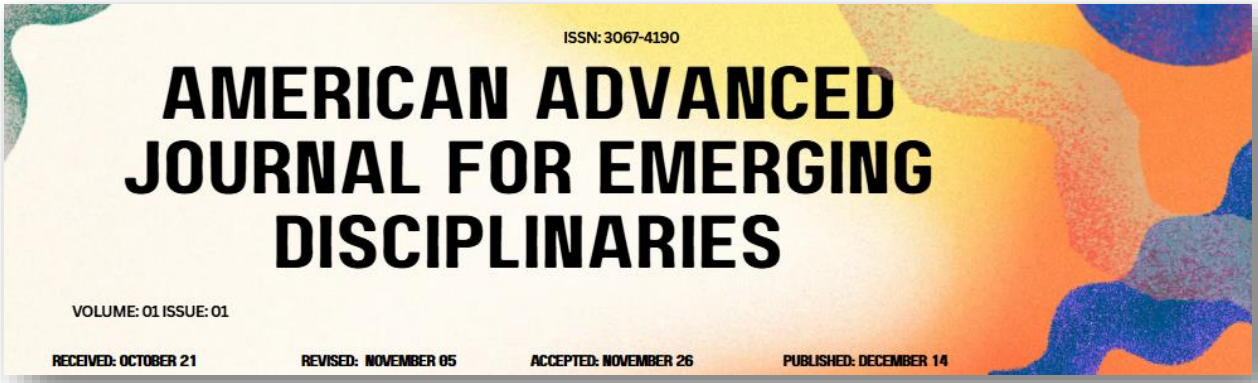


An Integrated Modernization Architecture for Data-Intensive Industries

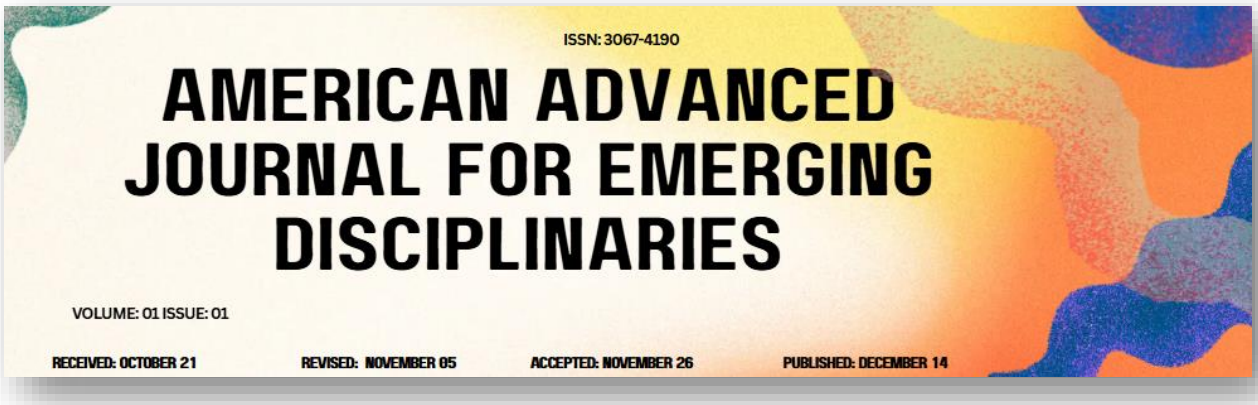
Dileep Valiki
Independent Researcher
valiki.dileep.researcher@gmail.com

Abstract

An AI-driven architectural framework for cloud-enabled data modernization in business domain was proposed. The framework helps organizations to efficiently endure the cross-sector convergence of cloud computing and Artificial Intelligence by addressing Data modernization on-demand and ensuring high-quality data for Artificial Intelligence algorithms. Data modernization leverages cloud computing in the Software as a Service delivery model, achieves Data product reliability and trustworthiness through Data governance, and meets usage requirements through scalability Data -including Supply, Processing, Storage, and Consumption layers- and Automation. Cross-sectoral deployment was validated by the foundation of the framework in three business domains (Retail, Healthcare, and Finance) and a detailed analysis of requirements. In parallel, the phased adoption of Data modernization was assessed through a maturity model and an Organizational and Team Readiness analysis.

Organizations from all business domains are converging towards the cloud, and so are several processes of Artificial Intelligence Data products. These two trends are fundamentally transforming the Data surrounding businesses by fueling Data modernization. Hyperscalers leverage their global network of Data centers to provide on-demand, scalable, and pay-per-use enablement of Data product development and operation. This unprecedented operational agility represents an attractive opportunity for organizations seeking fast response time and cost competitiveness. However, organizations often struggle in Data modernization, and Data products can suffer from inaccurate, incomplete, inconsistent, or outdated Data, thus risking unexpected behavior, poor prediction performance, or wrong decisions. AI-ready Data ecosystems –encompassing Data Supply, Processing, Storage, and Consumption– that ensure quality Data should be made available for low-latency AI-enabled analytics.

Keywords: AI-Driven Data Modernization, Cloud-Native Data Architecture, Intelligent Data Ecosystems, Data Product Engineering, AI-Ready Data Platforms, Scalable Data Pipelines, Data Governance Frameworks, SaaS Data Delivery, Cross-Domain Data Integration, Hyperscaler Cloud Systems, Data Quality Assurance, Low-Latency Data Analytics, Automated Data



Workflows, Data Lifecycle Management, Enterprise Data Transformation, Data Reliability Engineering, Adaptive Data Infrastructure, AI Data Readiness, Data Maturity Models, Organizational Readiness Analysis.

1. Introduction

Business data modernization services in cloud environments enable organizations to complement traditional data engineering efforts by integrating a new set of cross-domain data pathways between operational systems. These services allow unmediated access to purpose-designed data representations for artificial intelligence algorithms via feature stores, reducing the need for bespoke data engineering dedicated to a single model. Retail, healthcare, and finance sectors are prominent domains within the business sphere. A scalable and efficient data and analytics ecosystem for these domains requires specialized data pipelines, processing methods, and supporting design principles.

Recent years have seen a spike in digital analytical capabilities across organizations. The international economy is entering an intensified phase, with business data and artificial intelligence entering new areas, development moving from a low base to a mature level, project budgets and stakeholder demand growing rapidly, and digitalization driving data demand. Such increased interest in AI is typically accompanied by a large number of proofs of concepts, which place extra strain on data engineering teams responsible for the orchestration and preparation of data for a given model, or data product.

2. Theoretical Foundations of Data Modernization

Data modernization covers conceptually distinct yet interrelated adaptations of technology, processes, and data assets that support organizational and societal transformation. With data and Artificial Intelligence at the forefront of the next wave of digital transformation, cloud technologies enable broader adoption. Additionally, generative AI is ready to disrupt every industry yet needs high-quality data pipelines to work reliably. Successfully tackling the enabling factors is becoming increasingly critical, but the adoption of AI-ready data and cloud services remains low. To address this need, a data modernization framework for three cross-industry domains — retail, healthcare, and finance — was designed within Microsoft.

The factors that enable Google, Amazon, and Microsoft to grow and modernize require the same investment by others for the next wave of AI-powered cloud-enabled digital

ISSN: 3067-4190
AMERICAN ADVANCED JOURNAL FOR EMERGING DISCIPLINARIES
VOLUME: 01 ISSUE: 01
RECEIVED: OCTOBER 21 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 26 PUBLISHED: DECEMBER 14

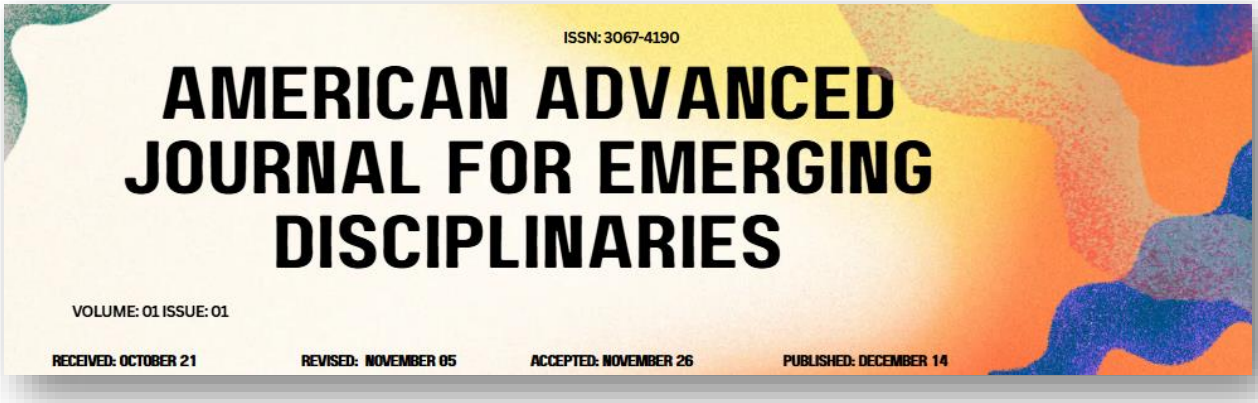
transformation. These foundational enablers cluster around five attributes considered essential for mass-market adoption: Data Lineage, Interoperability, Governance, Scalability, and Automation. Each attribute maps to concrete architectural build, operating, or engineering guidelines that provide specialized direction for practitioners focused on any of the three domains.

Fig 1: Benefits of Data Modernization

2.1. Key Concepts in Data Modernization Framework

Data modernization encompasses a variety of interconnected concepts defining data technology architecture, processes, operations, and stewardship to underpin a cloud-hosted data ecosystem capable of generating AI-ready data products or services spanning multiple organizations. Although shared across participating organizations, architecting, building, operating, and governing the complete data ecosystem requires fulfilling specific considerations in quality, speed, security, privacy, regulatory compliance, and cost for each organization at a given point in time.

Data lineage indicates the life cycle and usage of data items originating in source systems, traveling through the data pathway, being stored in data lakes and warehouses, and finally supplying target systems. Data lineage serves as an automation enabler supporting the above-mentioned architecting, building, operating, and governing of the data ecosystem by providing analytics capabilities that cover questions regarding privacy, security, internal regulations, and external compliance when products using or data-sharing packages transfer these data items. A secondary aspect of data lineage for steering the speed faculties relates to



latency requirements, describing the maximum elapsed time from the event generating the data until the delivery of a data product.

3. Methodology

A design-oriented research approach is adopted to develop the framework, incorporating the guidelines into an analytical tool. The researcher's organization operates in retail, healthcare, and finance, enabling cross-domain applicability evaluation by domain authorities and professionals. Validation involves case study scenarios within these sectors.

Scalability is a common requirement for modern operations and applications, as verified by organizational readiness and preparedness assessments. Organizations often operate data pipelines spanning public and private infrastructures and make usage-and-cost trade-offs at the processing-unit level. Architectural guidelines covering data movement, storage, processing, and preparation pathways facilitate a cloud-scale operating model, readily deployed in serverless or containerized environments.

Equation 1. Service Utilization

$$\text{Utilization} = \frac{\text{Actual Output or Actual Used Capacity}}{\text{Maximum Possible Output or Provisioned Capacity}}$$

Step-by-step derivation

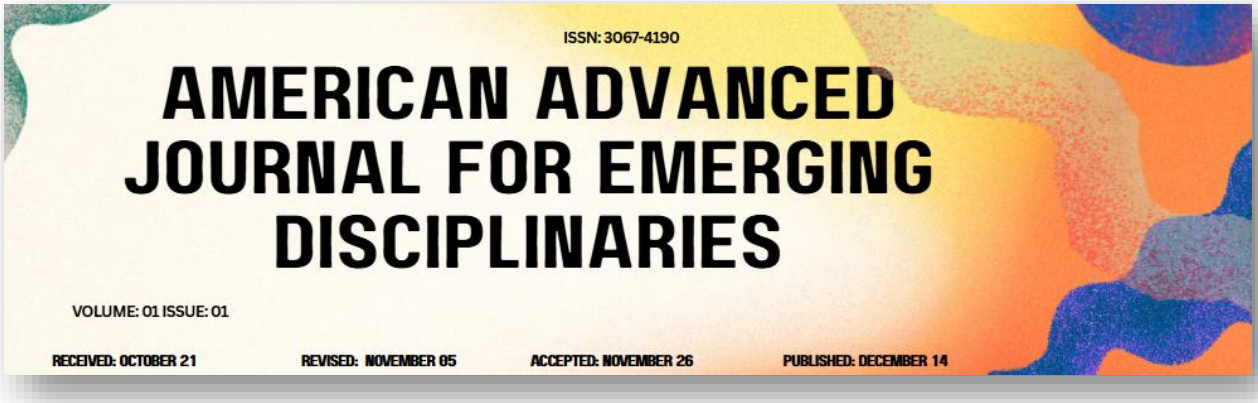
Let:

- A = actual used compute/storage/network capacity in a time window
- M = maximum available or provisioned capacity in the same time window

By definition, utilization is the fraction of available capacity that was truly used:

$$U = \frac{A}{M}$$

To express it as a percentage:



$$U_{\%} = \frac{A}{M} \times 100$$

Example

If a compute service can process 10,000 jobs/hour but actually processes 7,500 jobs/hour:

$U = \frac{7500}{10000} = 0.75$ $U_{\%} = 0.75 \times 100 = 75\%$

3.1. Architectural Guidelines for Efficient Cloud Infrastructure

Efficient and effective cloud infrastructure is essential for realizing AI-ready data ecosystems at scale. Achieving this relies partly on choosing an appropriate cloud engineering approach, supported by cross-domain requirements on modularity, elasticity, cost awareness, and fault tolerance. These principles guide practitioner decision-making in their use of cloud platforms—public, private, or hybrid—as well as tooling choices for implementing shared data paths, persistent data stores, compute models, and integration mechanisms.

Taking a multi-cloud or hybrid approach aids both choice and dispersion, circumventing vendor lock-in while allowing distinct services to underpin specific data tasks—such as big storage, low-latency delivery, or specialized application APIs—without a single provider being optimal across the board. Portability concerns further motivate this strategy, particularly for tasks needing infrequent burst capacity. Nevertheless, dispersing cloud activity requires careful governance, since different providers' controls and regulations often diverge.

