World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

AI-driven analytics products at scale: A managerial and technical perspective

Yuvachandra Marasani *

Director, Software Development at IQVIA, USA.

World Journal of Advanced Research and Reviews, 2024, 21(02), 2193-2211

Publication history: Received on 1 January 2024; revised on 18 February 2024; accepted on 27 February 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.21.2.0450

Abstract

Organizations continue to struggle with scaling AI-powered analytics products from successful research prototypes into enterprise-grade, production systems. Prior research typically addresses this challenge through two partially disconnected streams: a technical stream focused on MLOps, data infrastructure, and model lifecycle automation; and a managerial stream focused on governance, organizational capability, and data-driven decision-making. This separation limits both theory and practice by under-explaining how technical and managerial systems interact during scaling.

This study proposes an integrated socio-technical framework—the Integrated Scaling Architecture for AI Analytics (ISA-4)—that conceptualizes scaling as a co-evolutionary process across four interlinked layers: Data Infrastructure, Model Operations, Governance, and Organizational Alignment. Building on the Resource-Based View, Dynamic Capabilities, and Socio-Technical Systems theory, the framework defines scaling performance as a multiplicative function of resource endowments, reconfiguration capability, and the quality of socio-technical coupling across layers.

The study advances eight propositions and operationalizes coupling mechanisms as a measurable construct across three dimensions: artifact, process, and normative coupling. The framework is applied to a structured multi-case investigation of pioneering AI-driven firms, illustrating how different coupling configurations explain variation in scaling speed, reliability, and organizational coordination.

This work contributes (1) a socio-technical co-evolution theory of AI analytics scaling, (2) a structured framework linking technical and managerial subsystems, and (3) a practitioner-oriented diagnostic lens to identify bottlenecks across the AI scaling lifecycle and inform platform strategy.

Keywords: Scaling Architecture; AI Analytics; MLOps; Model Lifecycle

1. Introduction

1.1. Background and Motivation

AI-infused analytics products—such as recommender systems, fraud detection, demand forecasting, and generative AI copilots—are increasingly central to enterprise operations. Yet, despite rapid advances in machine learning methods and scalable computing, many organizations fail to convert promising prototypes into robust, sustained production systems. Research and industry evidence consistently show a persistent gap between successful AI experimentation and enterprise-level scaling, with many initiatives never reaching durable operational impact (Davenport & Ronanki, 2018; Enholm et al., 2022; Shollo et al., 2022).

A key reason is that scaling AI-powered analytics is fundamentally different from scaling traditional software. AI systems are probabilistic, data-dependent, and dynamic; performance depends not only on model accuracy, but also on

* Corresponding author: Yuvachandra Marasani

data quality, infrastructure reliability, operational monitoring, governance constraints, and organizational readiness. Consequently, scaling is not merely an engineering challenge—it is a socio-technical challenge requiring coordinated evolution of technical platforms and managerial systems.

1.2. Dispersion of prior work

Existing research on AI scaling has developed largely in isolation across two dominant streams. The **technical stream** draws on software engineering and data systems literature and emphasizes machine learning operations (MLOps), pipelines, feature stores, reproducibility, deployment automation, and lifecycle monitoring (Sculley et al., 2015; Paleyes et al., 2022; Kreuzberger et al., 2023). This work highlights practical issues such as hidden technical debt, model drift, deployment fragility, and challenges in observability and maintenance at scale.

The managerial stream, grounded in information systems and strategy research, focuses on governance, AI capabilities, organizational alignment, and value realization (Benbya et al., 2020; Mikalef & Gupta, 2021; Shollo et al., 2022). It emphasizes challenges related to accountability, trust, team structures, decision rights, and capability building.

Both streams offer important insights—but because they are typically studied separately, the literature under-specifies the mechanisms through which technical architectures and managerial choices jointly shape scaling outcomes.

