Hadoop/MapReduce	Provides local control over infrastructure and can process very large datasets using distributed storage and computation [71].	Requires investment of hardware, cluster administration, tuning and specialized staff. It is less flexible when workloads change quickly.	Large historical datasets, batch jobs and organizations with strict internal infrastructure requirements [6,71].
Parallel DBMS and data warehouse	Offers mature query optimization, structured reporting and strong support for business intelligence.	Less suitable for highly unstructured, semi-structured or rapidly changing data sources. It may also be expensive on a very large scale.	Structured reporting, OLAP workloads and controlled enterprise data environments [13,117].

Table 9. *Cont.*

Environment or Approach	Advantages	Disadvantages	Suitable Use
Cloud analytics	Provides elastic computing, managed services and faster deployment without full ownership of physical infrastructure.	Creates concerns related to provider dependency, data transfer costs, compliance and cloud security.	Organizations requiring scalable analytics services and variable workloads [37,68].
Stream processing	Supports near-real-time monitoring, anomaly detection and rapid response to changing events.	Requires careful management of latency, fault tolerance, ordering, state and continuous data quality.	Fraud detection, IoT monitoring, cybersecurity, transportation and time-sensitive decisions [42,124].
Lakehouse architecture	Combines data lake flexibility with warehouse-style management, supporting BI, data science and machine learning on shared data.	Still requires strong governance, metadata management and architecture maturity to avoid turning into an unmanaged data lake.	Unified analytics platforms, mixed structured/unstructured data and AI-ready data environments [13,52].
MLOps and DataOps	Improves model deployment, monitoring, reproducibility and collaboration between data science and operations teams.	Adds process complexity and requires automation maturity, version control and continuous monitoring.	Production analytics, deployed ML systems and reliable model lifecycle management [12,32].
Integrated governance	Clarifies ownership, standards, access rights, quality controls and accountability across the analytics lifecycle.	May be difficult to implement uniformly across departments, platforms and external data sources.	Regulated sectors, multi-platform environments and high-impact analytics decisions [15,33].

Table 10. Evolution of BDA architecture.

Architecture Generation	Dominant Design Logic	Typical Limitation	Current Relevance
Enterprise data warehouse	Centralized, structured and schema-driven analytical repository.	Less flexible for unstructured, streaming and high-variety data.	Still useful for governed reporting and stable business intelligence [13,114].
MapReduce and Hadoop ecosystem	Distributed storage and batch processing on commodity clusters.	High latency and limited suitability for iterative machine learning and interactive analytics.	Important historical foundation for large-scale processing [6,43].
In-memory and iterative processing	Faster processing through memory-based computation and reusable working sets.	Requires careful resource management and cluster tuning.	Supports machine learning, iterative analytics and large-scale data engineering [45,47].
Cloud-native analytics	Elastic services, managed infrastructure and scalable storage-compute separation.	Vendor dependency, cost governance and data sovereignty concerns.	Dominant model for flexible analytics deployment and rapid scaling [67,68].

Table 10. *Cont.*

Architecture Generation	Dominant Design Logic	Typical Limitation	Current Relevance
Streaming and real-time analytics	Continuous processing of events, logs, transactions and sensor streams.	Requires fault tolerance, low-latency design and continuous monitoring.	Essential for fraud detection, IoT, cybersecurity and operational decision making [42,124].
Lakehouse and AI-enabled platforms	Integration of data lake flexibility with warehouse governance and AI/ML workloads.	Still evolving in relation to standards, governance and performance benchmarking.	Promising direction for unified analytics, machine learning and data governance [13,52].

5.3. Design Propositions

Conceptual synthesis shows that the organizational value of BDA is created by a series of mechanisms that are interdependent with each other, not by BDA adoption alone. Timeliness, completeness, and relevance of analytical outputs are shaped by architectural integration and task–analytics; trust, interpretability, accountability, and legitimate use of analytical outputs are shaped by governance and human oversight; translation of analytical insight into action is shaped by organizational readiness; and decision quality is the proximal link to broader organizational outcomes. The relationships are presented below in the form of design propositions. The propositions are theoretical expectations based on synthesis and not as empirical causal effects.

Proposition 1. *Architectural integration and quality of decisions. The increased integration between data sources, data ingestion, data storage, data processing, data analysis, data visualization, and data-driven operational decision processes will lead to faster, more consistent, and more usable analytical evidence, which will in turn enable higher-quality decisions than if these processes were implemented separately [10,12,13,26,65].*

Proposition 2. *Task–analytics fit. The link between BDA capability and decision quality is likely to be stronger if the processing latency, model type, explanation requirements and degree of automation are aligned with the characteristics of the decision task. Batch analytics can be used for decisions that are made periodically and historically, while stream analytics can be used when the value of information diminishes quickly with delay [9–11,29,42].*

Proposition 3. *Governance Assurance and Adoption. Embedded governance (data quality, lineage, accountability, explainability, fairness, compliance, security, privacy) is expected to boost trust in analytical outputs and their uptake in the organization. Trust and perceived legitimacy are thus theorized as means by which governance can impact on realized value of BDA [14,15,33,34,65,69].*

Proposition 4. *Complementarity of human and organization. Analytical models are likely to play a greater role in improving the quality of decision making and organizational performance if the data are accessible, if the manager has experience in the domain, if the manager has good judgment, if there is cross-disciplinary collaboration, if there is leadership support, and if the organization has the capacity to turn the analytical output into action [7,18,23,30,35]. High-impact, uncertain, irreversible, and poorly explained decisions should be given more weight in human oversight.*

Proposition 5. *Continuous adaptation. Feedback mechanisms between decision outcomes and data pipelines, platform configuration, model performance, visualizations and governance controls are expected to be a key factor in the long-term effectiveness of BDA systems. Continued alignment*

with changing organizational conditions should be facilitated through monitoring drift, degradation, cost, risk and unintended consequences [12,15,32,125,126].

Proposition 6. *Contingency of organizations and institutions. The technical BDA capability and the value of the decisions made is likely to be contingent on organizational readiness, task criticality, sectoral risk, professional norms, and regulatory conditions. Organizations that are more knowledgeable, have more leadership, human, financial and engagement readiness should be better able to translate BDA capability into sustained use and measurable outcomes [7,35,36,127,128].*

Proposition 7. *Different security and privacy routes. Security and privacy are expected to impact BDA adoption and organizational outcomes in various ways. Security is more related to the integrity, availability and resilience of the system, while privacy is more related to the perception of legitimate use of data, trust in stakeholders and regulatory acceptability. Their impacts should therefore not be aggregated into a single risk construct [14,129].*

6. Proposed Integrated Big Data Analytics Decision-Making Framework

6.1. Conceptual Foundation

The proposed framework is the result of conceptual synthesis and design propositions that were presented in the previous sections. It is not a novel individual technology or another linear data pipeline, but rather a combination of five relationships that are typically disjointed in previous work: architecture-to-analytics integration, task–analytics fit, governance-to-trust and adoption, human and organizational complementarity, and outcome-to-system feedback. Decision value is thus considered a socio-technical phenomenon that is an indirect result of model accuracy and platform sophistication, but instead a product of coordinated data, infrastructure, analytical methods, governance mechanisms, organizational readiness, and human judgment [7,18,30,35].

Accordingly, the framework adopts a socio-technical and lifecycle perspective. It represents the analytical process as a sequence of interdependent stages, while governance and responsibility operate across the entire sequence. This design reflects recent developments in stream processing, cloud-native and Lakehouse platforms, MLOps, data catalogues, and responsible AI governance [12,13,15,34,42,68,69].