Table 1. Core framework components

Layer / Component	Purpose	Main function in the article	Key benefit
Data Supply	Bring data from operational sources into the platform	Ingestion from multiple systems and domains	Centralizes fragmented data
Data Processing	Clean, transform, normalize, enrich	Supports batch, near-real-time, and real-time paths	Improves AI-readiness

ISSN: 3067-4190
AMERICAN ADVANCED JOURNAL FOR EMERGING DISCIPLINARIES
VOLUME: 01 ISSUE: 01
RECEIVED: OCTOBER 21 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 26 PUBLISHED: DECEMBER 14

Layer / Component	Purpose	Main function in the article	Key benefit
Data Storage	Persist data economically and reliably	Data lakes, warehouses, feature stores, hot/warm/cold tiers	Balances cost, access, and latency
Data Consumption	Serve analytics, reports, and AI applications	BI, model training, model serving, downstream APIs	Delivers business value
Governance	Control trust, quality, ownership, compliance	Stewardship, lineage, policy enforcement	Raises reliability and compliance
Scalability	Handle growth in users, workloads, and data	Serverless, containers, hybrid/multi-cloud	Avoids bottlenecks
Automation	Reduce manual engineering overhead	Pipelines, CI/CD, metadata-driven operations	Faster and more repeatable delivery

4. Objectives of the Study

Data-intensive environments evolve rapidly; hence, architecture must support dev/test or proof-of-concept services but permit scaling as need, use-cases, or workloads announce. The adoption of cloud resources is thus incremental and governed by owned on-premises/colocation resources and TCO analytics. A maturity model helps consolidate the infrastructural landscape and assess cloud use across enablers, promoted skills and governance readiness, team setup, and maturity drivers. Expanding beyond TCO, an assessment of the data needed to train operational, support, and decision-making products exposes quality, quantity, signal rich/poor, or fresh/old issues.

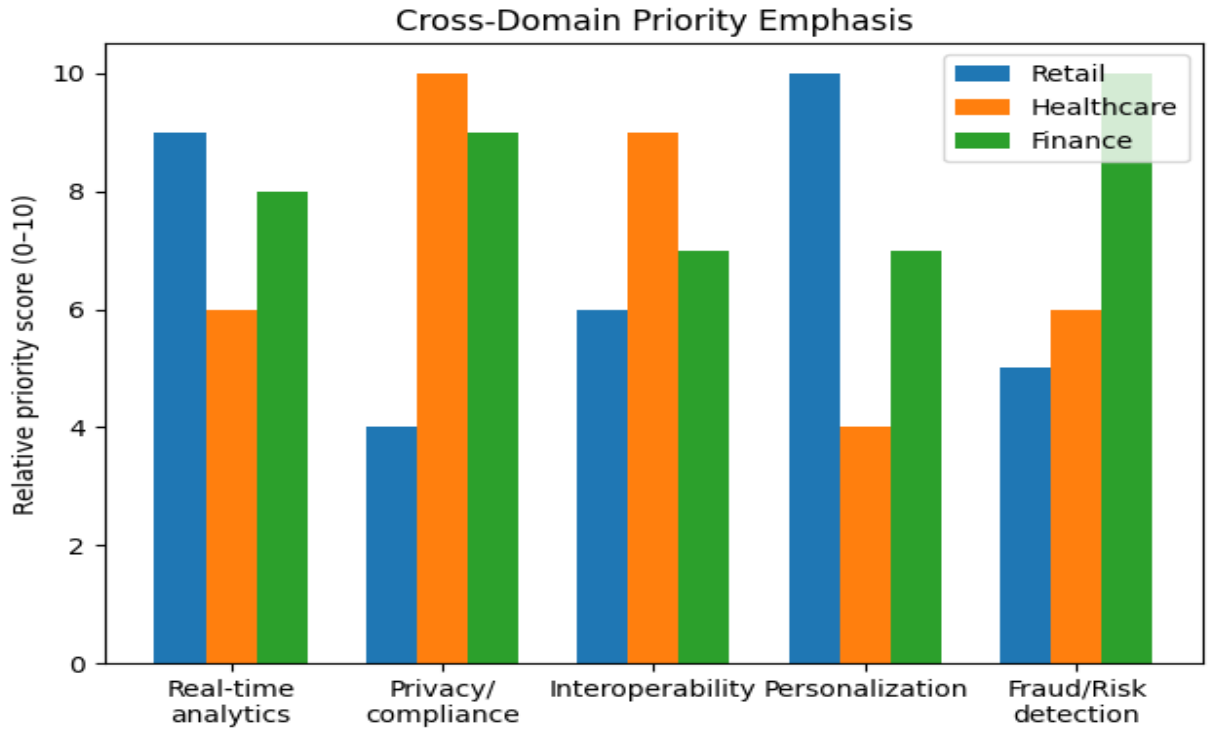
A data line with sufficient maturity serves products whose cost drivers relate to utilization—compute or storage resource costs, time-compliance aspects, and TCO, or business performance aspects and consistency in sales and ad revenue. Data line dev/test for specialized needs or trial purposes often impose latency penalties on other products—personalization, inventory optimization, and campaign performance analysis. As the ecosystem matures, it becomes instrumental in facing business requirements whose cost metrics fail to meet land, compliance, financial, or strategic requirements.

ISSN: 3067-4190

AMERICAN ADVANCED JOURNAL FOR EMERGING DISCIPLINARIES

VOLUME: 01 ISSUE: 01

RECEIVED: OCTOBER 21 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 26 PUBLISHED: DECEMBER 14



4.1. Study Aims and Research Questions

A synthetic answer to these questions directing the framework development is: provide a scalable architecture for cloud platforms that crosses cloud borders for delivering data products, with a special focus on AI-ready data. A good cloud architecture is modular and supports elasticity, low TCO, reduced complexity, maximum data utility, and automated operations.

Scaling solutions for AI-ready data that more than one organization consumes are extremely different for different data suppliers. These differences stem from the nature of the data suppliers and the data products they build, as is illustrated in the following in a retail setting. AI-Ready data products in retail are built for personalization, real-time analytics, or loyalty analytics. The data needed in emerging personalization pipelines come from multiple sources, have low latency requirements, and must be constantly fused. In contrast, the data inherent to

ISSN: 3067-4190

AMERICAN ADVANCED JOURNAL FOR EMERGING DISCIPLINARIES

VOLUME: 01 ISSUE: 01

RECEIVED: OCTOBER 21 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 26 PUBLISHED: DECEMBER 14

steering a loyalty program is analytical in nature and is normally generated in batch mode, with limited freshness requirements. The AI product in real-time pricing relies on perishable products and hence requires additional latency consideration, such as that imposed by the freshness of pricing signals from customers and competitors.

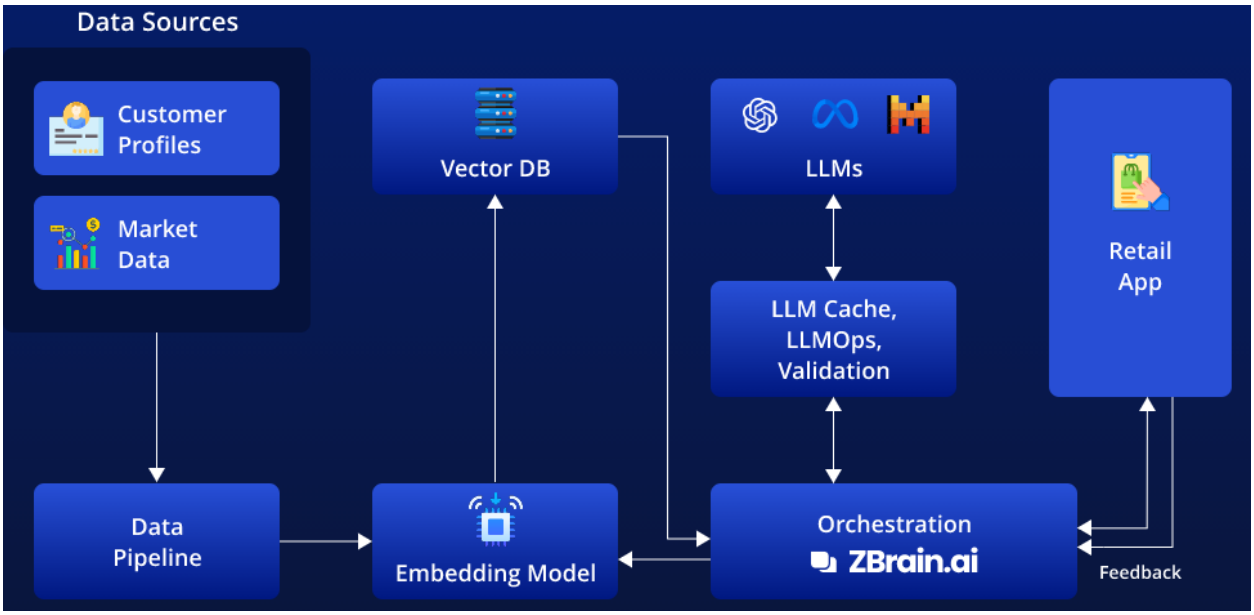
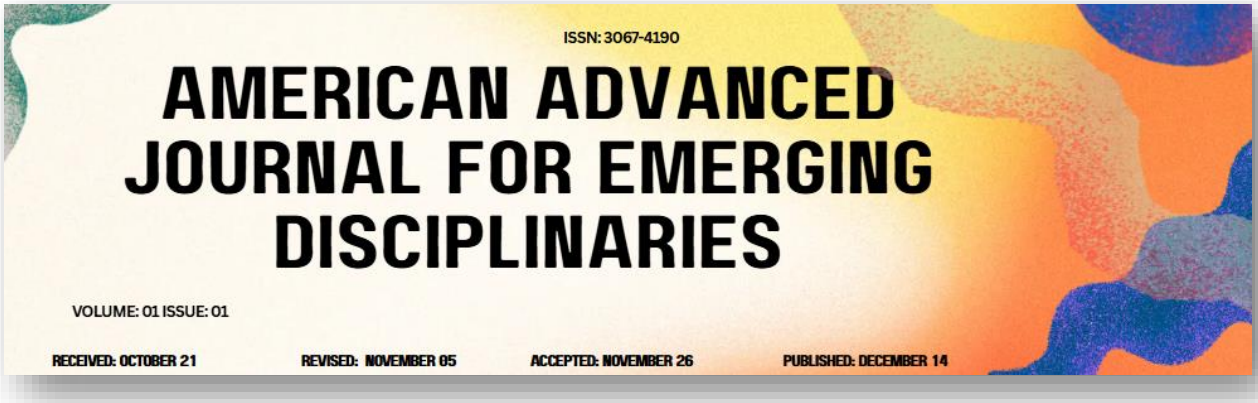


Fig 2: AI in retail

5. Research Summary

Cloud engineering plays an increasingly crucial role in supporting data strategy decisions that drive modern cloud-scale data pipelines, features, models, and AI solutions. The growing presence of data in everyday business life, coupled with the rise of cloud technologies, has led to efforts to leverage cloud deployment for entering the AI sphere. However, data still remains a limiting factor—without quality data that is present in large amounts, AI applications are unlikely to reach their full potential. For cloud-centered organizations aiming for AI, establishing and operating scalable cloud data pathways for pipelines or feature generation and technology for consumption is imperative. Several AI cloud data product pathways are mentioned as representative examples.



While cloud service providers have made enormous investments in different tools and technologies, organizations have been slow in leveraging cloud technologies to enable AI and building the additional cloud data and AI products to support their data-driven services. This resulted in a few guidelines to support growth and scaling of data pathways to enable a scalable data strategy for AI products. Building these guidelines requires a clear understanding of the requirements and demands of different domains, such as retail, healthcare, finance, and other domains. However, the wish for pursuing a cloud strategy combined with AI is not restricted to a single domain. Therefore, it is essential to lay out the common requirements that exist across domains for achieving an AI data ecosystem and what is needed, in more detail, to support implementation in a specific domain.

5.1. Core Architectural Guidelines for Cloud Scalability

For effective scaling of data pathways, storage, and processing, response time remains a paramount architectural goal. Consequently, data paths between data sources and sinks should be optimized for direct delivery, whether in real-time, near-real-time, or batch mode, such that the target system's freshness requirements are fulfilled. Storage tiers should be partitioned into hot storage, intended for data constantly required for operations, warm storage, composed of explode46 infrequently accessed datasets that remain relatively usable for exploration, and cold storage, hosting data only to fulfill regulatory requirements or infrequent inquiries. Finally, processing should be performed in function, workflow, or pipeline mode, as appropriate for the target environment, with a reasonable periodical latency that stays close to the target response quality.

In addition to maintaining response time, architectural design must ensure that the underlying cloud environment can absorb peak usages. However, overdimensioning resources implies a direct increase in operational costs, being generally unfeasible for systems that cannot recover their investments. Therefore, the ideal setup should optimize scale usage towards a low-cost, reliable, yet efficient architecture. For specific needs like ingestion, only volume-based charges such as those common in Big Data events should be considered.

ISSN: 3067-4190
AMERICAN ADVANCED JOURNAL FOR EMERGING DISCIPLINARIES
VOLUME: 01 ISSUE: 01
RECEIVED: OCTOBER 21 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 26 PUBLISHED: DECEMBER 14

Equation 2. Throughput

Step-by-step derivation

Let:

- D = total data processed
- T = total elapsed time

By definition, throughput is:

$$\text{Throughput} = \frac{D}{T}$$

If D is measured in GB and T in hours:

$$\text{Throughput} = \frac{\text{GB}}{\text{hour}}$$

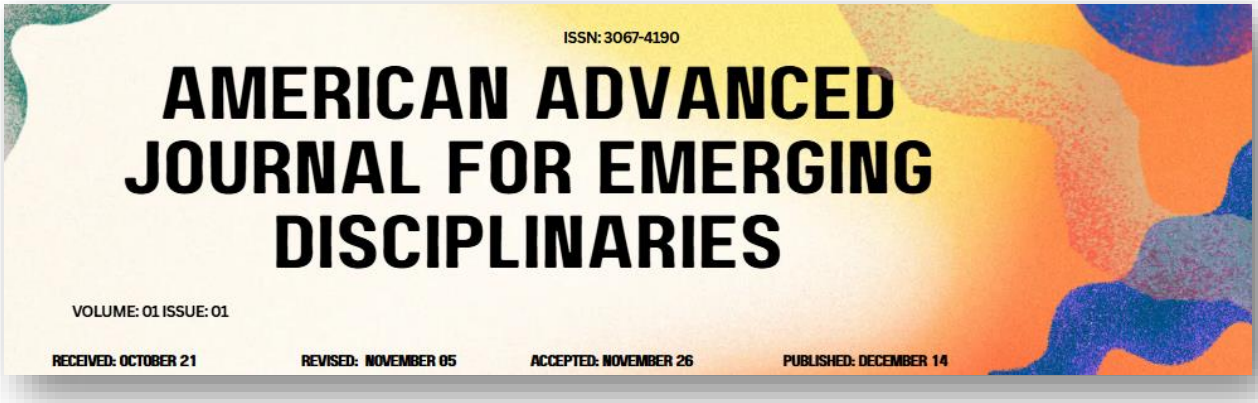
Example

If a pipeline processes 240 GB in 6 hours:

$$\text{Throughput} = \frac{240}{6} = 40 \text{ GB/hour}$$

6. Architectural Principles for Scalable Cloud Engineering

Organizational resources undergo steady erosion due to digital workplace demands and dispersed, elastic resources mismanagement. Well-governed, efficient and reliable management of such environments revolves around 14 architectural principles. Deployment of repetitive AI/ML models consumes time and expertise since traditional data management processes are not transformed. Hosting of AI/ML workloads with adequate business and technical data readiness is hindered by automation priorities. Environmental and resource considerations are sought for development and routine execution of AI/ML workloads.