1.3. Research Gap and Problem Statement

The separation between technical and managerial perspectives reveals a critical research gap: the lack of integrated frameworks that explain how technical architecture choices and managerial systems **co-evolve** to enable (or constrain) scaling of AI-driven analytics products. This gap is especially visible in persistent tensions found across both streams. Technical studies often advocate infrastructure standardization (to reduce debt and improve reproducibility) while also emphasizing team autonomy (to preserve agility). Managerial studies similarly recognize the need for centralized governance while also promoting decentralized decision-making for responsiveness.

These contradictions are not incidental - they reflect a deeper socio-technical dilemma about coordination versus flexibility across an AI product's lifecycle. This motivates the central research question of the paper:

How do technical and managerial systems co-evolve to facilitate (or hinder) the scaling of AI-powered analytics products?

1.4. Conceptual Framework and Approach

To address this question, the paper proposes the Integrated Scaling Architecture for AI Analytics (ISA-4)—a socio-technical framework that models scaling through four interdependent layers:

- Data Infrastructure
- Model Operations
- Governance
- Organizational Alignment

The framework draws on three complementary theoretical perspectives: the Resource-Based View (Barney, 1991), Dynamic Capabilities (Teece, 2007), and Socio-Technical Systems theory (Bostrom & Heinen, 1977; Sarker et al., 2019). Together, these lenses support a view of scaling as a co-evolutionary process rather than a simple maturity progression in technical sophistication.

A central modeling choice in ISA-4 is the definition of scaling performance as multiplicative across dimensions—implying that deficiencies in any one layer or coupling relationship can constrain scaling, even when other components are strong.

1.5. Context and Empirical Support

To ground the conceptual framework, the study applies ISA-4 to a structured multi-case analysis of leading AI-driven firms, including Netflix, Amazon, and Google. These cases provide rich publicly available evidence on technical architecture patterns, deployment practices, and governance approaches. The intent is illustrative rather than confirmatory, enabling identification of recurring patterns and trade-offs that inform the framework and propositions.

1.6. Contributions

This research makes three contributions.

- First, it advances a socio-technical co-evolution perspective on AI analytics scaling, framing scaling as emergent from interactions between technical and managerial subsystems.
- Second, it introduces the ISA-4 framework, specifying inter-layer coupling mechanisms and offering testable propositions for empirical research.
- Third, it provides a diagnostic lens for practitioners to identify scaling bottlenecks and harmonize technical architecture, governance practices, and organizational design across the AI product lifecycle.

1.7. Paper's Structure

The remainder of the paper is structured as follows. Section 2 reviews relevant literature across four areas and highlights tensions motivating an integrative approach. Section 3 introduces ISA-4 and its propositions. Section 4 presents the research design and methodology. Section 5 provides case analyses. Section 6 discusses implications and boundary conditions. Section 7 outlines limitations and future research directions. Section 8 concludes.

Table 1 Framing the Integration Gap in AI Scaling Research

Perspective	Primary Focus	Strength	Limitation
Technical (MLOps, Data Systems)	Infrastructure, pipelines, deployment	Deep technical rigor	Limited organizational context
Managerial (IS, Strategy)	Governance, capabilities, alignment	Strong organizational insight	Technical abstraction
ISA-4 (This Study)	Socio-technical integration	Joint explanation of scaling	Requires multi-layer analysis

This table highlights that existing research explains parts of the problem, but not the system as a whole—motivating the need for an integrative framework.