6.2. Architectural Structure of the Framework

The proposed framework comprises seven stages, each connected to the next, together with two cross-cutting mechanisms: cross-cutting governance and a continuous feedback loop, as shown in Figure 1. While the sequential structure is an expression of the movement from raw evidence to organizational action, the transversal mechanisms will provide trust, accountability, and adaptation during the analytical lifecycle. Data sources. The framework starts with the internal and external data that come from enterprises, transactions, sensors, IoT devices, digital platforms, web logs, social media and open-data repositories that are heterogeneous. The importance of this stage for the analysis is not defined by volume but rather by the fitness, representativeness, origin, and timeliness of the evidence for the decision problem. To minimize the risk that an increasing amount of data only adds more noise or historical bias to the analysis, modern big data applications demand explicit selection of data sources and data provenance management [4,49].

Ingestion and integration. The second stage is based on batch and streaming data gathered via ETL/ELT pipelines, APIs, message brokers and event-processing. Differences in syntax, semantic inconsistencies, and update frequency, identity resolution and metadata capture are all issues to be addressed in integration. Data catalogues and semantic

alignment are particularly relevant since they enable data to be found, understood and reused by different teams and platforms [65,66,69].

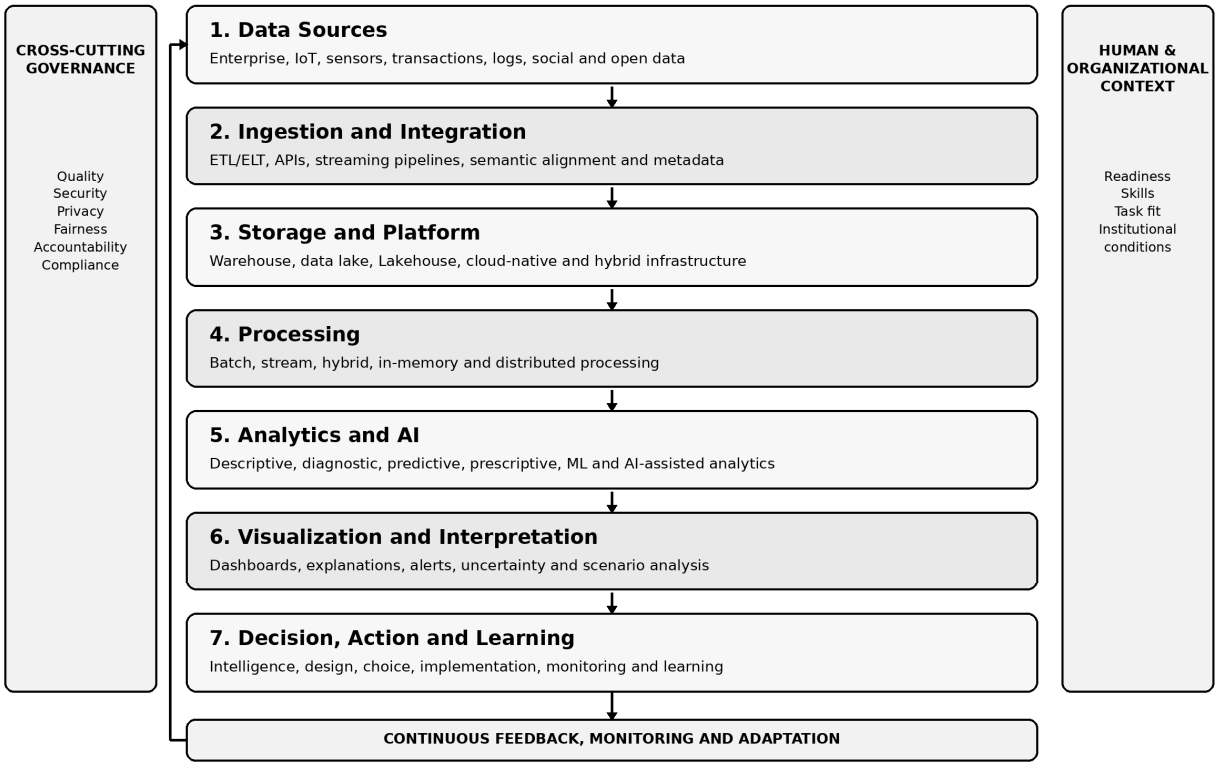


Figure 1. Integrated socio-technical BDA decision making framework with governance, organizational context, and continuous feedback.

Storage and platform. The third stage offers a scalable and controlled data foundation by implementing data warehouses, data lakes, Lakehouse platforms, cloud-native services or a hybrid setup. It is not a specific architecture that is prescribed in the framework, but the platform is to be selected based on the sensitivity of the data, latency, variability of workloads, cost, interoperability and maturity. Recent Lakehouse research shows that flexible storage and warehouse-style management can help with BI and ML on shared data, although the benefits depend on metadata quality, governance controls, and architectural discipline [13,53,68].

Processing. The fourth stage is the transformation of data via batch, stream, hybrid, in-memory and distributed processing. When the value of the insight fades quickly with time, stream processing is needed, but if the insight is only important at a specific point in time, batch processing remains appropriate for periodic reporting and historical model development. Modern systems thus demand to explicitly relate the processing latency with the decision's time urgency, as well as fault tolerance, state management, resource efficiency, and cost [42,45,46].

The use of analytics and artificial intelligence. The fifth stage takes processed data and transforms it into descriptive, diagnostic, predictive and prescriptive outputs. It can involve statistical modeling, machine learning, anomaly detection, optimization, simulation, and AI-powered analysis. The choice of model depends on the nature of the decision to be made and not because of novelty of the model itself. For production, MLOps practices need to handle model versions, reproducibility, deployment, monitoring, and retiring models, while recent architectural efforts highlight that all aspects of operational structures, processes, tools and responsibilities must be considered along the model's lifecycle [12,32].

Visualization and interpretation. The sixth stage converts the analytical results and outputs into formats that can be understood and questioned by experts and decision makers in the field. Managerial sensemaking is facilitated through dashboards, alerts, explanations, scenario analysis, uncertainty communication and interactive visualizations. This stage is intentionally independent of the generation of the model as it is only when the model is used in the appropriate way that it can be of predictive value. When decisions are high-impact, uncertain and controversial, explainability and contextual interpretation is critical [15,98,99].

Decision, action and learning. The final stage ties analytics to the action of the intelligence, design, choice, implementation and control stages of decision making. The analytical recommendations are merged with the organizational goals, domain knowledge, constraints and human judgment before action is taken. The results of the decision are then tracked to see if the intervention has the desired results. At this stage, analytics becomes an organizational learning platform, and quality of decision, adoption and value realized are the ultimate measures of success [8,11,20].

6.3. AI Enablement Across the Framework

In the proposed architecture, AI is not considered as an independent technology block. It focuses its processing power mainly during the analytics phase, but requires upstream data quality and integration, platform capacity, processing latency, and downstream interpretation, governance and human authorization. AI can be used to assist in anomaly detection, classification, prediction, suggestion, optimization, scenario creation, natural language communication with analytical systems, and explanation of model results, depending on the decision problem. Therefore, the right AI technique should be chosen based on the characteristics of the task, the required response time, uncertainty, criticality of the decision, and the availability of validated data, instead of novelty of the technology [12,15,31,32].