Data-related influence in operational and training model pathways demands the following specialized techniques and practices: an AI-ready ingestion of large-volume heterogeneous data with adequate governance; a unified data lake for efficient storage of data with diverse velocity; a feature store simplifying and accelerating model training; and a well-governed architecture for deploying predictive data products. With its operational models extending horizontally across digital business ecosystems and vertically across budgetary domains via hybrid allocation modes, scalability combines storage and processing decoupling via data and code virtualization in tiered, serverless and containerized settings. This independence fosters the selection of L1, L2, and L3 resources according to demand and available budget.

The technical guidance above serves to enable cross-domain pathways configured for eased data sharing and model development. Synthesis of domain representatives throughout cloud and data industrialization phases serves to define minimal-shared components, integration-and-sharing contracts, and connection-pHYHFDSSOLFDWLRQV-sOLFDWLRQV-sROLFVDV-t dedication and availability.

6.1. Data Governance and Stewardship

Effective data governance and stewardship are key to improving data quality, risk, and compliance within organizations. Governance establishes a high-level decision structure and accountability for entrusted data assets. Employing clear policies enables organizations to define their data quality targets, monitor the fulfillment of these goals, and initiate corrective actions when necessary. On the other hand, data stewardship infuses these policies with specialized expertise by assigning data owners to certain business areas, domains, or sources and defining the corresponding stewardship responsibilities. Good stewardship restores the balance between the need to share data and the quality and risk mitigation requirements that limit this sharing, especially in complex multi-stakeholder scenarios such as those in the healthcare and finance industries.

ISSN: 3067-4190

AMERICAN ADVANCED JOURNAL FOR EMERGING DISCIPLINARIES

VOLUME: 01 ISSUE: 01

RECEIVED: OCTOBER 21 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 26 PUBLISHED: DECEMBER 14

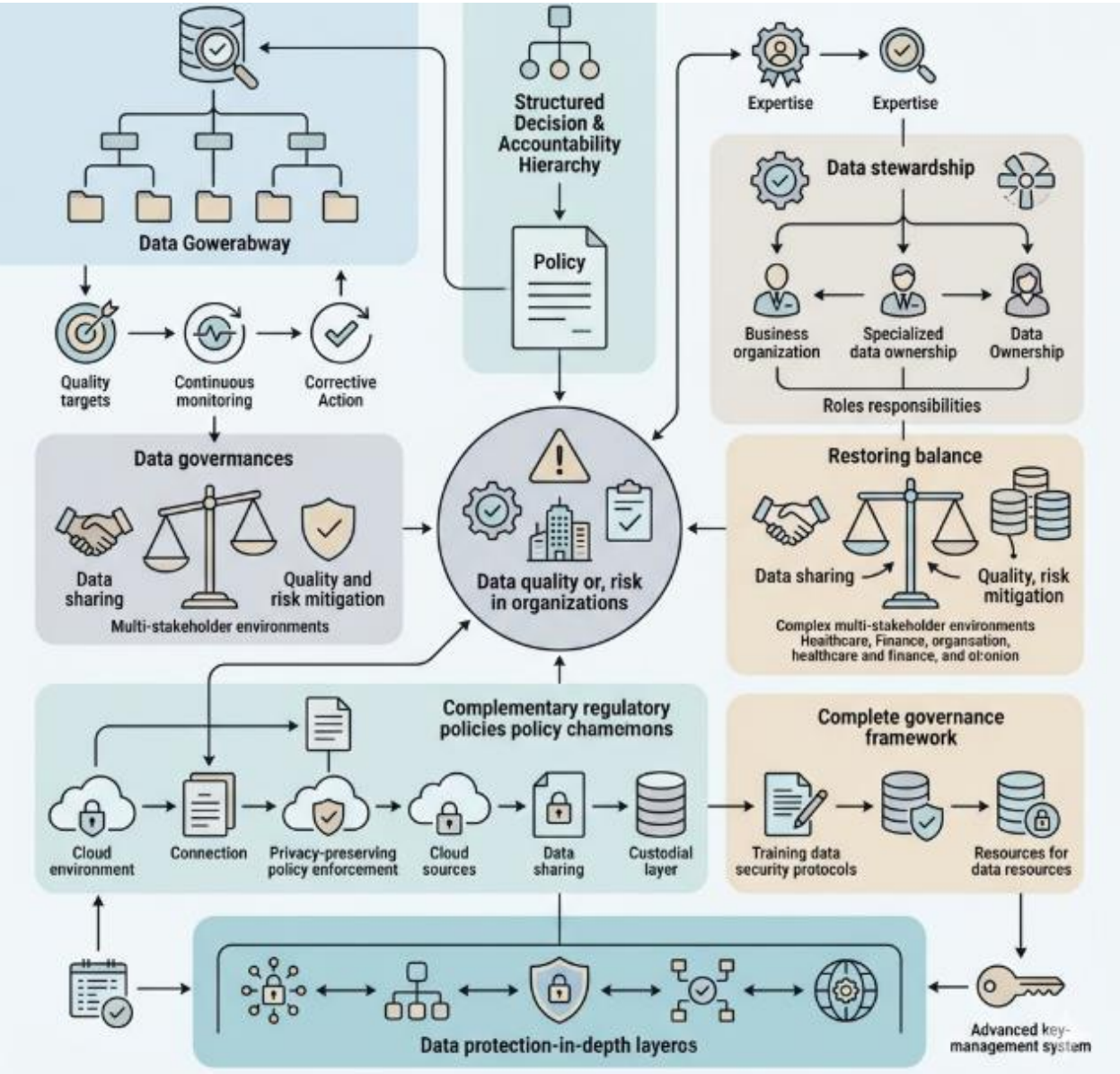
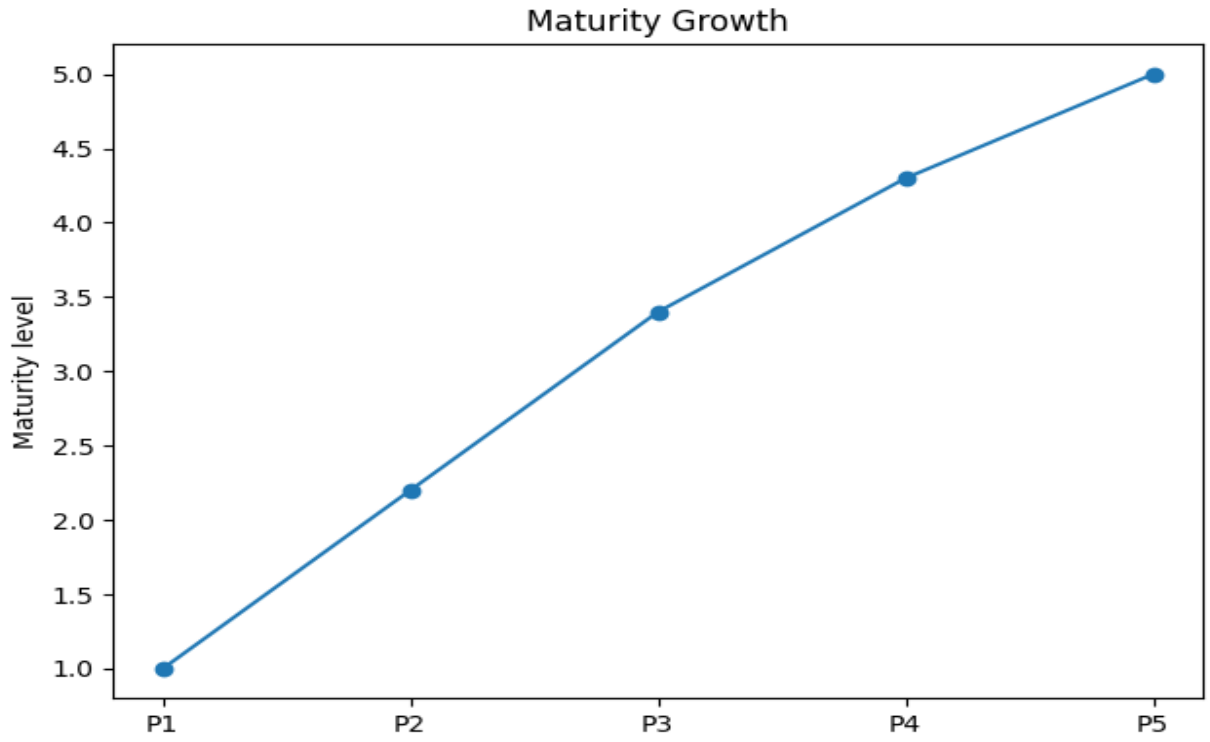


Fig 3: Synergizing Governance and Stewardship for High-Integrity, Compliance-Ready Data Systems

ISSN: 3067-4190
AMERICAN ADVANCED JOURNAL FOR EMERGING DISCIPLINARIES
VOLUME: 01 ISSUE: 01
RECEIVED: OCTOBER 21 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 26 PUBLISHED: DECEMBER 14

The accountability and responsibilities frameworks established through governance and stewardship should further be complemented with suitable regulatory policies that address the specific industry data characteristics. In the healthcare sector, privacy-preserving policy enforcement mechanisms become critical in data-sharing scenarios, especially over the cloud, and, as such, must be integrated with the custodial and decision-making layers required to form a complete governance framework. In any artificial-intelligence-supporting data ecosystem, training data is secured through set formal methods and intelligently sharing such data resources introduces a broad range of new requirements affecting both operational efficiency and user trust. At the core of this data management strategy is a comprehensive data protection-in-depth approach that employs specific techniques at different architectural levels, all supported by an advanced key-management system.



6.2. Metadata Management and Discoverability

ISSN: 3067-4190
AMERICAN ADVANCED JOURNAL FOR EMERGING DISCIPLINARIES
VOLUME: 01 ISSUE: 01
RECEIVED: OCTOBER 21 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 26 PUBLISHED: DECEMBER 14

Configurable metadata schemas provide a foundation for internal and external discoverability, while catalogs offer a centralized view of all data and artifacts. Tracking data lineage enhances understanding of data production and usage, enabling data quality assessment and targeted quality improvement efforts.

Data and algorithm catalogs act as user-facing layer for metadata. The catalog should comprise a unified view of all information products within the organization, including datasets, collection pipelines, stored features, AI models, and any other consumable information entity. The catalog should also expose properties published in the relevant metadata schema. For core datasets, a full-fledged data encyclopedia should be considered, capturing elements such as the dataset definition, associated use cases, data preparation procedures and code, data quality metrics and statistics, and so on.

Metadata management capabilities are required across all platforms, facilitating self-service data consumption while preventing the creation of data silos by monitoring data usage patterns and ensuring that frequently consumed datasets are easily discoverable. Furthermore, metadata should be appended to data and rules governing its use enforced automatically via data masking, encryption, and other measures. Besides information product cataloging, a key ingredient to efficient data consumption is data lineage monitoring. Capturing lineage information of data products in all systems enhances understanding of their production processes and supports data quality monitoring and identification of data preparation procedures that bring the most quality benefits for analysts and models consumers.

Table 2. Cross-domain requirements comparison

Domain	Primary drivers	Latency sensitivity	Compliance intensity	Typical AI/data use cases
Retail	Personalization, customer analytics, loyalty, pricing	High	Medium	Recommendations, churn, campaign targeting, real-time pricing
Healthcare	Privacy, interoperability, compliant sharing	Medium to High	Very High	Clinical analytics, federated learning, secure sharing

ISSN: 3067-4190
AMERICAN ADVANCED JOURNAL FOR EMERGING DISCIPLINARIES
VOLUME: 01 ISSUE: 01
RECEIVED: OCTOBER 21 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 26 PUBLISHED: DECEMBER 14

Domain	Primary drivers	Latency sensitivity	Compliance intensity	Typical AI/data use cases
Finance	Fraud detection, AML, transaction intelligence, personalization	High	Very High	Risk scoring, fraud monitoring, compliance analytics

7. Cross-Domain Requirements for Retail, Healthcare, and Finance

Despite considerable variation within the domains, certain cross-sector data modernization requirements have emerged. At a high level, the attributed importance of key factors varies between domains. For instance, organizations within the customer-facing retail industry are more likely to consider personalization and real-time analytics among the top three features helping drive customer loyalty. In addition, personalization in retail is more about interaction and thus has a requirement for small-latency events. For the healthcare sector, patient data privacy is paramount, and the ability to encrypt data, as well as comply with HIPAA and GDPR regulations, greatly supports data sharing with the aim of accelerating scientific discovery and advancing operational analytics.

Organizations within the finance industry must comply with strict anti-money laundering regulations; thus, having the ability to integrate third-party transaction data for real-time detection and prevention remains a top driver. At the same time, there is a need for the use of customer signals—such as sentiment analysis on social media, transactional activities, and compliance with Bank Secrecy Act/Anti Money Laundering regulations—at micro levels to generate personalized financial products. The external ability and compliance with any initiative by a different government also help organizations provide better offerings and solutions to their customers without directly being involved in the process.

Equation 3. End-to-End Latency

Step-by-step derivation

Let:

- t_s = source event creation time

ISSN: 3067-4190
AMERICAN ADVANCED JOURNAL FOR EMERGING DISCIPLINARIES
VOLUME: 01 ISSUE: 01
RECEIVED: OCTOBER 21 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 26 PUBLISHED: DECEMBER 14

- t_i = ingestion completion time
- t_p = processing completion time
- t_c = consumer delivery/availability time

Total end-to-end latency is:

$$L = t_c - t_s$$

It can also be decomposed:

$$L = (t_i - t_s) + (t_p - t_i) + (t_c - t_p)$$

So:

$$L = L_{\text{ingest}} + L_{\text{process}} + L_{\text{serve}}$$

Example

If:

- ingestion takes 2 sec,
- processing takes 4 sec,
- serving takes 1 sec,

then:

$$L = 2 + 4 + 1 = 7 \text{ sec}$$

7.1. Retail: Personalization, Real-Time Analytics, and Loyalty Data

In retail, customer signals are an essential part of understanding shopping behavior as they indicate intentions such as return, purchasing, and abandonment. For online shops, road signs indicate recent website visits, region of origin, product preferences, and so on. For brick-and-mortar stores, traffic counters can provide information on store visits, store and aisle dwell

ISSN: 3067-4190
AMERICAN ADVANCED JOURNAL FOR EMERGING DISCIPLINARIES
VOLUME: 01 ISSUE: 01
RECEIVED: OCTOBER 21 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 26 PUBLISHED: DECEMBER 14

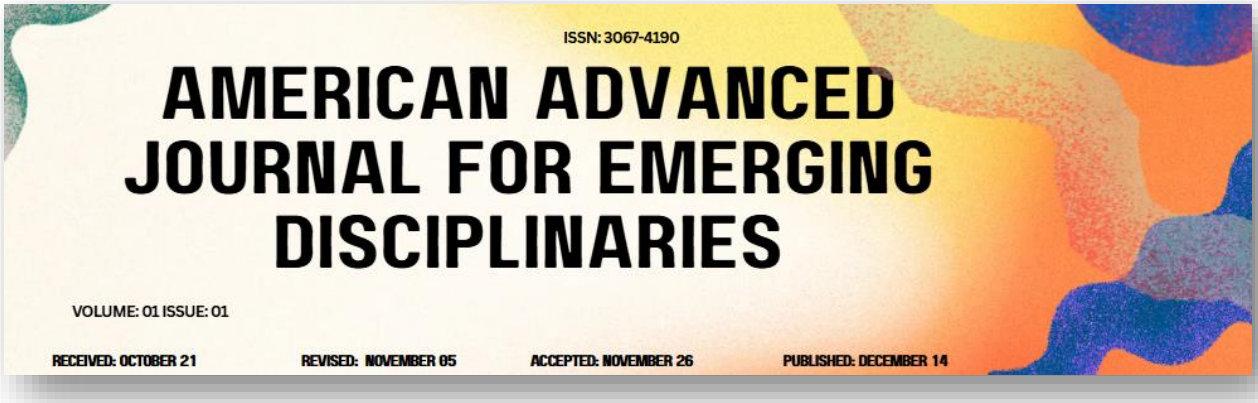
time, and the demographics of store visitors. These data, however, are usually captured in separate silos. The goal is thus to perform data fusion on those data so that new insights can emerge. Although classification and regression are the more straightforward use cases, these signals open up other exciting use cases such as sentiment analysis from review and feedback data, competitive analysis using web scraping solutions, and churn prediction from web, electronic communication, and loyalty data fusion. In practice, the challenge is that many of these signals are captured from the above-mentioned systems and databases but do not reach the category of real time.