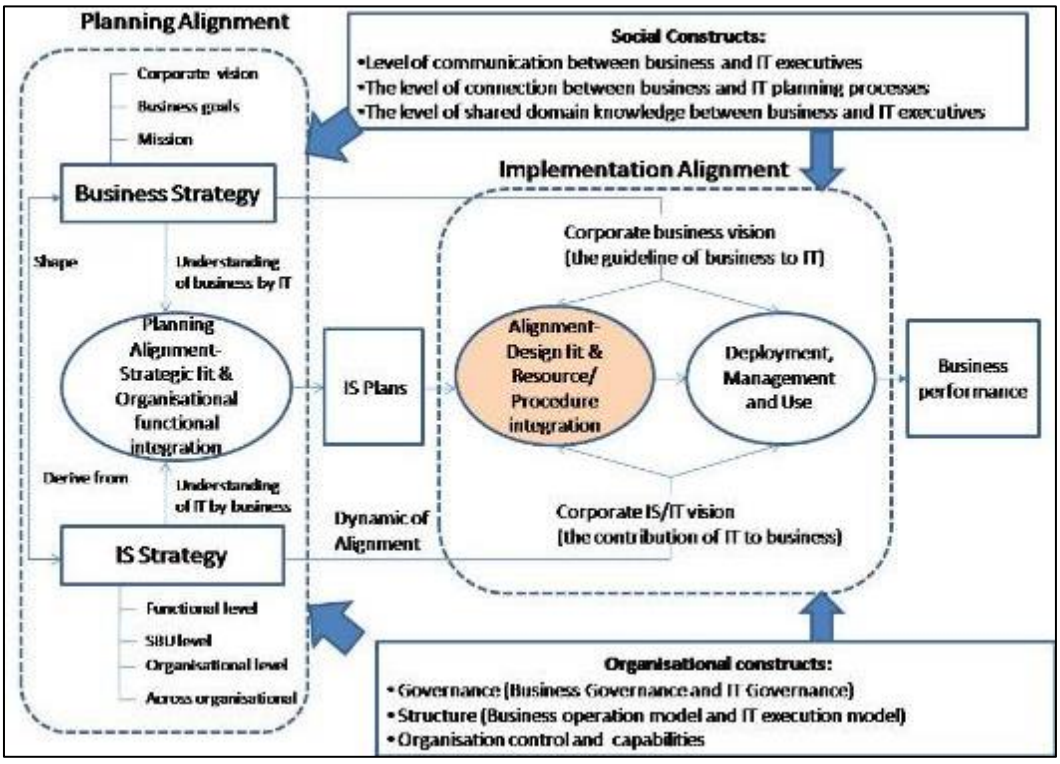

Figure 1 Conceptual Positioning of ISA-4 within AI Scaling Research

2. Literature Review

2.1. Review Methodology and Focus

This literature review synthesizes four adjacent research areas that collectively shape the ability to scale AI-based analytics products: (i) MLOps and the model lifecycle, (ii) AI product lifecycle management, (iii) data infrastructure scalability, and (iv) organizational alignment and AI governance. The purpose is not to provide an exhaustive catalog of studies, but to identify recurring tensions and inconsistencies within and across streams that motivate an integrated socio-technical framework.

A narrative review approach was adopted. First, candidate studies were identified through keyword searches (e.g., "MLOps," "AI governance," "data infrastructure," "AI capability") in major academic databases. Second, citation-based expansion was used to locate influential or foundational works. Third, the resulting studies were thematically coded to surface dominant constructs, contradictions, and scaling challenges. The review is therefore integration-driven: it emphasizes cross-stream relationships and unresolved tensions rather than completeness of coverage.

2.2. MLOps and Model Lifecycle

The MLOps literature extends DevOps principles into machine learning systems, emphasizing automation, reproducibility, continuous delivery, and operational monitoring. A common lifecycle is now widely described: data ingestion, feature engineering, training, validation, deployment, monitoring, and retraining—framed as continuous rather than one-time activities (Sculley et al., 2015; Paleyes et al., 2022; Kreuzberger et al., 2023).

A major contribution of this stream is its analysis of hidden technical debt in machine learning systems. Foundational work highlights feedback loops, hidden dependencies, configuration complexity, and mismatch between training and serving environments—issues that intensify as systems scale (Sculley et al., 2015; Amershi et al., 2019). Subsequent research emphasizes continuous integration and training, robust monitoring, and operational feedback loops to preserve reliability and performance under drift and changing data distributions (Lwakatare et al., 2020; Kreuzberger et al., 2023).

A second recurring theme is traceability and reproducibility: establishing lineage across data, features, code, and model artifacts is difficult in distributed environments, yet foundational for auditability and reliable rollback (Schelter et al., 2018; Polyzotis et al., 2018).