AI-powered parts should be controlled by explicit lifecycle controls for operational use. Training and evaluation data should be traceable, model versions and features, prompts and configuration changes should be documented, performance and drift should be monitored after deployment, and high-impact outputs should remain under human control. Generative AI can also be used to query data, summarize evidence of analysis, create alternative scenarios or explain the results of models, but this should not be used as authoritative if there is doubt regarding the source provenance, factual accuracy, privacy or consequences of decisions. This puts AI in the role of decision augmentation in a governed BDA lifecycle, and not as an independent substitute for managerial or professional decision making [15,98,99,104].

6.4. Cross-Cutting Governance, Human Oversight, and Organizational Readiness

Governance is not a final control step but rather a cross-cutting mechanism. From the acquisition of the data to post-deployment monitoring, data quality, access control, lineage, fairness, explainability, accountability, regulatory compliance, sustainability, security, and privacy must be built in [15,34]. This governance mechanism explicitly distinguishes security from privacy concerns. Security protects the integrity, availability and resilience of data and analytical services, while privacy regulates the lawful collection, linkage, inference, retention and sharing of sensitive data. This separation allows for empirical studies to examine if the two dimensions have different impacts on trust, adoption and organizational outcomes [14,129].

Human oversight offers domain interpretation, ethical judgment, contextual knowledge, exception handling and clear point of accountability. The framework considers

automation as a tool to augment decisions, not to replace them. The level of human engagement should be commensurate with the criticality of the decision, the reversibility of the decision, the uncertainty of the decision, the number of stakeholders who will be affected by the decision and the availability of meaningful explanations. Studies on professional use of technology also suggest that knowledge and professional judgment can influence intention to use advanced technologies, which means that human factors should be incorporated as mechanisms, and not as an afterthought to implementation [127].

Whether the architecture can be translated into routine use is dependent on the organizational readiness. The organization's ability to absorb analytical systems and act based on the results is impacted by leadership commitment, human capability, financial resources, engagement, data literacy, and cross-functional coordination. Recent evidence on AI and BDA adoption reveals that the readiness dimensions are interdependent and should be developed as a whole and not as stand-alone resources [128]. Thus, in the framework, the extent of the relationship between the technical capability and the decision value realized depends on readiness.

6.5. Feedback, Monitoring, and Continuous Adaptation

A feedback loop connects decision outcomes to all preceding stages. Operational monitoring should identify data drift, concept drift, model degradation, changing user behavior, emerging risks, cost escalation, and unintended organizational consequences. When deviations are detected, the organization may revise source selection, integration rules, platform configuration, model assumptions, visualization design, or governance controls. This continuous adaptation is particularly important in cloud-native and AI-enabled environments, where data, software dependencies, business conditions, and regulatory expectations change over time. MLOps and DataOps provide the operational mechanisms for this feedback process, while governance defines escalation paths, review responsibilities, and acceptable risk thresholds [12,32,69].

6.6. Framework Evaluation and Validation Strategy

The proposed framework has not yet been empirically validated because this study is conceptual. The revised paper introduces a multi-level validation protocol to connect each conceptual component to observable indicators and suitable empirical approaches, and thus to make the framework evaluable. Validation should be done from technical feasibility to decision-use evaluation to organizational outcomes. This will enable future studies to evaluate both the ability of the architecture to function as designed, and the effectiveness of the architecture in enhancing the decision process in particular situations.

Table 11 is an operationalization of this evaluation logic. The technical layers can be evaluated based on latency, throughput, reliability, interoperability, resource utilization and cost. Assessment of analytical components can be done predictively or prescriptively, in terms of robustness, drift, interpretability and fairness. Usability, trust, data literacy, role clarity, adoption and readiness are examples of human and organizational dimensions. Auditability, effectiveness of access-control, compliance with privacy, security incidents, accountability and review procedures are all examples of ways to evaluate governance. Relevant outcomes at the decision level are decision speed, decision quality, adoption, reversibility, error rates and realized organizational value. Examples of appropriate methods are proof of concept benchmarking, expert review, controlled experiments, surveys, comparative case studies and longitudinal evaluations [12,15,23,32,34].

A rigorous empirical evaluation should therefore be context-specific and should not assume that all indicators improve simultaneously. For example, lower latency may increase infrastructure costs, stronger privacy protection may reduce data granularity, and

increased automation may require stronger oversight. These trade-offs are part of the framework's evaluation logic and are explicitly treated as empirical questions rather than predetermined outcomes.

Table 11. Evaluation and validation protocol for the proposed framework.

Evaluation Domain	Illustrative Indicators	Validation Method	Expected Evidence
Technical architecture	Latency, throughput, reliability, interoperability, resource use, cost	Benchmarking; proof-of-concept deployment	Evidence that the selected architecture meets workload and timing requirements
Analytics and AI	Accuracy or utility, robustness, drift, interpretability, fairness	Offline validation; stress testing; model monitoring	Evidence that analytical outputs remain useful, stable, and understandable
Human and organizational	Usability, trust, data literacy, role clarity, readiness, adoption	Expert review; user studies; surveys	Evidence that users can interpret, challenge, and act on analytical outputs
Governance, security and privacy	Auditability, access control, incidents, compliance, privacy risk, accountability	Control assessment; audit; threat/privacy review	Evidence that data and models are used within acceptable risk and accountability boundaries
Decision process	Decision speed, quality, error rate, adoption, reversibility	Scenario tests; controlled studies; case comparison	Evidence that architecture supports better decision processes in the target context
Organizational outcomes	Cost, productivity, risk reduction, service quality, resilience, realized value	Longitudinal case study; before/after comparison	Evidence of sustained value beyond technical performance

6.7. Comparison with Existing Big Data Analytics Frameworks

There are several streams of existing BDA frameworks, some of which overlap, such as technical architecture, organizational capability models and governance-oriented approaches. Initial technology-centric approaches focused on data collection, distributed storage, processing engines and analytical tools, and provided less attention to organizational decision processes, post-deployment model management and embedded governance [24,26,37]. Capability-based frameworks then extended this perspective by adding in the technological, human, organizational, and intangible resources needed to create business value [18,30,35]. These approaches tend to be more descriptive of what an organization needs to do, however, and do not provide as clear of an explanation of how data pipelines, real-time processing, model operations, governance controls and managerial decisions interact across the entire decision lifecycle.

A more recent framework tackles important but more limited issues of this problem. To achieve this, a new architecture was developed, known as the lakehouse architecture, which brings together the flexibility of a data-lake with the reliability and support of a data-warehouse and supports shared business-intelligence and machine-learning workloads [13,52,68]. The goal of the MLOps architecture is to deploy, version, monitor, reproduce and maintain models [12,32]. Modern governance models focus on ownership, metadata, access control, accountability, responsible AI and data catalogues [15,65,69]. While these contributions are valuable, they are usually offered as platform, model-lifecycle, or governance solutions, but not as an integrated architecture that connects analytical operations with the formulation of decisions, their implementation, and learning. Table 12

outlines these differences by summarizing the key framework streams, their focus and limitations, and the unique contribution of the proposed framework.

Table 12. Comparison of the proposed framework with existing BDA framework streams.