They are processed on a periodic basis, such as daily, weekly, monthly, or yearly; it is, therefore, imperative that the latency requirements are determined, and only the signals that need to be fused on a real-time basis are selected as such. For these signals, online learning algorithms can be considered to deliver predictive models in real time and capture the data distribution drift. The detected drift can in turn trigger the standard periodic processing and retraining of models in a batch fashion. Chatbot marketing is another area that financial organizations are exploring; some model deployment strategies allow for automated training of the chatbot alongside natural language processing testing and training with test-driven development paradigms.

7.2. Healthcare: Patient Privacy, Compliance, and Interoperability

The health industry places the greatest requirements on sensitive data, as patient privacy is paramount. Data is often governed by regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in the European Union. Organizations active in the health industry need to ensure that data does not leave the control of the organization to other untrusted environments; when images are sent to the cloud for inferencing, all other data must come from the same cloud. Unlike for the retail industry, where certain types of analytics naturally have a low-latency requirement, the health industry must also pay attention to the latency of data models. Furthermore, certain aspects such as low-quality data and conclusions reached on a low data quality lifecycle are more critical than in the retail industry. The Healthcare Industry, especially in Europe, cannot afford such low-quality conclusions and models.

Due to the nature of the sensitive data being processed, compliance with the regulations is often the endpoint of the analysis. GDPR and HIPAA consist of dozens of individual rules, making it difficult to quickly ascertain whether an environment is compliant. Analysis against



these regulations is often neither complete nor thorough enough, especially when considering not only compliance but also the impact of an income statement on future models and income flows. As privacy regulations evolve in the near future to make these regulations even stricter, setting up an environment with data that comply with the rules combined with an analytics pipeline that generates results even for privileged users is becoming more important.

8. AI-Driven Data Modernization Techniques

Preparing data for AI product development is an important aspect of an AI data product lifecycle. The frameworks address the ingestion of data into a unified data lake and the construction of AI data product feature stores, which support supervised learning tasks but often are necessary for a variety of other AI products. By addressing the ingestion of all types of data, the construction of feature stores, and the exposure of those features for operational consumption, the techniques are applicable to a wide range of AI data products.

A unified data lake provides a central repository for all data in the enterprise. However, such environments often have different characteristics for the data that reaches the lake and for the processing of that data. The establishment of a common set of preferred operational data formats goes a long way in preventing a messy and complex environment. The ingestion pipelines that normalize data coming from different sources and systems into a common structure make the data more usable.

8.1. Data Ingestion and Unified Data Lakes

Data readiness requires careful planning and execution. Data ingestion pipelines are often modular, processing data independently of others. Data acquired from various sources might not exist in the same format, especially when the source systems have been developed by different teams separated by not only the business unit but also by such factors as geographical boundaries. The errors and the noise associated with the data from these different systems might make simply dumping them in the same location an incorrect proposition. Hence, data normalization is often carried out in the data ingestion pipelines. If, however, the data from the systems is of a relatively good quality, then data does not need to undergo this step. Regardless of whether data normalization is undertaken in the ingestion pipeline, trying to store all of this data in standard low-cost storage is often advisable. Data that is not in required shape is directed to stored standard cloud storage solutions to easily and cost-effectively perform any run-time

ISSN: 3067-4190
AMERICAN ADVANCED JOURNAL FOR EMERGING DISCIPLINARIES
VOLUME: 01 ISSUE: 01
RECEIVED: OCTOBER 21 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 26 PUBLISHED: DECEMBER 14

data transformations and within the data lake itself rather than on supercomputing resources used for model training.

The unified data lake becomes the master data store for all AI and ML modeling and analytics requirements. Data from the feature store drives AI &ML model inference and online real-time analytical requirements. This ensures maximum sharing of data at the lowest latency and also provides the single view of the customer for loyalty management, custom ads, and recommendations. The data product requirement from the business traversing these pathways creates an active data ecosystem for managing data lifecycle.

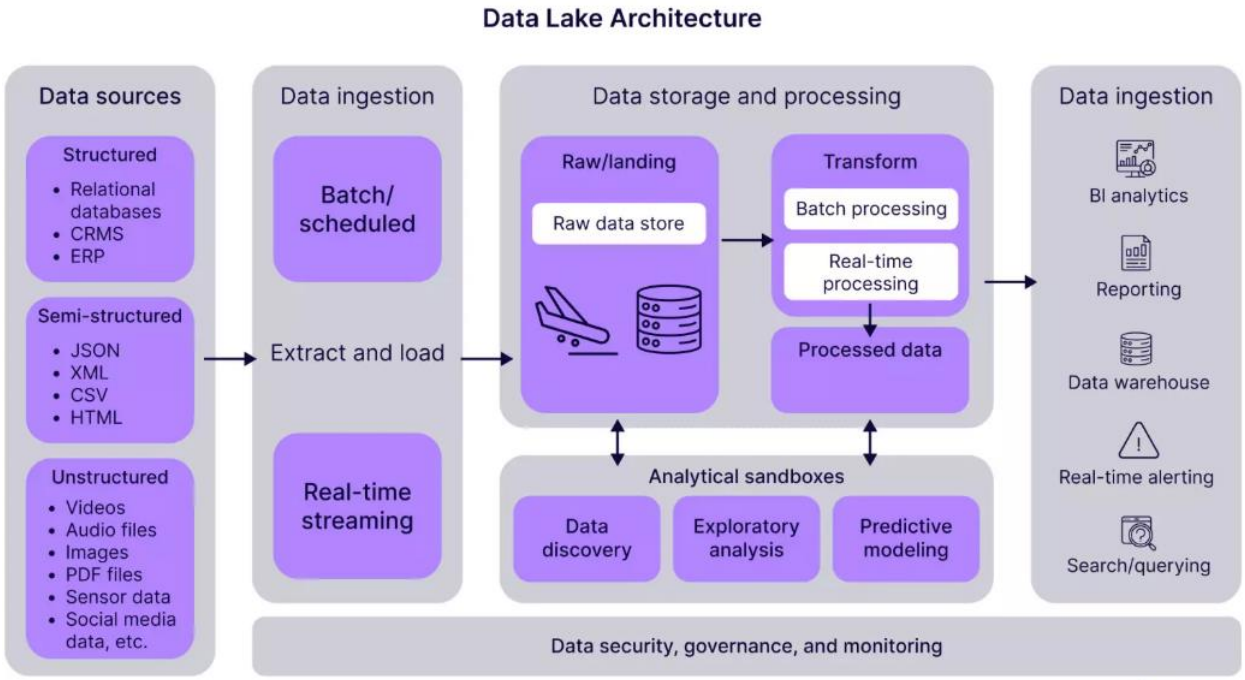
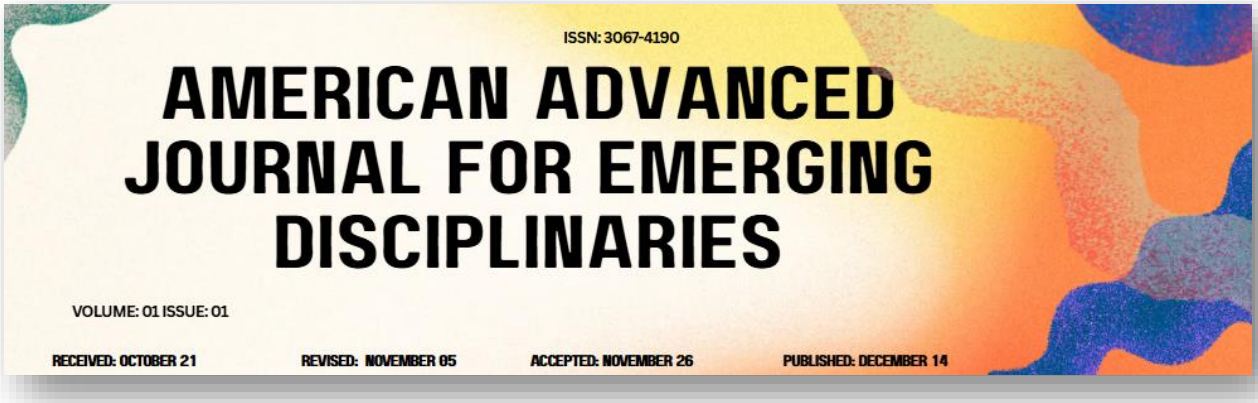


Fig 4: Modern Data Lake Architecture

8.2. Feature Stores and Model Serving

To be exploitable in machine-learning applications, data from various sources must be unified, transformed into appropriate representations, and hosted in environments easily



accessible by model's training and serving components. The process of supporting these requirements constitutes the feature-store ecosystem. A feature store is an AI-ready management system for data generated for artificial-intelligence purposes, supporting the full life cycle of the data products that are machine-learning features. Dedicated storage and specialized-serving technologies can be provided to ensure performance and low latency during the consumption of the features by models that are deployed in production. Feature stores are widely considered necessary to establish a successful feature-data management framework and its establishment is, however, often underestimated.

Data need to be prepared before training machine-learning models. The data products consumed at this phase are features—predictive or explanatory factors used to predict target variables—and are produced by dedicated feature-induction pipelines or directly by a feature store. Features can simply be the raw data from different data sources or the output of data-transformation pipelines that may include processes like outlier detection, normalization, or encoding. However, such a scenario is not product scalable since, when the same feature set is constantly produced and made available, the process can easily become a bottleneck during the training phase. This situation is what a feature store aims to avoid. A feature store can be seen as a dedicated life-cycle-management system for features.

Equation 4. Data Freshness

The emphasizes freshness requirements for target systems.

Freshness is how old the delivered data is at the moment of use.

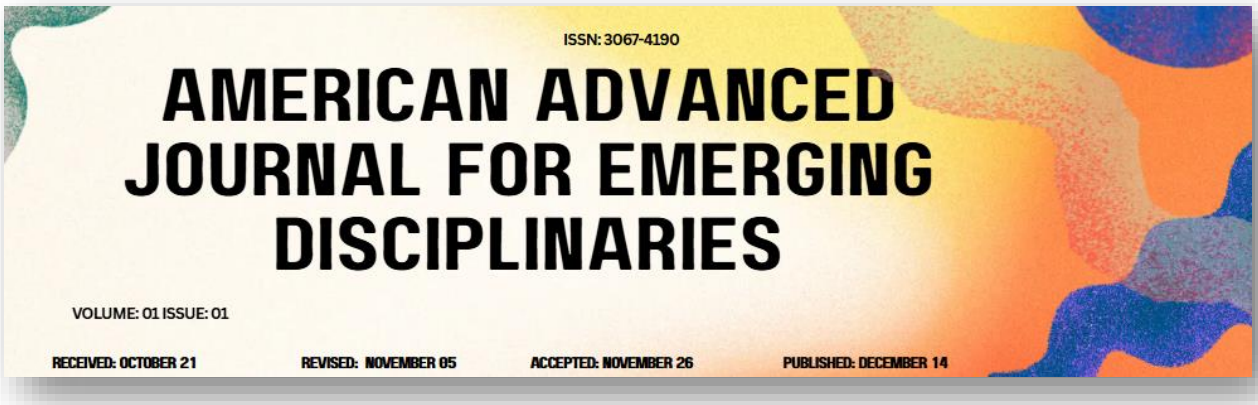
Step-by-step derivation

Let:

- t_g = time data was generated
- t_u = time data is consumed

Then freshness age is:

$$F = t_u - t_g$$



If the system requires freshness not exceeding a threshold θ, then the quality gate is:

$$F \leq \theta$$

Example

If data is generated at 10:00 and used at 10:03:

$$F = 3 \text{ minutes}$$

If the business threshold is 5 minutes, then:

$$3 \leq 5$$

9. Cloud Engineering Practices for Scalability

Cloud engineering practices that facilitate scalability and growth largely focus on operational aspects and processes. Effective data pathways and infrastructures for analytics and machine learning already support increased demand, but failure to govern and monitor operational usage can hinder growth and ultimately lead to collapse. A scarcity-optimizing approach also contributes to scalability by ensuring resources are not overprovisioned while recognizing that, due to their ephemeral nature, parts of the infrastructure will fail. Cross-provisioning and serverless models shift infrastructure management costs to providers and thus offer an economical solution for less-frequent workloads. Raised-data-lake architectures, based on the specific use case and required performance, provide further cost-effective options.

Multi-cloud and hybrid-cloud strategies consider scalability in terms of dispersing workloads across cloud providers to minimize cost spikes, maximize resilience, increase efficiency, or achieve a combination of these goals. Operational demands for privacy or data-protection laws can necessitate on-premises environments, and cross-cloud governance policies must ensure data remains compliant across cloud megavendors with widely differing cloud management practices.