Framework Stream	Primary Focus	Typical Limitation	Distinctive Contribution of the Proposed Framework
Technology-oriented BDA frameworks	Data acquisition, distributed storage, processing engines, and analytical tools.	Limited integration with managerial decision stages, embedded governance, and post-deployment learning.	Links the technical analytics lifecycle directly to decision formulation, implementation, monitoring, and feedback [24,26,37].
BDA capability frameworks	Technological, human, organizational, and intangible resources that enable analytics value.	Explain required capabilities more clearly than the operational interactions among pipelines, models, governance, and decisions.	Translates capability requirements into an end-to-end socio-technical reference architecture [7,18,30,35].
Lakehouse and cloud-native architectures	Unified storage, reliable data management, scalable computation, and shared BI/ML workloads.	Primarily platform-centered; decision processes, human interpretation, and organizational learning remain outside the core architecture.	Use Lakehouse or cloud services as one architectural layer within a wider governed decision system [13,52,66,68].
Stream-processing frameworks	Continuous ingestion and low-latency processing for event-driven analytics.	Optimize processing speed but do not fully specify how real-time outputs are interpreted, governed, acted upon, and evaluated.	Combines batch, streaming, and hybrid processing with interpretation, human oversight, and outcome monitoring [42,45,46].
MLOps and model-lifecycle frameworks	Model deployment, versioning, reproducibility, monitoring, maintenance, and technical reliability.	Model-centered scope may underrepresent data governance, managerial choice, accountability, and feedback from implemented decisions.	Connects DataOps/MLOps with explainability, governance, decision accountability, and closed-loop adaptation [12,15,32].
Data and responsible-AI governance frameworks	Ownership, metadata, quality, privacy, security, fairness, accountability, and compliance.	Governance is often treated as a specialized control domain rather than a mechanism embedded across the entire decision architecture.	Positions governance and human oversight as cross-cutting controls operating across all technical and managerial layers [15,34,65,69].
Technology-adoption and organizational-readiness models	Knowledge, perceived benefits and barriers, leadership, human capability, financial resources, engagement, and contextual conditions shaping adoption.	Often explain willingness or readiness to adopt a technology without specifying how adoption conditions interact with end-to-end BDA architecture and post-decision learning.	Uses readiness, task characteristics, and institutional conditions as configurational factors that moderate how the seven-stage architecture is adopted and converted into decision value [127,128].

The proposed framework has the following six differences from previous approaches, as summarized in Table 12. First, it integrates the technical analytics lifecycle into the managerial decision-making lifecycle, instead of viewing infrastructure and decision use as distinct areas. Second, it provides batch, streaming and hybrid processing, allowing a single conceptual architecture to meet the needs of different types of decisions with varying temporal demands. Third, governance becomes a transversal type of assurance, rather than just a compliance exercise at the end. Fourth, DataOps and MLOps principles are connected to explainability, with human oversight, outcome monitoring and organizational accountability. Fifth, the framework is closed-loop, where the results of decisions feed back into the data quality, platform configuration, model performance, governance controls, and future decision rules. Sixth, organizational readiness, task characteristics, and institutional conditions are considered as boundary conditions that shape the architecture, but not as implementation details that are outside of the architecture [127,128].

Accordingly, the novelty of the proposed framework does not depend on introducing a new individual analytics technology. Its contribution lies in integrating previously fragmented technical, analytical, organizational, governance, and decision-making components into a unified socio-technical architecture. This integration distinguishes the framework from technology stacks, capability models, Lakehouse architectures, MLOps lifecycles, and governance frameworks, while providing a clearer basis for future empirical validation of trustworthy and measurable decision value [7,15,34,66].

6.8. Boundary Conditions, Scalability, and Configurability

The architecture is not designed to be applied in the same way by all organizations. It has boundary conditions such as organizational size, analytics maturity, criticality of the task, sensitivity of data, scale of workload, current infrastructure, budget, staff capability, industry risk and regulatory obligations. Small organizations can have multiple functions in managed services or in a small open-source stack, while large or highly regulated organizations can have the same functions split up into various platforms, teams, and control processes. The framework is thus conceptually scalable, maintaining the necessary functions and relationships without specifying several servers, tools or organizational units [7,35,36,128].

Scalability should also be understood in various ways. Technical scaling is needed for data volume and velocity, operational scaling is needed for user, model and business unit growth, and governance scaling is needed for expansion into regulated or high-impact decisions. Technical growth can be met with horizontal distributed processing, elastic cloud resources, workload partitioning, and tiered storage, but at a larger scale, the amount of metadata, lineage, access-control, observability, skills, and cost-management requirements also increases. Thus, for organizations of various sizes, adoption is only possible if the implementation is commensurate with the size of the decision problem and with the readiness of the organization, if the implementation is not too complex, and if it avoids unnecessary complexity.

The primary boundary of the framework is that it defines functional relationships, not necessarily the best configuration. For a low-risk use case, the level of AI, the governance structure, and the frequency of processing might be less robust, whereas a real-time healthcare, financial, or public-sector application might demand more stringent security, privacy evaluations, explainability, audit trails, redundancy, and human authorization. Thus, the framework should not be context-independent and future empirical research should test the propositions separately for different organization sizes, sectors, decision types and regulatory environments.

6.9. Practical Feasibility, Implementation Challenges, and Open-Source Pathways

Practical implementation can be achieved as a step-by-step architecture instead of a single big transformation project. Organizations can start with a well-defined decision use case, add the data sources needed for that use case, set minimum quality requirements, set access controls, and choose the processing and analytical components based on the latency and complexity requirements. After stabilization, other sources, models, business units and governance controls may be added. This staged approach minimizes the risk of over-engineering and enables testing prior to investing in an enterprise platform.

While open-source components can reduce software-license expenditure and vendor dependency, they do not eliminate infrastructure, integration, security, maintenance, training, or governance costs. Table 13 shows an illustrative low-cost path for implementation. The options listed are examples only and organizations should choose the tools based on their current capabilities, interoperability requirements, support requirements, workload size, and regulatory requirements. Technologies like Kafka, Spark, Flink, HDFS, Iceberg, MLflow, Kubeflow, and Apache Atlas are already featured in the technical and MLOps literature mentioned in this paper [12,32,42,45–48,69,71].

Table 13. Illustrative open-source implementation options for the proposed framework.

Framework Level	Illustrative Open-Source Options	Cost-Conscious Implementation Approach	Practical Challenges
Data sources	PostgreSQL, MariaDB, Apache Cassandra, object/file stores	Reuse existing operational sources; expose only required data through controlled connectors	Data ownership, source quality, schema drift, and access permissions
Ingestion and integration	Apache Kafka, Apache NiFi, Airbyte	Start with scheduled ingestion; add streaming only where decision latency requires it	Connector maintenance, semantic mismatch, duplicate events, metadata capture
Storage and platform	HDFS, MinIO, Apache Iceberg, Delta Lake	Use commodity or cloud object storage; separate hot and archival data	Capacity planning, backup, data lifecycle, metadata and governance overhead
Processing	Apache Spark, Apache Flink	Use batch processing for non-urgent workloads; reserve streaming resources for time-sensitive cases	Cluster tuning, state management, fault tolerance, compute cost
Analytics and AI	Python/scikit-learn, H2O.ai, MLflow, Kubeflow	Begin with interpretable baseline models; automate deployment only after stable validation	Model drift, reproducibility, explainability, skills, GPU/compute demand for advanced AI
Visualization and interpretation	Apache Superset, Grafana	Provide role-specific dashboards and alerts rather than broad unrestricted reporting	Poor visual design, information overload, misinterpretation of uncertainty
Governance and control	Apache Atlas, OpenMetadata, Keycloak, OpenBao	Prioritize identity/access control, metadata, lineage, logging, and high-risk data first	Configuration complexity, policy ownership, audit effort, cross-platform consistency
Decision, action and learning	Workflow tools plus existing ticketing/ERP/BI processes	Integrate recommendations into existing approval and monitoring processes before adding automation	Change management, accountability, adoption, feedback quality

There are still some practical restrictions. The cost of licensing may be shifted to engineering and operations, heterogeneous tools may lead to an increased burden in integration and observability, skilled personnel may be harder to retain, cloud elasticity may result in unpredictable usage charges, advanced AI may require expensive computing, and governance requirements may hinder experimentation. Security and privacy also need to be implemented and not just a result of the architecture. Therefore, to achieve cost-effective implementation, the architecture should be simple, existing systems should be reused, workloads should be prioritized, monitoring should be automated, data-retention policies should be established, proportionate governance should be implemented, and the total cost of ownership should be explicitly assessed.