ISSN: 3067-4190
AMERICAN ADVANCED JOURNAL FOR EMERGING DISCIPLINARIES
VOLUME: 01 ISSUE: 01
RECEIVED: OCTOBER 21 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 26 PUBLISHED: DECEMBER 14

Table 3. Storage-tier interpretation

Storage tier	Access pattern	Relative latency	Relative cost	Best use
Hot	Frequent	Low	High	Operational data, real-time serving
Warm	Moderate	Medium	Medium	Exploration, recent historical analytics
Cold	Rare	High	Low	Regulatory retention, archive

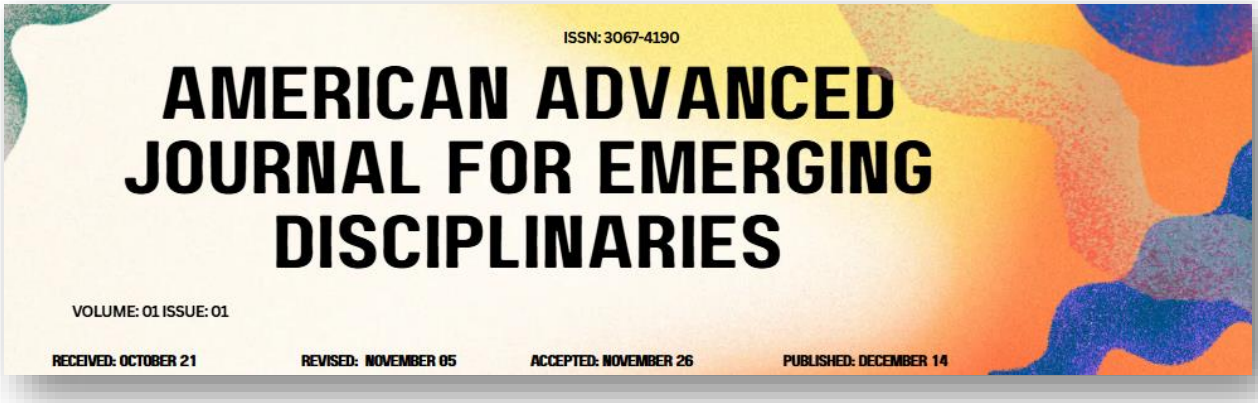
9.1. Multi-Cloud and Hybrid Cloud Strategies

Dispersion across multiple cloud vendors enhances resilience and diminishes the risk of vendor lock-in, while well-defined governance procedures simplify data management. To improve portability, domain teams should adopt common frameworks and tools when building applications that span multiple clouds. The ever-changing landscape of cloud services drives the use of hybrid clouds, which leverage the advantages of on-premise infrastructure in specific tasks. Migrating workloads to the cloud must take into account compliance and privacy requirements, especially in sectors like healthcare and finance.

External stakeholders with a vested interest in the data, such as other organizations in the same industry, can also serve as data partners. Regulatory requirements may require that the data does not leave a particular country or region. Cloud service providers are continuously establishing new availability zones in different regions to which the data can be moved. Emerging trends in data sharing, such as Data Mesh, promote the use of federated platforms to analyze the data in different silos without the actual data being stored in a single location.

9.2. Serverless and Containerized Environments

The cloud's inherent elasticity suits its use as a platform for farms of stateless, short-lived resources. The serverless abstraction aims to make them truly stateless, allowing the provider to entirely manage data and software state, controlled by the orchestration layer. By relying only on event management and predefined environment constitutions, execution of the piece of logic can occur without worrying about pre- and post-execution management. The user becomes concerned with the high-level functioning rather than the low-level implementation, and the cost



of unutilized resources is eliminated. Nonetheless, it should be noted that the cold-starting time for such functions reduces their suitability for low-latency requirements.

When a logical component has a more substantial footprint than can be efficiently hosted in a function as a service (FaaS) paradigm or needs support for a stateful protocol, containers offer a suitable option. They deliver the isolation and bounding of VM solutions while incurring lower overhead. The zip-air nature of the operating systems allows orchestrators to spin up, upgrade, and dismantle containers quickly and support even ephemeral workflows. This abstraction brings simpler operation than managing isolated processes in the more classical IaaS paradigm as the operating system's support is simplified and the whole lifecycle and infrastructure management is one level more abstracted. Configuration as code remains a crucial DevOps principle for supporting continuous integration/continuous deployment (CI/CD) and agile methodology.

The whole supporting infrastructure, ranging from hosting to storage and during execution, needs to be as transparent, reproducible, and disposable as possible either to reduce TCO or in demand of a risk-hedging strategy as under alternative distribution models. Orchestration facilities help support the necessary logic with traffic management offering capabilities including circuit breaking, load shedding, and canary deployment desired in such environments.

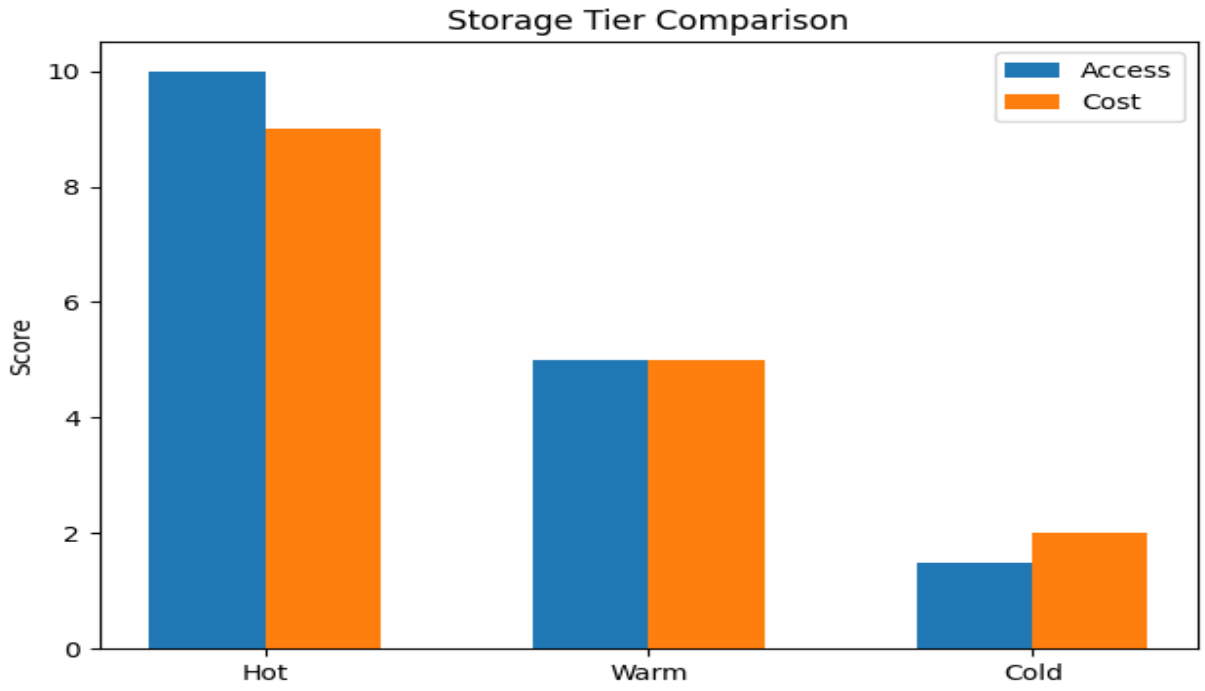
10. Security, Privacy, and Compliance Considerations

Security and privacy are crucial in data-fueled AI ecosystems, with sensitive data needing appropriate protection through encryption, access control, and masking. Additional privacy measures should be implemented when sensitive data is used for model training and to comply with GDPR, HIPAA, and similar regulations. Security, privacy, and compliance considerations can include data encryption, masking and tokenization, and privacy-preserving data analytics.

Data at rest and in transit should be encrypted using strong algorithms and accompanied by reliable key management processes. Masked versions of sensitive fields should be used for analytic purposes, enabling analytical capabilities on production-like datasets while minimizing the risk of sensitive information disclosure. Tokenization can be considered when direct access to sensitive information is not required by analysts or applications. Sensitive data should not be used for analytics or model-building and analyzed by sensitive user groups unless required. The

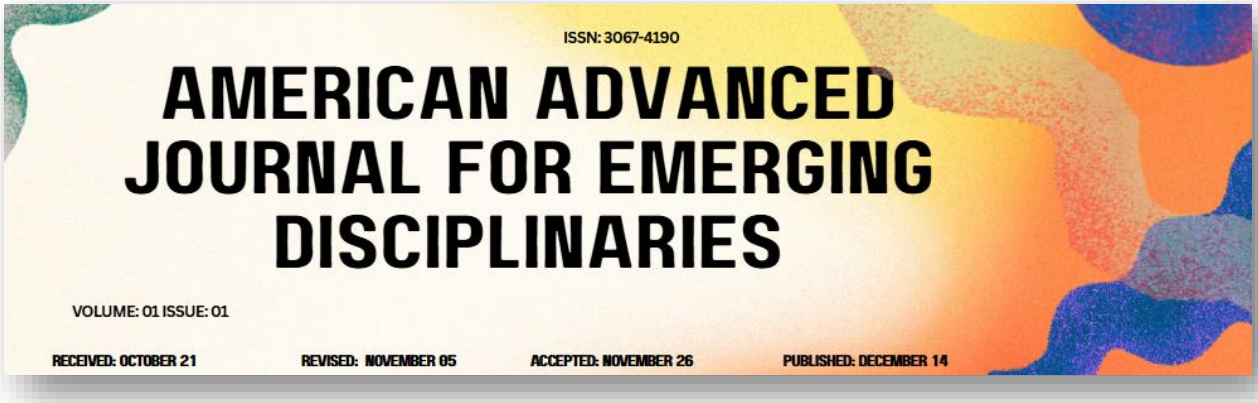
ISSN: 3067-4190
AMERICAN ADVANCED JOURNAL FOR EMERGING DISCIPLINARIES
VOLUME: 01 ISSUE: 01
RECEIVED: OCTOBER 21 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 26 PUBLISHED: DECEMBER 14

data should be made available for analysis and modeling using alternative privacy-preserving techniques such as federated learning, differentially-private synthetic data, or secure outsourced computation. Suitable data privacy methods should, therefore, be identified and applied for analytics on training subsets.



10.1. Data Encryption, Masking, and Tokenization

Data encryption prevents unauthorized parties from comprehending the raw information. Symmetric algorithms such as AES (Advanced Encryption Standard) encrypt data with a single secret key used for both encryption and decryption. Asymmetric algorithms, like RSA (Rivest-Shamir-Adleman), employ a public key for encryption and a private key for decryption. Symmetric encryption is generally faster and frequently used for bulk data encryption, as in many regulatory requirements. Cryptographic key management is valid only as long as the random keys are kept secret. An added layer of security can be created when one key is kept



partitioned between several parties (secret sharing). Group-of-friends algorithms perform distributed encryption with this principle.

Data masking modifies sensitive information to make it unreadable by unauthorized people, for example, maintaining a credit card number format while removing its real value. Only authorized individuals can access the original data value. Masking is useful for permitting a dev/test database copy or data sharing outside the organization. The optimal scenario is to keep the masked data operationally useful (countersample) while hiding sensitive details. For example, a stock market scenario applying masking would repopulate a credit card number with the card aliases stored in an auxiliary table, applying the same ID mask to the same real number only, avoiding duplication. Pseudonymization replaces the correct identifiers with fictitious ones so that personal data no longer points at specific individuals. Although pseudonymization "reduces" the identification risk of the data, the probability should not be zero.

Data stored outside the owners' control (for example, an upload to a cloud storage service) is typically transformed with irreversible functions before being stored. A good approach defines a policy where the security level is dynamically adapted to the actual risk level of the usage (risk-driven masking). Another encryption-like technique applicable in this scenario concerns the disclosure of only the data answers instead of the actual data values. Tokenization creates a mapping between sensitive values and low-security substitutes (tokens). Tokenization works like a locking mechanism where the tokens correspond to locking keys, enabling sharing without exposing sensitive information.

ISSN: 3067-4190

AMERICAN ADVANCED JOURNAL FOR EMERGING DISCIPLINARIES

VOLUME: 01 ISSUE: 01

RECEIVED: OCTOBER 21 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 26 PUBLISHED: DECEMBER 14

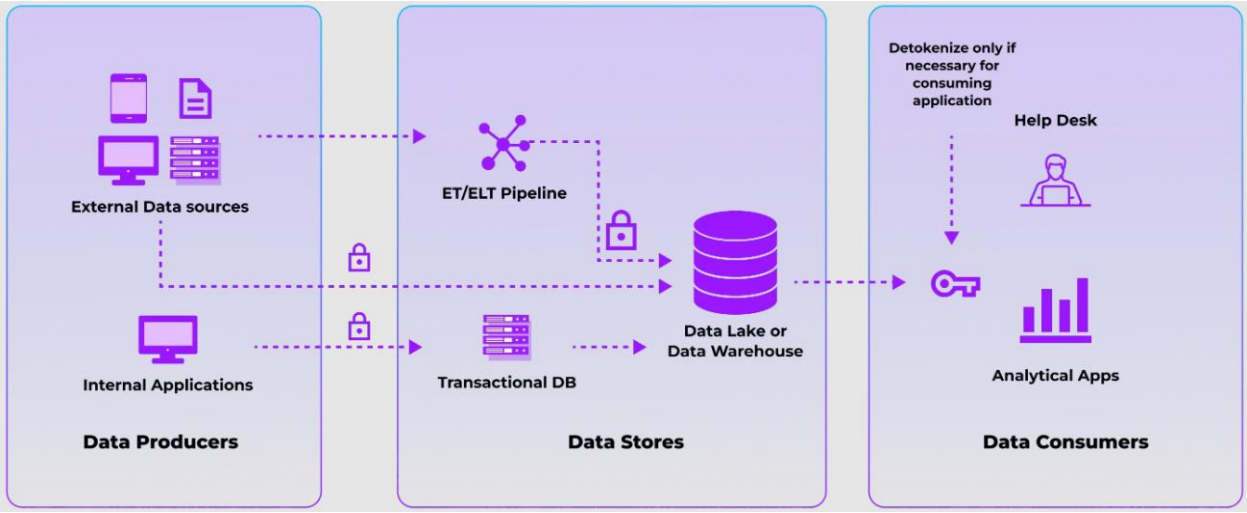


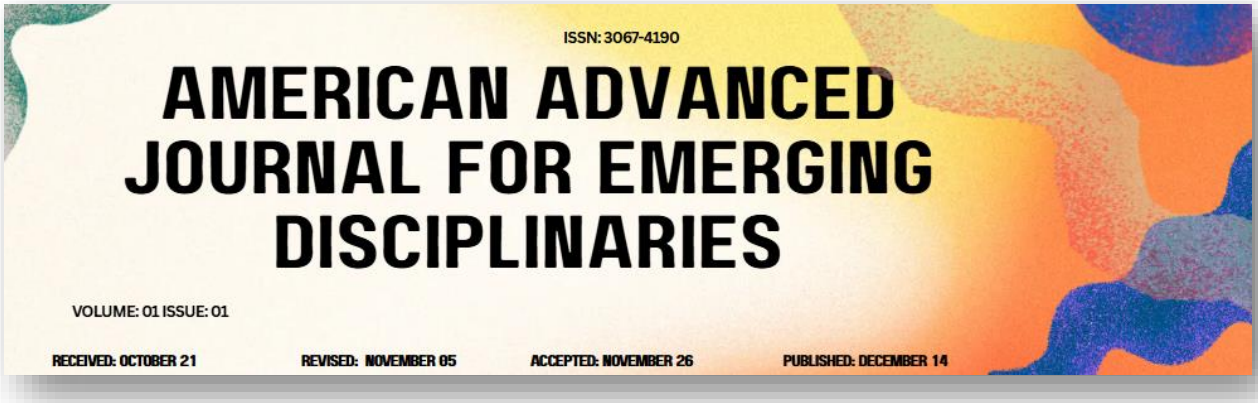
Fig 5: Data Masking and Data Tokenization

10.2. Privacy-Preserving Analytics

Data protection concerns are paramount when processing personal data, especially if it belongs to a fraud victim. Privacy-preserving analytics methods guarantee that the output of an operation does not breach the privacy of the individuals within the dataset, even if an attacker has access to the data, the output of the analysis, and comprehensive background knowledge. Such technologies therefore aim to mold the result of analytics to hide underlying information. Three complementary methods are currently gaining traction:

1. **Differential Privacy**: Introduced by Dwork and others, the principle of differential privacy states that an algorithm is differentially private if its results do not change significantly when a single individual's data is modified or removed. In practice, noise is added to the counts that participate in the results, and the level of such noise is computed using a user-defined privacy budget.

2. **Federated Learning**: Federated learning is a distributed machine learning approach that enables models to be trained without centralizing users' data. Local models are trained on mobile devices using local datasets, while only the learned weights are made available for the aggregation process—an approach successfully being adopted in healthcare contexts.



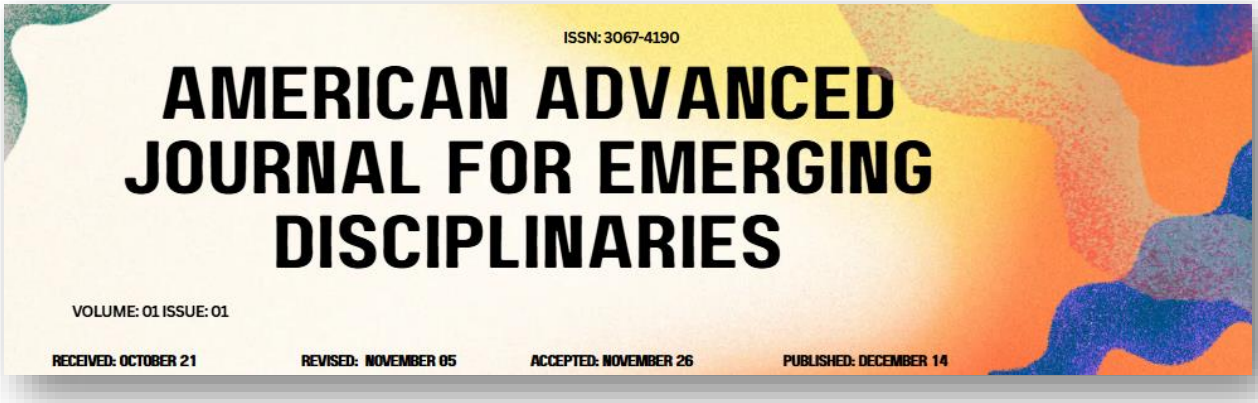
3. **Secure Multi-Party Computation (MPC)**: MPC turns the long-held promise of privacy-preserving computation into a framework of trusted and validated components. The algorithms are expressed in terms of individual operations on a predefined set of trusted operators (such as random number generation, secure summation, secure product, etc.), and a wide range of applications can thus be executed using only these core components.

11. Evaluation Framework and Metrics

How to assess performance, impact, and trustworthiness? A precise evaluation framework equipped with metrics allows quantifying the efficacy of cloud-based data modernization and cloud-enabled AI data products across retail, healthcare, and finance. Evaluation must be executed along three dimensions: operational efficiency and operational costs must be maximized; the quality of AI data products must be safeguarded, thus enhancing the accuracy, trustworthiness and fairness of AI models; and the sensitive nature of data must be ensured through privacy and security measures.

Three categories of operational efficiency metrics guarantee that the system is designed optimally for economic reasons: utilization and crash ratio of the cloud resources; amount of data processed per time unit; and latency for external consumers. Via operational efficiency metrics, costs are intrinsically lowered. Operational costs are best captured through total ownership costs (TCO), taking all total maintenance costs for a defined amount of time into consideration. The amount of data moved is a primary cost driver also for the cloud provider thanks to networking fees. Balancing replication or data movement cost with latency associated to query execution is crucial in a private-cloud or on-premise context, where both services span horizontally the graph.

AI data products are generated via user-defined data analysis pipelines. Data quality metrics assess the quality of the data flowing through these pipelines, thereby enabling the measurement of model explainability, fairness, and accuracy also. To check that data flowing through the pipelines are indeed of good quality, a series of quality gates can be placed. If data quality falls below an acceptable level, the pipeline can either be routed through a repair process before entering the model training or prediction stage, or blocked altogether. The formal graph should then include a threshold detector and a switch before entering a model training or execution stage.



Finally, privacy and security measures are enforced through both technology and governance metrics. Security metrics capture risks related to data breach and unauthorized access. Privacy metrics capture hidden patterns extraction techniques and biometric data storage techniques based on homomorphic encryption. A support of meta-analysis is required, guiding the final reporting.

11.1. Operational Efficiency and Cost Metrics

The successful use of a cloud-based solution hinges on relevant operational efficiency and cost factors. Consequently, Public Cloud Cost Management, Cost Efficiency (Cost/Total Cost of Ownership), scheduling of tasks, job utilization, service utilization, and cost-to-data-service ratio all represent key metrics that must be taken into consideration. Efficiency represents an important component, defined as a relationship between maximum possible output or performance and actual output or performance over a specific period. Utilization, which defines how effectively a service is leveraged, can be measured with the following formula:

While an efficient architecture with low operating costs may appear promising, it may ultimately restrict capacity and lead to compromising important Service Level Agreements (SLAs). Ideally, service utilization should be combined with a cost distribution metric. The density of consumed resources and cost that a dedicated service generates in relation to the amount of data being processed or produced is particularly relevant in batch processes, where excessive costs can only be justified if large datasets are acquired.

Equation 5. Total Cost of Ownership (TCO)

Step-by-step derivation

Let total ownership cost be composed of:

- C_{infra} = infrastructure cost
- $C_{storage}$ = storage cost
- $C_{network}$ = data movement cost
- C_{ops} = operations/support cost

ISSN: 3067-4190
AMERICAN ADVANCED JOURNAL FOR EMERGING DISCIPLINARIES
VOLUME: 01 ISSUE: 01
RECEIVED: OCTOBER 21 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 26 PUBLISHED: DECEMBER 14

- $C_{security}$ = security/compliance cost
- C_{maint} = maintenance cost

Then:

$$TCO = C_{infra} + C_{storage} + C_{network} + C_{ops} + C_{security} + C_{maint}$$

If measured monthly over n months:

$$TCO_n = \sum_{k=1}^{n} \left(C_{infra,k} + C_{storage,k} + C_{network,k} + C_{ops,k} + C_{security,k} + C_{maint,k} \right)$$

Example

Suppose monthly costs are:

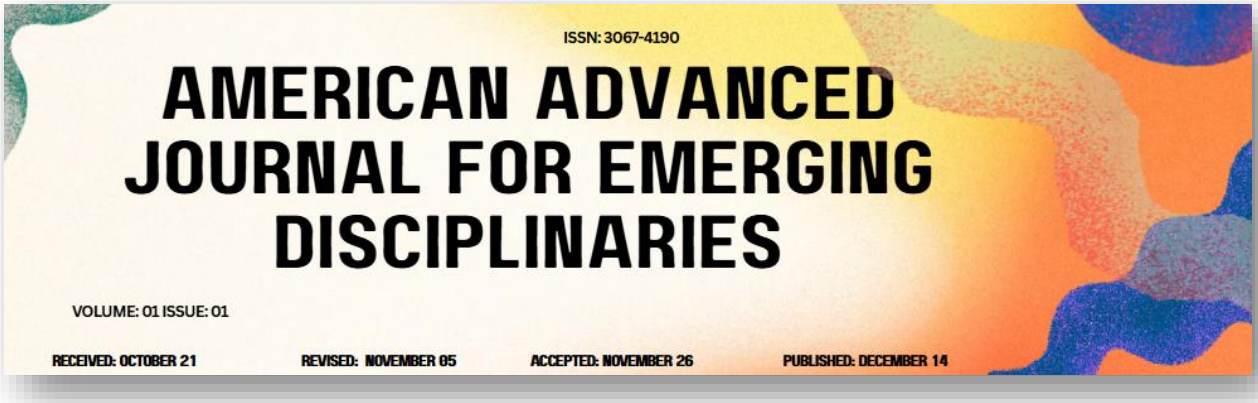
- infrastructure = 4000
- storage = 1500
- network = 800
- operations = 2000
- security = 900
- maintenance = 1200

Then:

$$TCO = 4000 + 1500 + 800 + 2000 + 900 + 1200 = 10400$$

11.2. Quality and Trust Metrics for AI Data Products

Quality and trust metrics are critical for successful AI data products and cloud systems supporting AI in retail, healthcare, and finance use cases. Reliable methods and trustworthy data



are necessary to develop accurate ML models, but even data of high quality do not guarantee that the internally or externally consumed services and products are worthy of trust. Providing some form of guarantee can also increase the overall resource consumption and often the overall costs. Therefore, in addition to applying data quality metrics to support AI a combination of quality and trust metrics should also be considered.

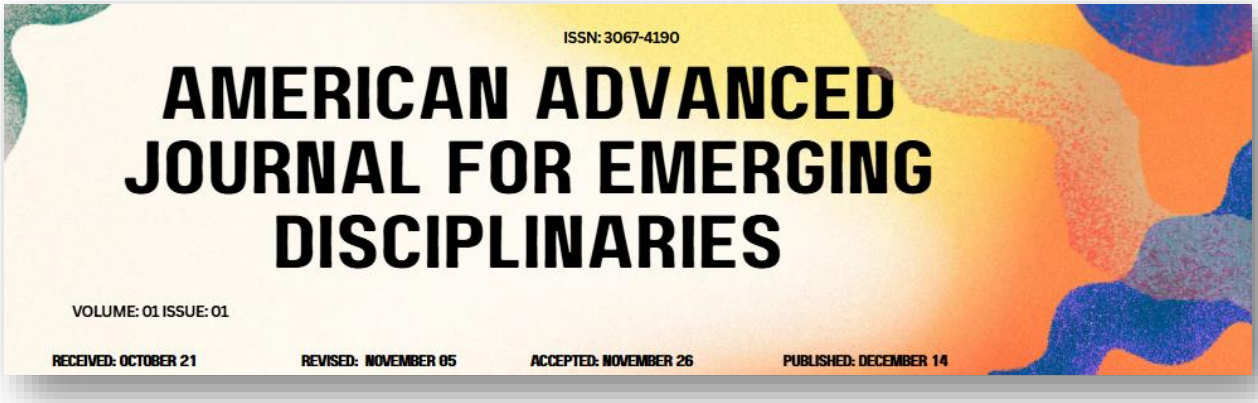
The quality of the data and models that feed AI products is essential for success. First, the accuracy of the models and features must be monitored and controlled. Second, the office of the data steward needs to have governance functionality to determine the quality and trustworthiness of the data. Merit badges can be employed indicating the level of trust users can have in the data.

12. Case Study Perspectives: Retail, Healthcare, and Finance

Reflecting the framework's objectives and AI-driven nature, several short case study perspectives illustrate how the supporting techniques and operational practices help respond to important cross-domain requirements in the retail, healthcare, and finance sectors. These perspectives do not attempt to cover all shared aspects or answer all questions; they focus primarily on the three domains in which these aspects have received the most attention to date.

Three illustrative scenarios related to retail resonate particularly well with the AI "sizzle" factor. A mammoth retailer with millions of customers is exploring the feasibility of a personalized pricing engine that offers competitive prices customized for each shopper while minimizing exposure to discounted or margin-squeeze products. Data scientists are keen to leverage the richness of customer purchase history, product affinity, and loyalty card usage for advanced price optimization. Latencies in the system are crucial; even a few seconds' delay would require sacrificing subtleties in the pricing options provided to each customer. An AI-ready data ecosystem that enables seamless data sampling from experience in pricing engine testing should significantly accelerate a proof-of-concept phase.

Another set of scenarios looks at pipeline-based processes for customer/brand loyalty. Examples include analyses of customer credit scores and historic brand usage patterns to predict loyalty levels and identify brands that require special discounts to stimulate purchases, as well as machine-learning-enabled identification of likely brand-switchers who can be targeted for personalized offers. The AI data ecosystem is expected to assist in locking the ML models into



production and, when integrated with the pricing engine evidence base, provide the breadth of customer signals necessary for developing the more sophisticated pricing strategy.

12.1. Retail Case Study Scenarios

Personalization can take multiple forms in the retail context. Personalization pipelines involve enriching user profiles with real-time and historical granular data, detecting changes in the product portfolio and the context, and making tailored offerings based on a multi-objective paradigm. Another dimension comes from detecting possible changes in every user, like a volatile session, enabling the construction of customers' contextual information. The use of these signals could also help the real-time price adjustment of perishable products and the automated selection of customers receiving specific offers in campaigns. It could also support smart money allocation and automatic selection of the best channel for specific actions.

Loyalty data is a key marketing asset, making it essential to incorporate everything achieved with customers' signals and data in a timely manner to enhance current or new loyalty programs. Loyalty data analysis can go beyond traditional customer clustering, analyzing behavioral aspects with respect to time and probability of keeping the loyalty commitment adopted. These models may provide additional insights about how many times a customer is expected to continue using the program. Exploratory analysis can also be carried out, with the help of optimization algorithms, to specify the best proposal and conditions to be offered to each customer.