7. Conclusions and Future Directions

In this paper, an integrated socio-technical framework for BDA-supported decision making was developed. Instead of viewing BDA as a set of tools, the framework is a set of tools that are linked to each other in a configurable lifecycle, including processing, platform, analytics, AI, governance, organizational readiness, human judgment, and decision processes. The proposed architecture illustrates the coordination and cross-cutting governance of data sources, ingestion and integration, storage and platform services, processing, analytics and AI, visualization and interpretation, and decision, action, and learning, with continuous feedback. The design propositions explain the anticipated connections between the architectural integration, task–analytics fit, governance assurance, human and organizational complementarity, security and privacy, contextual conditions and continuous adaptation. The paper also outlines a validation protocol and an incremental open-source implementation pathway to make the architecture empirically testable and practically interpretable. The framework is still conceptual and not guaranteed to result in universal performance gains. Future work should validate its constructs and trade-offs in proof-of-concept implementations, expert evaluations, comparative case studies, surveys, longitudinal designs and design-science studies in a variety of organization sizes, sectors, regulatory contexts and analytics-maturity levels. Total cost of ownership, security and privacy controls, AI governance, human oversight, scalability and when more technology means more decision value should be considered.

Author Contributions: Conceptualization, W.Z.A. and M.Y.M.; methodology, W.Z.A. and F.N.A.-A.; investigation, W.Z.A., F.N.A.-A., A.Q., N.A.K., H.A. and M.Y.M.; writing—original draft preparation, W.Z.A.; writing—review and editing, W.Z.A., F.N.A.-A., A.Q., N.A.K., H.A. and M.Y.M.; visualization, W.Z.A.; supervision, M.Y.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding authors.

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

References

1. Chen, H.; Chiang, R.H.L.; Storey, V.C. Business intelligence and analytics: From big data to big impact. *MIS Q.* **2012**, *36*, 1165–1188. [CrossRef]
2. Labrinidis, A.; Jagadish, H.V. Challenges and opportunities with big data. *Proc. VLDB Endow.* **2012**, *5*, 2032–2033. [CrossRef]
3. Agarwal, R.; Dhar, V. Editorial-Big data, data science, and analytics: The opportunity and challenge for IS research. *Inf. Syst. Res.* **2014**, *25*, 443–448. [CrossRef]
4. Badshah, A.; Daud, A.; Alharbey, R.; Banjar, A.; Bukhari, A.; Alshemaimri, B. Big data applications: Overview, challenges and future. *Artif. Intell. Rev.* **2024**, *57*, 290. [CrossRef]

5. Jamarani, A.; Haddadi, S.; Sarvizadeh, R.; Haghi Kashani, M.; Akbari, M.; Moradi, S. Big data and predictive analytics: A systematic review of applications. *Artif. Intell. Rev.* **2024**, *57*, 176. [CrossRef]
6. Tosi, D.; Kokaj, R.; Roccetti, M. 15 years of Big Data: A systematic literature review. *J. Big Data* **2024**, *11*, 73. [CrossRef]
7. Huynh, M.T.; Nippa, M.; Aichner, T. Big data analytics capabilities: Patchwork or progress. A systematic review of the status quo and implications for future research. *Technol. Forecast. Soc. Change* **2023**, *197*, 122884. [CrossRef]
8. Rodepeter, E.; Gschnaidtner, C.; Hottenrott, H. Big data-based management decisions and start-up performance. *Small Bus. Econ.* **2026**, *67*, 361–399. [CrossRef]
9. Wixom, B.; Yen, B.; Relich, M. Maximizing value from business analytics. *MIS Q. Exec.* **2013**, *12*, 111–123.
10. Elgendy, N.; Elragal, A. Big data analytics in support of the decision-making process. *Procedia Comput. Sci.* **2016**, *100*, 1071–1084. [CrossRef]
11. Chatterjee, S.; Chaudhuri, R.; Gupta, S.; Sivarajah, U.; Bag, S. Assessing the impact of big data analytics on decision-making processes, forecasting, and performance of a firm. *Technol. Forecast. Soc. Change* **2023**, *196*, 122824. [CrossRef]
12. Kreuzberger, D.; Kuhl, N.; Hirschl, S. Machine Learning Operations (MLOps): Overview, definition, and architecture. *IEEE Access* **2023**, *11*, 31866–31879. [CrossRef]
13. Harby, A.A.; Zulkernine, F. Data lakehouse: A survey and experimental study. *Inf. Syst.* **2025**, *127*, 102460. [CrossRef]
14. Rajan, A.A.; Vetriselvi, V. Systematic survey: Secure and privacy-preserving big data analytics in cloud. *J. Comput. Inf. Syst.* **2024**, *64*, 136–156. [CrossRef]
15. Papagiannidis, E.; Mikalef, P.; Conboy, K. Responsible artificial intelligence governance: A review and research framework. *J. Strateg. Inf. Syst.* **2025**, *34*, 101885. [CrossRef]
16. Shmueli, G.; Koppius, O.R. Predictive analytics in information systems research. *MIS Q.* **2011**, *35*, 553–572. [CrossRef]
17. Chen, C.L.P.; Zhang, C.Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Inf. Sci.* **2014**, *275*, 314–347. [CrossRef]
18. Gupta, M.; George, J.F. Toward the development of a big data analytics capability. *Inf. Manag.* **2016**, *53*, 1049–1064. [CrossRef]
19. Ghasemaghaei, M.; Hassanein, K.; Turel, O. Increasing firm agility through the use of data analytics: The role of fit. *Decis. Support Syst.* **2018**, *101*, 95–105. [CrossRef]
20. Phan, T.; Baird, K. The use of big data analytics in performance management: The antecedents and role in enhancing performance measurement system effectiveness. *J. Manag. Control* **2026**, *37*, 111–139. [CrossRef]
21. Janssen, M.; van der Voort, H.; Wahyudi, A. Factors influencing big data decision-making quality. *J. Bus. Res.* **2017**, *70*, 338–345. [CrossRef]
22. Muller, O.; Fay, M.; vom Brocke, J. The effect of big data and analytics on firm performance: An econometric analysis considering industry characteristics. *J. Manag. Inf. Syst.* **2018**, *35*, 488–509. [CrossRef]
23. Ghasemaghaei, M. Does data analytics use improve firm decision making quality. The role of knowledge sharing and data analytics competency. *Decis. Support Syst.* **2019**, *120*, 14–24. [CrossRef]
24. Chen, M.; Mao, S.; Liu, Y. Big data: A survey. *Mob. Netw. Appl.* **2014**, *19*, 171–209. [CrossRef]
25. Gandomi, A.; Haider, M. Beyond the hype: Big data concepts, methods, and analytics. *Int. J. Inf. Manag.* **2015**, *35*, 137–144. [CrossRef]
26. Oussous, A.; Benjelloun, F.Z.; Lahcen, A.A.; Belfkih, S. Big data technologies: A survey. *J. King Saud Univ. Comput. Inf. Sci.* **2018**, *30*, 431–448. [CrossRef]
27. Elgendy, N.; Elragal, A. Big data analytics: A literature review paper. In *Advances in Data Mining. Applications and Theoretical Aspects*; Springer: Cham, Switzerland, 2014; pp. 214–227.
28. Provost, F.; Fawcett, T. Data science and its relationship to big data and data-driven decision making. *Big Data* **2013**, *1*, 51–59. [CrossRef] [PubMed]
29. Wang, Y.; Kung, L.; Byrd, T.A. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technol. Forecast. Soc. Change* **2018**, *126*, 3–13. [CrossRef]
30. Akter, S.; Wamba, S.F.; Gunasekaran, A.; Dubey, R.; Childe, S.J. How to improve firm performance using big data analytics capability and business strategy alignment. *Int. J. Prod. Econ.* **2016**, *182*, 113–131. [CrossRef]
31. Cao, J. Intelligent decision-making in business management: Integrating artificial intelligence and big data analytics for strategic optimization in enterprise operations. *Sustain. Comput. Inform. Syst.* **2026**, *51*, 101382. [CrossRef]
32. Amou Najafabadi, F.A.; Bogner, J.; Gerostathopoulos, I.; Lago, P. An architectural perspective on MLOps: Structures, processes, tools, and stakeholders. *Inf. Softw. Technol.* **2026**, *193*, 108029. [CrossRef]
33. Abraham, R.; Schneider, J.; vom Brocke, J. Data governance: A conceptual framework, structured review, and research agenda. *Int. J. Inf. Manag.* **2019**, *49*, 424–438. [CrossRef]
34. Xia, H.; Chen, H.; Zhang, J.Z.; Kamal, M.M. Exploring the impact of responsible AI governance on corporate performance: A quasi-natural experiment. *Technol. Forecast. Soc. Change* **2026**, *223*, 124425. [CrossRef]