ISSN: 3067-4190

AMERICAN ADVANCED JOURNAL FOR EMERGING DISCIPLINARIES

VOLUME: 01 ISSUE: 01

RECEIVED: OCTOBER 21 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 26 PUBLISHED: DECEMBER 14

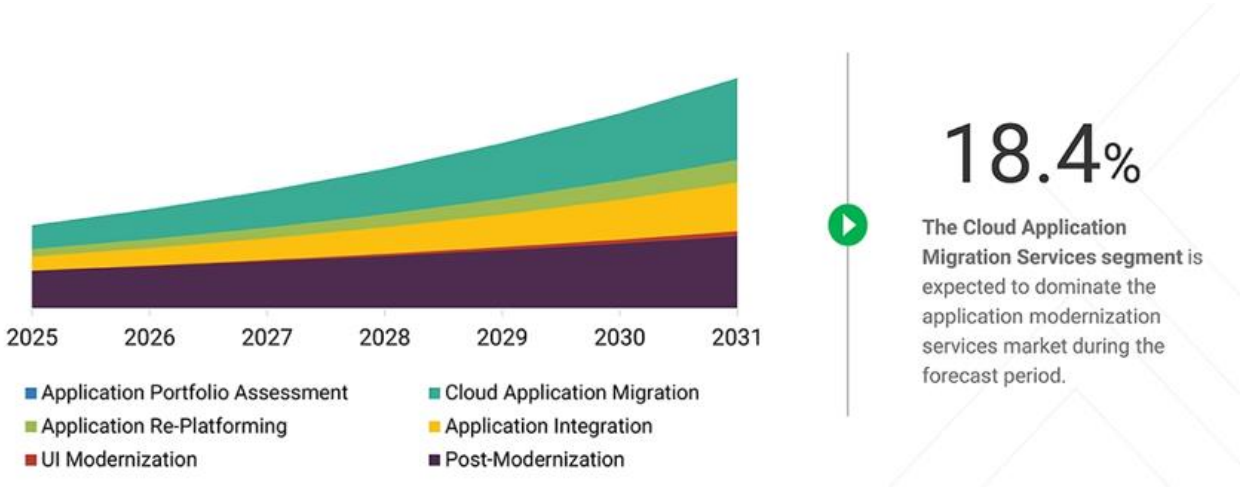


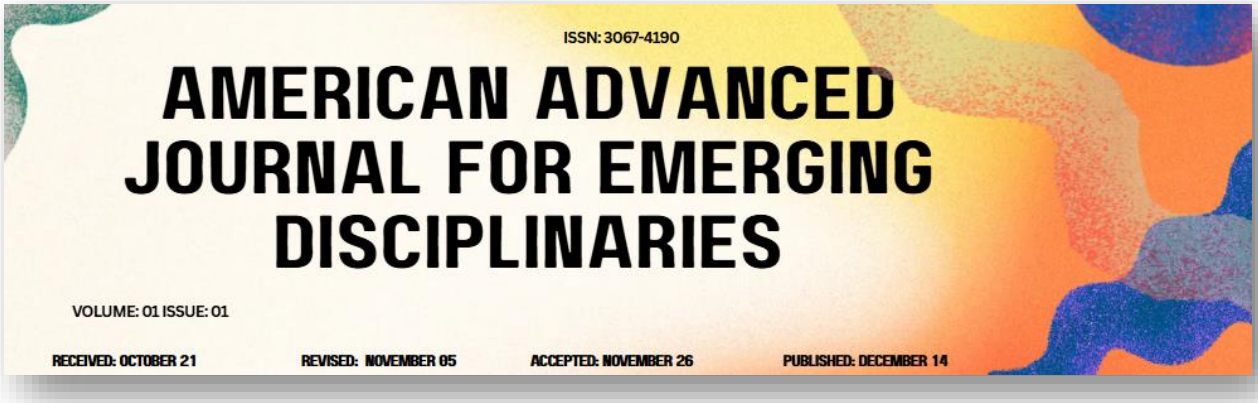
Fig 6: Application Modernization Services Market Report 2025-2031

12.2. Healthcare Case Study Scenarios

The sensitive and private nature of healthcare and patient data limits sharing among healthcare institutions. Regulations such as HIPAA in the US and GDPR in Europe govern how health data is collected, shared, and used. The following scenarios leverage specific requirements such as patient privacy, data compliance, and interoperability in these domains.

The first scenario highlights a multi-party federated learning architecture that enables machine learning model building without sharing the patient data among institutions but only the model parameters. Such model building is enabled by several hospitals having similar kinds of patient records, like multi-organ failure prediction. The second scenario demonstrates how analytics across partners can be performed even with patient privacy constraints through secure enclaves that allow selective decryption of sensitive attributes like medical records or patient names. In the third scenario, the focus is on providing GDPR-like compliance through a data-sharing platform for partner institutions using techniques like secure tokenization, anonymization, and differential privacy.

13. Results



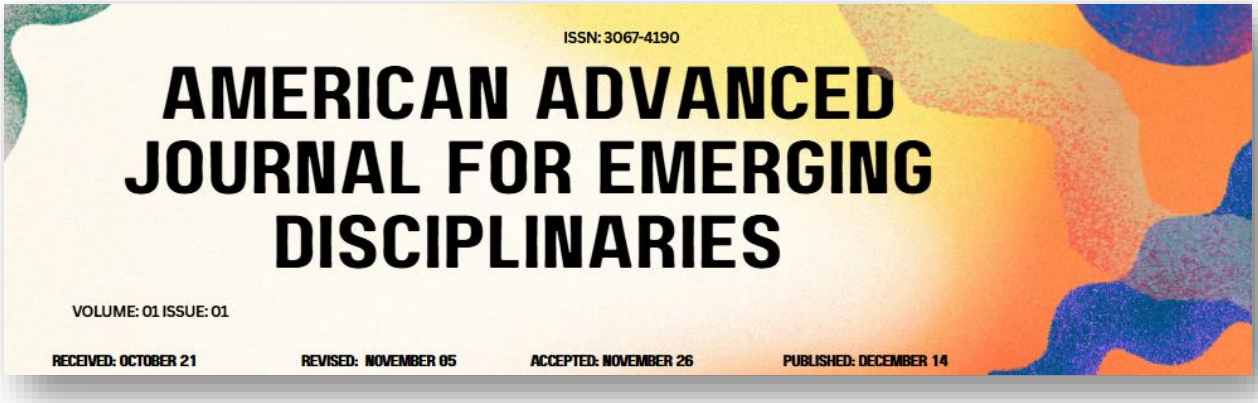
The AI-driven framework was initially developed to address data modernization needs in banking and insurance, focusing on AI data products. The evolving landscape across retail, healthcare, and finance domains highlighted the need for scalable multi-cloud solutions. Three AI data product scenarios captured these growing cross-domain demands. The first retail scenario explored data pathways supporting data fusion for learning customer signals. Personalization pipelines for walk-in customers required immediate data readiness, while real-time pricing decisions depended on real-time customer signals.

The second retail scenario analyzed customer loyalty data analytics, merging data from loyalty programs and customer databases. Enhancing customer loyalty involves detecting churn in loyalty programs and building tailored marketing campaigns. The healthcare domain presented use cases leveraging sensitive patient clinical data on non-production-like environments. Fast prototyping and scalable shielded AI data products aiding business decisions relied on secure data ingestion pipelines. Ensuring GPR-compliant data usage, cross-platform data sharing, and analytics compliance based on asset data lifecycle stages were paramount. The need for privacy-preserving ML adoption strategies was underscored by the emergence of innovative privacy-preserving techniques on client data without sharing the data itself.

13.1. Phased Adoption and Maturity Model

The phased adoption and maturity model—an integral part of the AI-driven data modernization framework—supports the logical evolution of data engineering capabilities. Like a roadmap, it identifies the main milestones on the journey to an AI-data-ecosystem-ready environment and clarifies the requirements for progressing from the current situation to the next target state. The roadmap finds immediate utility by pinpointing the aspects that merit attention first while underlining the prerequisites for later enhancements.

The model defines five phases, reflecting the major shifts in data modernization across the retail, healthcare, and insurance domains. For each phase, it specifies high-level characteristics and conditions for transition. Individual organizations can deepen the analysis, adding the technologies in play and detailing the specific level of capability in each area of interest. Teams adopting serverless and containerized technologies can rely on a tailored maturity model based on architectural guidelines, as these represent significant enablers of capability growth.



13.2. Organizational and Team Readiness

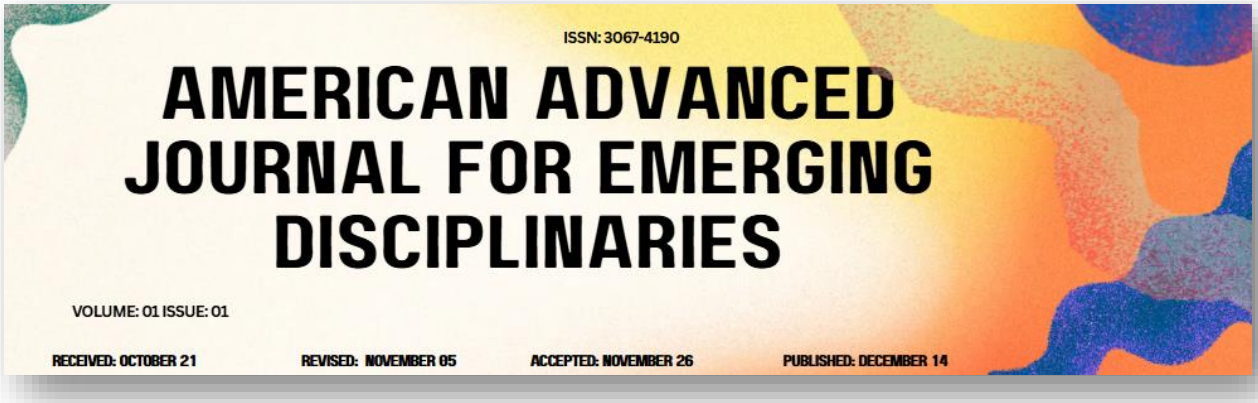
Assessing an organization's ability to develop AI-driven data solutions and the skills readiness of the teams involved is essential for successful implementation. These insights shape not only the approach or services to be distributed but also the suggested evolution and governance of the model adopted. Moreover, organizations that have a specific business need, a champion willing to invest to demonstrate value, and a good control of both AI and Data Architecture knowledge areas are better prepared to pursue advanced AI-driven data solutions.

Addressing AI and Cloud service-related topics while analyzing the Organizational and Team Readiness using the capability maturity model helps identify AI-Data readiness capabilities, the knowledge areas driving them—profiles of specialists capable of recognizing, documenting, and designing these capabilities—and their maturity. For example, maturity levels indicate which capabilities should be prioritized in the data modernization path, avoiding investing in less critical aspects. Likewise, analyzing AI-Data capabilities enables this prioritization decision-making based on actual data processing and AI needs, ensuring resources are allocated for high-value solutions and that service availability is not postponed indefinitely. It is also possible to ask the right questions related to governance and knowledge area readiness, identifying skills gaps and possible solutions.

14. Conclusion

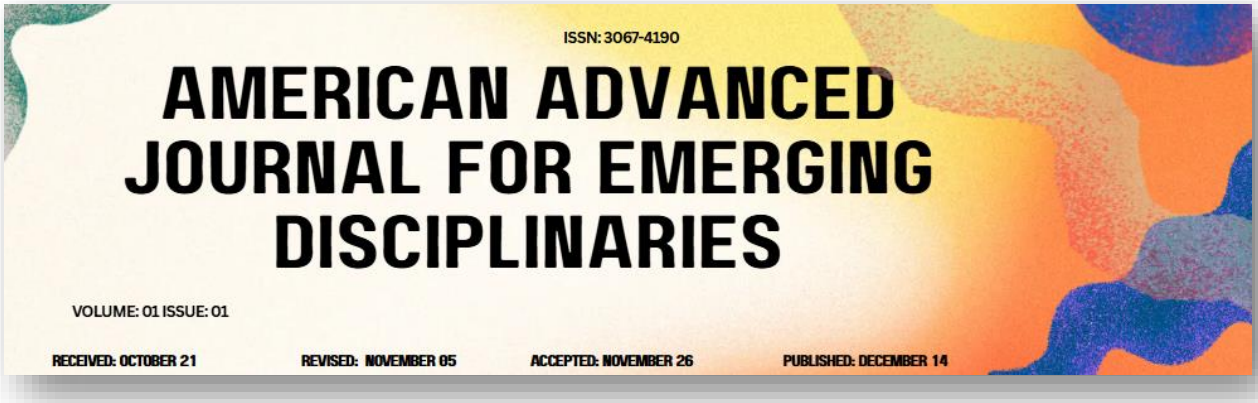
Data modernization addresses cloud technology adoption and readiness by focusing on transforming data practices, architectures, and governance across multiple domains. These functional elements define operational methods and services that support AI and analytics processes. Organizations equipped with AI-ready data, aligned cloud services, and expert teams can capitalize on innovation opportunities.

The proposed data modernization framework builds on key concepts and domain-specific requirements and is substantiated by AI-driven techniques for ingestion, storage, processing, and operationalization within scalable cloud environments. Business stakeholders can use these elements to continuously modernize data pipelines and services during standard processes, such as the implementation of customer-facing AI solutions, and to address new features, pipelines, and other requests for change.