35. Mikalef, P.; Pappas, I.O.; Krogstie, J.; Giannakos, M. Big data analytics capabilities: A systematic literature review and research agenda. *Inf. Syst. e-Bus. Manag.* **2018**, *16*, 547–578. [CrossRef]
36. Mikalef, P.; Krogstie, J. Examining the interplay between big data analytics and contextual factors in driving process innovation capabilities. *Eur. J. Inf. Syst.* **2020**, *29*, 260–287. [CrossRef]
37. Hashem, I.A.T.; Yaqoob, I.; Anuar, N.B.; Mokhtar, S.; Gani, A.; Khan, S.U. The rise of big data on cloud computing: Review and open research issues. *Inf. Syst.* **2015**, *47*, 98–115. [CrossRef]
38. Assuncao, M.D.; Calheiros, R.N.; Bianchi, S.; Netto, M.A.S.; Buyya, R. Big Data computing and clouds: Trends and future directions. *J. Parallel Distrib. Comput.* **2015**, *79–80*, 3–15. [CrossRef]
39. Kune, R.; Konugurthi, P.K.; Agarwal, A.; Chillarige, R.R.; Buyya, R. The anatomy of big data computing. *Softw. Pract. Exp.* **2016**, *46*, 79–105. [CrossRef]
40. Yaqoob, I.; Hashem, I.A.T.; Gani, A.; Mokhtar, S.; Ahmed, E.; Anuar, N.B.; Vasilakos, A.V. Big data: From beginning to future. *Int. J. Inf. Manag.* **2016**, *36*, 1231–1247. [CrossRef]
41. Marjani, M.; Nasaruddin, F.; Gani, A.; Karim, A.; Hashem, I.A.T.; Siddiqa, A.; Yaqoob, I. Big IoT data analytics: Architecture, opportunities, and open research challenges. *IEEE Access* **2017**, *5*, 5247–5261. [CrossRef]
42. Marcu, O.-C.; Bouvry, P. *Big Data Stream Processing*; Technical Report; University of Luxembourg: Luxembourg, 2024.
43. Dean, J.; Ghemawat, S. MapReduce: Simplified data processing on large clusters. *Commun. ACM* **2008**, *51*, 107–113.
44. Sakr, S.; Liu, A.; Batista, D.M.; Alomari, M. A survey of large scale data management approaches in cloud environments. *IEEE Commun. Surv. Tutor.* **2011**, *13*, 311–336. [CrossRef]
45. Henning, S.; Hasselbring, W. Benchmarking scalability of stream processing frameworks deployed as microservices in the cloud. *J. Syst. Softw.* **2024**, *208*, 111879. [CrossRef]
46. Almeida, A.; Brás, S.; Sargento, S.; Pinto, F.C. Time series big data: A survey on data stream frameworks, analysis and algorithms. *J. Big Data* **2023**, *10*, 83. [CrossRef] [PubMed]
47. Zaharia, M.; Xin, R.S.; Wendell, P.; Das, T.; Armbrust, M.; Dave, A.; Meng, X.; Rosen, J.; Venkataraman, S.; Franklin, M.J.; et al. Apache Spark: A unified engine for big data processing. *Commun. ACM* **2016**, *59*, 56–65.
48. Carbone, P.; Katsifodimos, A.; Ewen, S.; Markl, V.; Haridi, S.; Tzoumas, K. Apache Flink: Stream and batch processing in a single engine. *IEEE Data Eng. Bull.* **2015**, *38*, 28–38.
49. Ogrizović, M.; Drašković, D.; Bojić, D. Quality assurance strategies for machine learning applications in big data analytics: An overview. *J. Big Data* **2024**, *11*, 156. [CrossRef]
50. Grolinger, K.; Higashino, W.A.; Tiwari, A.; Capretz, M.A.M. Data management in cloud environments: NoSQL and NewSQL data stores. *J. Cloud Comput.* **2013**, *2*, 22. [CrossRef]
51. Nargesian, F.; Zhu, E.; Miller, R.J.; Pu, K.Q.; Arocena, P.C. Data Lake management: Challenges and opportunities. *Proc. VLDB Endow.* **2019**, *12*, 1986–1989.
52. Armbrust, M.; Ghodsi, A.; Xin, R.S.; Zaharia, M. Lakehouse: A new generation of open platforms that unify data warehousing and advanced analytics. *Proc. CIDR* **2021**, *8*, 28.
53. Pohl, M.; Wijemanne, N.D.; Staegemann, D.; Haertel, C.; Daase, C.; Dreschel, D.; Walia, D.S.; Osterthun, A.; Reibert, J.; Turowski, K. Data Lakehouse for Time Series Data: A Systematic Literature Review. In Proceedings of the 2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 15–18 December 2024; pp. 5833–5842. [CrossRef]
54. Ji, S.; Li, Q.; Cao, W.; Zhang, P.; Muccini, H. Quality assurance technologies of big data applications: A systematic literature review. *Appl. Sci.* **2020**, *10*, 8052. [CrossRef]
55. Buyya, R.; Ilager, S.; Arroba, P. Energy-efficiency and sustainability in new generation cloud computing: A vision and directions for integrated management of data centre resources and workloads. *Softw. Pract. Exp.* **2024**, *54*, 24–38. [CrossRef]
56. Wu, X.; Zhu, X.; Wu, G.Q.; Ding, W. Data mining with big data. *IEEE Trans. Knowl. Data Eng.* **2014**, *26*, 97–107. [CrossRef]
57. Qiu, J.; Wu, Q.; Ding, G.; Xu, Y.; Feng, S. A survey of machine learning for big data processing. *EURASIP J. Adv. Signal Process.* **2016**, *2016*, 67. [CrossRef]
58. Najafabadi, M.M.; Villanustre, F.; Khoshgoftaar, T.M.; Seliya, N.; Wald, R.; Muharemagic, E. Deep learning applications and challenges in big data analytics. *J. Big Data* **2015**, *2*, 1. [CrossRef]
59. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. [CrossRef] [PubMed]
60. Witten, I.H.; Frank, E.; Hall, M.A.; Pal, C.J. *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed.; Morgan Kaufmann: Burlington, MA, USA, 2016.
61. Castro, A.; Villagrá, V.A.; García, P.; Rivera, D.; Toledo, D. An ontological-based model to data governance for big data. *IEEE Access* **2021**, *9*, 109943–109959. [CrossRef]
62. Khatri, V.; Brown, C.V. Designing data governance. *Commun. ACM* **2010**, *53*, 148–152. [CrossRef]
63. Otto, B. A morphology of the organisation of data governance. In Proceedings of the 19th European Conference on Information Systems (ECIS), Helsinki, Finland, 9–11 June 2011; p. 272.