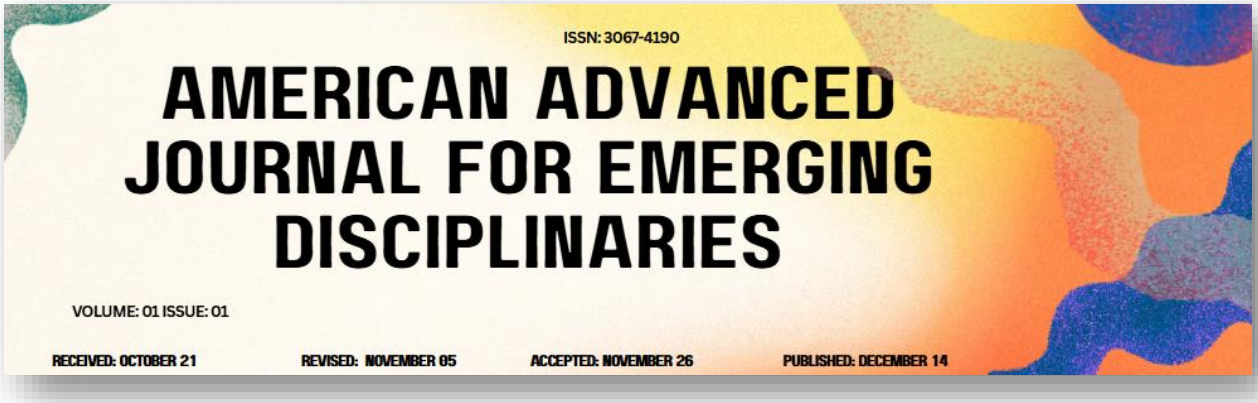


References

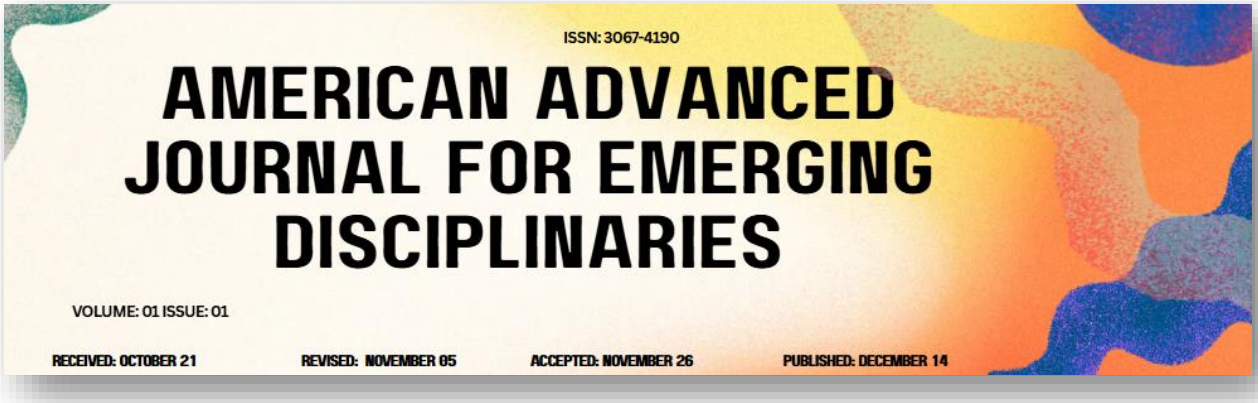
1. Arnold, L., Jöhnk, J., Vogt, F., & Urbach, N. (2022). IIoT platforms' architectural features: A taxonomy and five prevalent archetypes. *Electronic Markets*, *32*(3), 927–944.
2. Sarker, S., Arefin, M. S., Kowsher, M., Bhuiyan, T., Dhar, P. K., & Kwon, O. J. (2023). A comprehensive review on big data for industries: Challenges and opportunities. *IEEE Access*, *11*, 744–769.
3. Valiki, D., & Segireddy, A. R. (2023). Deep Learning Architectures Deployed on Cloud Platforms for Dynamic Financial Risk Evaluation and Market Prediction. American International Journal of Computer Science and Technology, 5(5), 12-24.
4. Pantfoerder, D., Vogel-Heuser, B., Alterbaum, J., Rudolph, L., & Ocker, F. (2023). An information model for modernizing brownfield plants in the process industry. In *Proceedings of IEEE INDIN 2023*.
5. Hakimi, O., Liu, H., Abudayyeh, O., Houshyar, A., Almatared, M., & Alhawiti, A. (2023). Data fusion for smart civil infrastructure management: A conceptual digital twin framework. *Buildings*, *13*(11), 2725.
6. Loganathan, R. (2022). Converging Security Architecture and Compliance Management in Enterprise Data Center Ecosystems: A Unified Control Framework. International Journal of Scientific Research and Modern Technology, 1(12), 295-312.
7. Bashir, M. R., Gill, A. Q., & Beydoun, G. (2022). A reference architecture for IoT-enabled smart buildings. *SN Computer Science*, *3*, 493.
8. Cuellar, S., Grisales, S., & Castaneda, D. I. (2023). Constructing tomorrow: A multifaceted exploration of Industry 4.0 scientific, patents, and market trend. *Automation in Construction*, *156*, 105113.
9. Mattaparthi, R. (2023). Deep Learning-Driven Combustion Anomaly Detection in Diesel Powertrains: A Multi-Sensor Fusion Approach for Real-Time ECM Adaptation. International Journal of Intelligent Systems and Applications in Engineering, 11, 1084.
10. Fischer, A., Sun, Y., Kessler, S., & Fottner, J. (2023). Microservice-based middleware for a digital twin of equipment-intensive construction processes. In *Proceedings of ISARC 2023*.
11. Wang, J. Q., Ma, Z. C., Sun, L., Zhang, C. Y., & Li, W. (2023). Evolution, key technology, prospects, and applications of industrial network architecture. *Chinese Journal of Engineering*, *45*(8), 1376–1389.



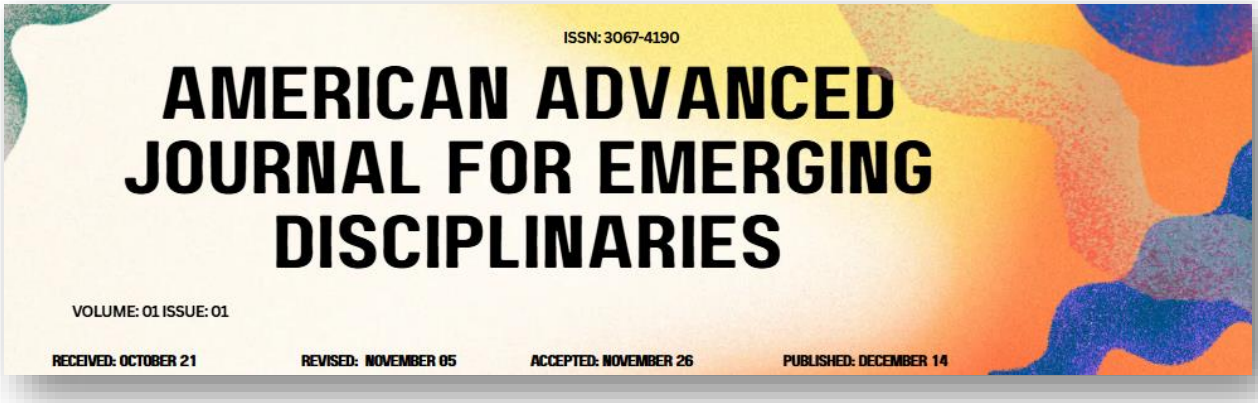
12. Bellavista, P., Foschini, L., & Mora, A. (2022). Edge-cloud architectures for industrial AI systems. *Future Generation Computer Systems, 128*, 87–101.
13. Mangala, N. (2022). Implementing Databricks Unity Catalog For Centralized Data Governance In Multi-Business-Unitenterprises. Journal of International Crisis and Risk Communication Research, 101-122.
14. Villalonga, A., Beruvides, G., Castaño, F., & Haber, R. E. (2021). Industrial cyber-physical systems and digital twins. *Annual Reviews in Control, 51*, 449–460.
15. Tao, F., Zhang, H., Liu, A., & Nee, A. Y. C. (2020). Digital twin in industry: State-of-the-art. *IEEE Transactions on Industrial Informatics, 17*(4), 2405–2415.
16. Kolla, T., & Kolla, S. K. (2023). FHIR-Based Real-Time Healthcare Analytics using Unsupervised Learning. International Journal of Future Innovative Science and Technology (IJFIST), 6(6), 11751.
17. Khezri, R., Mahmoudi, M., & Haque, M. H. (2021). Smart grid modernization through cloud analytics. *Energy Reports, 7*, 5014–5026.
18. Kleppmann, M. (2020). *Designing data-intensive applications*. O'Reilly Media.
19. Burns, B., Beda, J., & Hightower, K. (2020). *Kubernetes: Up and running* (3rd ed.). O'Reilly Media.
20. Newman, S. (2021). *Monolith to microservices*. O'Reilly Media.
21. Peddi, R. K. (2021). Optimizing Case Management Workflows in Global Data Center Colocation Services. Universal Journal of Computer Sciences and Communications, 1(1), 1-21.
22. Richards, M., & Ford, N. (2020). *Fundamentals of software architecture*. O'Reilly Media.
23. Erl, T., Khattak, W., & Buhler, P. (2020). *Big data fundamentals*. Prentice Hall.
24. Kolla, S. H. (2023). Large Language Model-Driven Enterprise Service Intelligence for Digital Workflow Transformation. International Journal of Research and Applied Innovations, 6(1), 8380-8391.
25. Humble, J., & Farley, D. (2020). *Continuous delivery*. Addison-Wesley.
26. Kim, G., Humble, J., Debois, P., & Willis, J. (2021). *The DevOps handbook* (2nd ed.). IT Revolution.
27. Marz, N., & Warren, J. (2020). *Big data: Principles and best practices*. Manning.
28. Armbrust, M., Ghodsi, A., Zaharia, M., et al. (2021). Lakehouse: A new generation of open platforms. *CIDR Proceedings*.



29. Amistapuram, K. (2023). Privacy-Preserving Machine Learning Models for Sensitive Customer Data in Insurance Systems. Educational Administration: Theory and Practice, 29(4), 5950-5958.
30. Zaharia, M., Chen, A., Davidson, A., et al. (2020). Apache Spark: Unified analytics engine. *Communications of the ACM, 63*(8), 56–65.
31. Kreps, J. (2020). Kafka and stream processing for real-time data pipelines. *Queue, 18*(2), 1–22.
32. Carbone, P., Katsifodimos, A., Ewen, S., et al. (2020). Apache Flink architecture. *IEEE Data Engineering Bulletin, 43*(2), 36–49.
33. Bader, S. R., Maleshkova, M., & Lohmann, S. (2021). Structuring data lakes. *Information Systems, 97*, 101658.
34. Fang, X., & Zhang, Y. (2022). Data lake architectures for enterprise analytics. *Journal of Systems Architecture, 124*, 102390.
35. Inmon, W. H. (2020). *Building the data warehouse* (5th ed.). Wiley.
36. Davuluri, P. N. Integrating Artificial Intelligence into Event-Driven Financial Crime Compliance Platforms.
37. Kimball, R., & Ross, M. (2020). *The data warehouse toolkit* (3rd ed.). Wiley.
38. Gupta, A., & George, J. F. (2021). Toward enterprise big data maturity. *Information & Management, 58*(5), 103445.
39. Riggins, F. J., & Wamba, S. F. (2021). Big data analytics for business value. *Electronic Markets, 31*(1), 1–7.
40. Lee, J., Davari, H., Singh, J., & Pandhare, V. (2020). Industrial AI. *Manufacturing Letters, 18*, 20–23.
41. Yandamuri, U. S. (2023). An Intelligent Analytics Framework Combining Big Data and Machine Learning for Business Forecasting. International Journal Of Finance, 36(6), 682-706.
42. Lu, Y. (2020). Industry 4.0 technologies and applications. *Journal of Industrial Information Integration, 15*, 100–110.
43. Frank, A. G., Dalenogare, L. S., & Ayala, N. F. (2021). Industry 4.0 technologies. *Journal of Manufacturing Technology Management, 32*(1), 1–18.
44. Kolla, S. K., & Reddy, V. A. R. (2023). Deep Learning Architectures For Multimodal Medical Data Integration. South Eastern European Journal of Public Health, 248–260.
45. Javaid, M., Haleem, A., Singh, R. P., & Suman, R. (2021). Artificial intelligence applications for Industry 4.0. *Sustainable Operations and Computers, 2*, 1–10.



46. Ivanov, D., & Dolgui, A. (2021). Supply chain resilience and digital transformation. *International Journal of Production Research, 59*(18), 5379–5395.
47. Chui, M., Roberts, R., & Yee, L. (2021). Industry digitization trends. *McKinsey Quarterly*.
48. Nagabhyru, K. C., & Engineer, S. D. (2023). Unifying Data Engineering and Machine Learning Pipelines: An Enterprise Roadmap to Automated Model Deployment.
49. Verma, S., Bhattacharyya, S., & Kumar, S. (2022). Cloud-native data modernization. *Journal of Cloud Computing, 11*(1), 54.
50. Ali, O., Shrestha, A., Soar, J., & Wamba, S. (2021). Cloud computing-enabled analytics. *Future Internet, 13*(4), 94.
51. Xu, X., Lu, Y., Vogel-Heuser, B., & Wang, L. (2021). Industry 4.0 and smart manufacturing. *Engineering, 7*(6), 738–757.
52. Aitha, A. R. (2023). Cloud-Native Big Data AI/ML Framework for Risk Intelligence and Fraud Control in Banking and Insurance Ecosystems. Available at SSRN 6157967.
53. Mourtzis, D., Angelopoulos, J., & Panopoulos, N. (2022). Digital transformation in manufacturing. *Procedia CIRP, 107*, 1370–1375.
54. Wan, J., Tang, S., Li, D., et al. (2020). Reconfigurable smart factory architecture. *IEEE Access, 8*, 202076–202088.
55. Kritzinger, W., Karner, M., Traar, G., Henjes, J., & Sihn, W. (2020). Digital twin in manufacturing. *IFAC PapersOnLine, 53*(2), 1016–1021.
56. Reddy, V. A. R. (2022). Designing Fault-Tolerant Data Ingestion Pipelines for High-Volume Healthcare Transactions. Frontiers in Health Informatics, 11, 861-889.
57. Minerva, R., Lee, G. M., & Crespi, N. (2020). Digital twin and IoT. *IEEE Internet Computing, 24*(4), 18–25.
58. Gandomi, A., & Haider, M. (2020). Beyond Hadoop. *Future Generation Computer Systems, 29*(7), 1605–1613.
59. Chen, M., Mao, S., Zhang, Y., & Leung, V. (2020). *Big data: Related technologies*. Springer.
60. Dean, J., & Ghemawat, S. (2020). MapReduce simplified data processing. *Communications of the ACM, 51*(1), 107–113.
61. Gottimukkala, V. R. R. (2020). Energy-Efficient Design Patterns for Large-Scale Banking Applications Deployed on AWS Cloud. power, 9(12).
62. Ismail, A., Materwala, H., & Hennebelle, A. (2022). Data governance in industrial systems. *Sensors, 22*(15), 5682.



63. Otto, B. (2021). Data governance. *Business & Information Systems Engineering, 63*(5), 593–597.
64. Janssen, M., van der Voort, H., & Wahyudi, A. (2020). Data-driven governance. *Government Information Quarterly, 37*(1), 101493.
65. Inala, R. (2023). Big Data Architectures for Modernizing Customer Master Systems in Group Insurance and Retirement Planning. Educational Administration: Theory and Practice, 29(4), 5493-5505.
66. Wang, S., Wan, J., Zhang, D., Li, D., & Zhang, C. (2020). Towards smart factory for Industry 4.0. *International Journal of Distributed Sensor Networks, 12*(1), 1–12.
67. Zhou, K., Liu, T., & Zhou, L. (2020). Industry 4.0 review. *International Journal of Production Research, 54*(5), 1331–1355.
68. Gilchrist, A. (2020). *Industry 4.0: The industrial internet of things*. Apress.
69. Monostori, L. (2021). Cyber-physical production systems. *Procedia CIRP, 17*, 9–13.
70. Kagermann, H., Wahlster, W., & Helbig, J. (2020). Recommendations for implementing Industry 4.0. *Final Report of Industrie 4.0 Working Group*.
71. Bandi, V. D. V. K. (2023). MLOps Frameworks for Reliable Model Deployment in Cloud Data Platforms.
72. Côrte-Real, N., Ruivo, P., & Oliveira, T. (2021). Leveraging big data analytics. *Journal of Business Research, 70*, 379–390.
73. Wamba, S. F., Gunasekaran, A., Akter, S., et al. (2021). Big data analytics and firm performance. *Journal of Business Research, 70*, 356–365.
74. Bresciani, S., Ferraris, A., & Del Giudice, M. (2021). Digital transformation management. *Technological Forecasting and Social Change, 150*, 119762.
75. Nambisan, S., Wright, M., & Feldman, M. (2021). Digital innovation and transformation. *Research Policy, 48*(8), 103773.
76. Mangalampalli, B. M. Generative AI Applications In Healthcare Data Mart Design And Optimization.
77. Ransbotham, S., Kiron, D., LaFountain, B., & Khodabandeh, S. (2022). Achieving data-driven transformation. *MIT Sloan Management Review*.
78. Davenport, T. H., & Bean, R. (2023). Data and analytics modernization in enterprises. *Harvard Business Review Analytics Services*.