64. Tallon, P.P.; Ramirez, R.V.; Short, J.E. The information artifact in IT governance: Toward a theory of information governance. *J. Manag. Inf. Syst.* **2013**, *30*, 141–178. [CrossRef]
65. Bližnák, K.; Munk, M.; Pilková, A. A systematic review of recent literature on data governance (2017–2023). *IEEE Access* **2024**, *12*, 149875–149888. [CrossRef]
66. Gieß, A.; Hutterer, A. The future of data management: A delimitation of data platforms, data spaces, data meshes, and data fabrics. *Inf. Syst. e-Bus. Manag.* **2025**, *23*, 971–997. [CrossRef]
67. Armbrust, M.; Fox, A.; Griffith, R.; Joseph, A.D.; Katz, R.; Konwinski, A.; Lee, G.; Patterson, D.; Rabkin, A.; Stoica, I.; et al. A view of cloud computing. *Commun. ACM* **2010**, *53*, 50–58. [CrossRef]
68. AbouZaid, A.; Barclay, P.J.; Chrysoulas, C.; Pitropakis, N. Building a modern data platform based on the data lakehouse architecture and cloud-native ecosystem. *Discov. Appl. Sci.* **2025**, *7*, 166. [CrossRef]
69. Tonnarelli, M.; Kumara, I.; Driessen, S.; Tamburri, D.A.; van den Heuvel, W.J.; Oor, P. Data catalog tools: A systematic multivocal literature review. *J. Syst. Softw.* **2025**, *230*, 112584. [CrossRef]
70. Kreps, J.; Narkhede, N.; Rao, J. Kafka: A distributed messaging system for log processing. *Proc. NetDB* **2011**, *11*, 1–7.
71. Shvachko, K.; Kuang, H.; Radia, S.; Chansler, R. The Hadoop Distributed File System. In Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), Incline Village, NV, USA, 3–7 May 2010; pp. 1–10.
72. Wixom, B.H.; Watson, H.J. The BI-based organization. *Int. J. Bus. Intell. Res.* **2010**, *1*, 13–28. [CrossRef]
73. Sivarajah, U.; Kamal, M.M.; Irani, Z.; Weerakkody, V. Critical analysis of big data challenges and analytical methods. *J. Bus. Res.* **2017**, *70*, 263–286. [CrossRef]
74. Demchenko, Y.; Ngo, C.; de Laat, C.; Membrey, P.; Gordijenko, D. Big security for big data: Addressing security challenges for the big data infrastructure. In *Proceedings of the Workshop on Secure Data Management*; Springer: Cham, Switzerland, 2014; pp. 76–85.
75. Katal, A.; Wazid, M.; Goudar, R.H. Big data: Issues, challenges, tools and good practices. In Proceedings of the 2013 Sixth International Conference on Contemporary Computing (IC3), Noida, India, 8–10 August 2013; pp. 404–409.
76. Khan, N.; Yaqoob, I.; Hashem, I.A.T.; Inayat, Z.; Ali, W.; Alam, M.; Shiraz, M.; Gani, A. Big data: Survey, technologies, opportunities, and challenges. *Sci. World J.* **2014**, *2014*, 712826. [CrossRef] [PubMed]
77. Batini, C.; Cappiello, C.; Francalanci, C.; Maurino, A. Methodologies for data quality assessment and improvement. *ACM Comput. Surv.* **2009**, *41*, 1–52. [CrossRef]
78. Cai, L.; Zhu, Y. The challenges of data quality and data quality assessment in the big data era. *Data Sci. J.* **2015**, *14*, 2. [CrossRef]
79. Merino, J.; Caballero, I.; Rivas, B.; Serrano, M.; Piattini, M. A data quality in use model for big data. *Future Gener. Comput. Syst.* **2016**, *63*, 123–130. [CrossRef]
80. Taleb, I.; Serhani, M.A.; Dssouli, R. Big data quality: A survey. In Proceedings of the IEEE International Congress on Big Data, San Francisco, CA, USA, 2–7 July 2018; pp. 166–173.
81. Begoli, E.; Goethert, I.; Knight, K. A lakehouse architecture for the management and analysis of heterogeneous data for analytics and AI. In Proceedings of the 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 15–18 December 2021; pp. 5317–5326.
82. Dinh, H.T.; Lee, C.; Niyato, D.; Wang, P. A survey of mobile cloud computing: Architecture, applications, and approaches. *Wirel. Commun. Mob. Comput.* **2013**, *13*, 1587–1611. [CrossRef]
83. Ranjan, R. Streaming big data processing in datacenter clouds. *IEEE Cloud Comput.* **2014**, *1*, 78–83. [CrossRef]
84. Buyya, R.; Srirama, S.N.; Casale, G.; Calheiros, R.; Simmhan, Y.; Varghese, B.; Gelenbe, E.; Javadi, B.; Vaquero, L.M.; Netto, M.A.S.; et al. A manifesto for future generation cloud computing: Research directions for the next decade. *ACM Comput. Surv.* **2019**, *51*, 1–38.
85. Toosi, A.N.; Calheiros, R.N.; Buyya, R. Interconnected cloud computing environments: Challenges, taxonomy, and survey. *ACM Comput. Surv.* **2014**, *47*, 7. [CrossRef]
86. Hai, R.; Geisler, S.; Quix, C. Constance: An intelligent data lake system. In Proceedings of the 2016 International Conference on Management of Data (SIGMOD '16), San Francisco, CA, USA, 26 June–1 July 2016; pp. 2097–2100. [CrossRef]
87. Zuech, R.; Khoshgoftaar, T.M.; Wald, R. Intrusion detection and big heterogeneous data: A survey. *J. Big Data* **2015**, *2*, 3. [CrossRef]
88. Kshetri, N. Big data's impact on privacy, security and consumer welfare. *Telecommun. Policy* **2014**, *38*, 1134–1145. [CrossRef]
89. Suthaharan, S. Big data classification: Problems and challenges in network intrusion prediction with machine learning. *ACM SIGMETRICS Perform. Eval. Rev.* **2014**, *41*, 70–73.
90. Dwork, C. Differential privacy. In *Automata, Languages and Programming*; Springer: Berlin, Germany, 2006; pp. 1–12.
91. Kairouz, P.; McMahan, H.B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A.N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. Advances and open problems in federated learning. *Found. Trends Mach. Learn.* **2021**, *14*, 1–210. [CrossRef]
92. Li, T.; Sahu, A.K.; Talwalkar, A.; Smith, V. Federated learning: Challenges, methods, and future directions. *IEEE Signal Process Mag.* **2020**, *37*, 50–60. [CrossRef]

93. Halevy, A.; Korn, F.; Noy, N.F.; Olston, C.; Polyzotis, N.; Roy, S.; Whang, S.E. Goods: Organizing Google's datasets. In *Proceedings of the 2016 International Conference on Management of Data*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 795–806.
94. Dong, X.L.; Srivastava, D. Big data integration. *Synth. Lect. Data Manag.* **2015**, *7*, 1–198. [CrossRef]
95. Lenzerini, M. Data integration: A theoretical perspective. In *Proceedings of the Twenty-First ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*; Association for Computing Machinery: New York, NY, USA, 2002; pp. 233–246.
96. Stonebraker, M.; Ilyas, I.F. Data integration: The current status and the way forward. *IEEE Data Eng. Bull.* **2018**, *41*, 3–9.
97. Doan, A.; Halevy, A.; Ives, Z. *Principles of Data Integration*; Morgan Kaufmann: Waltham, MA, USA, 2012.
98. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A survey of methods for explaining black box models. *ACM Comput. Surv.* **2018**, *51*, 1–42. [CrossRef]
99. Barredo Arrieta, A.; Diaz-Rodriguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
100. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef] [PubMed]
101. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv.* **2021**, *54*, 1–35. [CrossRef]
102. Holstein, K.; Vaughan, J.W.; Daume, H.; Dudik, M.; Wallach, H. Improving fairness in machine learning systems: What do industry practitioners need. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*; Association for Computing Machinery: New York, NY, USA, 2019; pp. 1–16.
103. Mittelstadt, B.; Russell, C.; Wachter, S. Explaining explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*; Association for Computing Machinery: New York, NY, USA, 2019; pp. 279–288.
104. Raji, I.D.; Smart, A.; White, R.N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; Barnes, P. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 33–44.
105. Davenport, T.H.; Patil, D.J. Data scientist: The sexiest job of the 21st century. *Harv. Bus. Rev.* **2012**, *90*, 70–76. [PubMed]
106. Chiang, R.H.L.; Goes, P.; Stohr, E.A. Business intelligence and analytics education, and program development: A unique opportunity for the information systems discipline. *ACM Trans. Manag. Inf. Syst.* **2012**, *3*, 1–13.
107. Debortoli, S.; Muller, O.; vom Brocke, J. Comparing business intelligence and big data skills. *Bus. Inf. Syst. Eng.* **2014**, *6*, 289–300. [CrossRef]
108. Vidgen, R.; Shaw, S.; Grant, D.B. Management challenges in creating value from business analytics. *Eur. J. Oper. Res.* **2017**, *261*, 626–639. [CrossRef]
109. Jobin, A.; Ienca, M.; Vayena, E. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* **2019**, *1*, 389–399. [CrossRef]
110. Floridi, L.; Cowls, J. A unified framework of five principles for AI in society. *Harv. Data Sci. Rev.* **2019**, *1*, 1–15.
111. Mittelstadt, B. Principles alone cannot guarantee ethical AI. *Nat. Mach. Intell.* **2019**, *1*, 501–507. [CrossRef]
112. Schwartz, R.; Dodge, J.; Smith, N.A.; Etzioni, O. Green AI. *Commun. ACM* **2020**, *63*, 54–63. [CrossRef]
113. Hu, H.; Wen, Y.; Chua, T.S.; Li, X. Toward scalable systems for big data analytics: A technology tutorial. *IEEE Access* **2014**, *2*, 652–687. [CrossRef]
114. Inmon, W.H. *Building the Data Warehouse*, 4th ed.; Wiley: Indianapolis, Indiana, 2005.
115. Kimball, R.; Ross, M. *The Data Warehouse Toolkit*, 3rd ed.; Wiley: Indianapolis, Indiana, 2013.
116. Chaudhuri, S.; Dayal, U. An overview of data warehousing and OLAP technology. *ACM SIGMOD Rec.* **1997**, *26*, 65–74. [CrossRef]
117. Stonebraker, M.; Abadi, D.J.; DeWitt, D.J.; Madden, S.; Paulson, E.; Pavlo, A.; Rasin, A. MapReduce and parallel DBMSs: Friends or foes. *Commun. ACM* **2010**, *53*, 64–71.
118. Pavlo, A.; Paulson, E.; Rasin, A.; Abadi, D.J.; DeWitt, D.J.; Madden, S.; Stonebraker, M. A comparison of approaches to large-scale data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*; Association for Computing Machinery: New York, NY, USA, 2009; pp. 165–178.
119. Zhang, Q.; Cheng, L.; Boutaba, R. Cloud computing: State-of-the-art and research challenges. *J. Internet Serv. Appl.* **2010**, *1*, 7–18. [CrossRef]
120. Vellido, A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Comput. Appl.* **2020**, *32*, 18069–18083. [CrossRef]
121. Zaharia, M.; Chowdhury, M.; Franklin, M.J.; Shenker, S.; Stoica, I. Spark: Cluster computing with working sets. In *Proceedings of the 2nd USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 10)*; USENIX Association: Berkeley, CA, USA, 2010; pp. 1–7.
122. Bu, Y.; Howe, B.; Balazinska, M.; Ernst, M.D. HaLoop: Efficient iterative data processing on large clusters. *Proc. VLDB Endow.* **2010**, *3*, 285–296.

123. Weber, K.; Otto, B.; Osterle, H. One size does not fit all: A contingency approach to data governance. *J. Data Inf. Qual.* **2009**, *1*, 1–27. [CrossRef]
124. Kolajo, T.; Daramola, O.; Adebiyi, A. Big data stream analysis: A systematic literature review. *J. Big Data* **2019**, *6*, 47. [CrossRef]
125. Sculley, D.; Holt, G.; Golovin, D.; Davydov, E.; Phillips, T.; Ebner, D.; Chaudhary, V.; Young, M.; Crespo, J.F.; Dennison, D. Hidden technical debt in machine learning systems. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 2503–2511.
126. Amershi, S.; Begel, A.; Bird, C.; DeLine, R.; Gall, H.; Kamar, E.; Nagappan, N.; Nushi, B.; Zimmermann, T. Software engineering for machine learning: A case study. In *Proceedings of the 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*; IEEE: Piscataway, NJ, USA, 2019; pp. 291–300.
127. Juma'h, A.H.; Li, Y. The effects of auditors' knowledge, professional skepticism, and perceived adequacy of accounting standards on their intention to use blockchain. *Int. J. Account. Inf. Syst.* **2023**, *51*, 100650. [CrossRef]
128. Hossain, M.K.; Srivastava, A.; Oliver, G.C.; Islam, M.E.; Jahan, N.A.; Karim, R.; Kanij, T.; Mahdi, T.H. Adoption of artificial intelligence and big data analytics: An organizational readiness perspective of the textile and garment industry in Bangladesh. *Bus. Process Manag. J.* **2024**, *30*, 2665–2683. [CrossRef]
129. Alnsour, Y.; Juma'h, A.H. The effect of political environment on security and privacy of contact tracing apps evaluation. *Int. J. Public Sect. Manag.* **2024**, *37*, 864–879. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.