However, the technical stream contains a persistent inconsistency: it simultaneously calls for infrastructure standardization (e.g., shared pipelines and feature stores to reduce debt and improve consistency) and for team autonomy (to preserve agility and reduce bottlenecks). Standardization reduces variability and improves reliability, but can introduce coordination overhead and slow responsiveness; autonomy increases speed locally but can fragment system design globally. This tension is recognized but not resolved within purely technical accounts—suggesting that organizational structure and governance are inseparable from scaling outcomes.

2.3. Product Management of AI

Research on AI product lifecycle management emphasizes that AI-based products differ from conventional software products because they are probabilistic, data-dependent, and continuously learning. Consequently, product value depends on sustained data quality, monitoring, and iterative model improvement rather than a one-time release (Agrawal et al., 2019; Benbya et al., 2020).

A key managerial challenge highlighted in this stream is development uncertainty. AI development is often characterized by uncertain feasibility, uncertain performance ceilings, and unpredictable iteration cycles, which complicate roadmap planning and decision rights (Berente et al., 2021). This also reshapes the role of product management: product leaders must increasingly bridge technical constraints, data realities, and business objectives while navigating ongoing model evolution (Zhang et al., 2021).

A recurring conceptual limitation is the unit of analysis. AI capabilities are often deployed as reusable services, platforms, and shared components rather than discrete products, making boundaries ambiguous. This ambiguity complicates lifecycle governance and accountability: when models, features, and pipelines are shared across domains, "ownership" becomes distributed and scaling becomes a platform coordination problem rather than a single-product execution problem. This boundary ambiguity is critical to enterprise practice and directly motivates the need for frameworks that explicitly model cross-layer coupling and organizational alignment.

2.4. Data Infrastructure Scalability

Data infrastructure scalability is a core technical foundation for AI analytics products. Research points to modern architectural patterns such as lakehouse platforms (unifying warehousing and lake flexibility) and streaming architectures enabling near-real-time processing—both critical for operational analytics products that require freshness, low latency, and cost control at scale (Armbrust et al., 2021; Akidau et al., 2015).

The emergence of feature stores addresses a practical scaling need: shared, governed feature definitions that improve consistency between training and serving and reduce duplicated feature engineering across teams (Schelter et al., 2018). In parallel, data mesh proposes decentralized domain ownership, treating data as a product and distributing accountability to domain teams (Dehghani, 2022; Machado et al., 2022).

Yet this stream also contains a central paradox: centralization (data lakes, centralized feature stores) improves consistency, governance, and economies of scale, while decentralization (data mesh) improves autonomy, responsiveness, and ownership. Both are empirically viable, but neither is universally optimal. The appropriate choice depends on organizational structure, governance maturity, and coordination needs—indicating that infrastructure decisions are socio-technical rather than purely technical.

2.5. Alignment and AI Governance

The managerial stream on AI emphasizes governance and alignment as prerequisites for sustained value creation. A prominent construct is AI capability, understood as an integrated bundle of technological, human, and organizational resources that enable effective AI use and performance outcomes (Mikalef & Gupta, 2021).

Governance is increasingly framed around trust, fairness, transparency, and accountability, supported by practices such as model audits, bias management, and compliance processes (Rai et al., 2019; Kellogg et al., 2020). Ethical and regulatory pressures further reinforce the need for formal governance mechanisms and documented operating principles (Floridi et al., 2018; Jobin et al., 2019).

This stream also emphasizes organizational design and team structure. Enterprises use varied configurations—centralized data science teams, distributed domain teams, or hybrid models—each introducing different trade-offs in coordination cost, speed, and control (Grønsund & Aanestad, 2020; Davenport & Mittal, 2022).

However, a notable limitation is that managerial literature frequently treats technical systems as a "black box." Governance is often conceptualized as rules applied to technology, without specifying how governance becomes embedded into pipelines, artifacts, and operational workflows. This weak integration with technical architecture limits explanatory power for scaling outcomes in production settings where governance must be implemented as enforceable mechanisms, not merely policy statements.

2.6. Theoretical Foundations

Three theoretical streams provide foundations for integrating the above literature into a coherent scaling explanation.

First, the **Resource-Based View (RBV)** explains how firm-specific resources—such as proprietary data, infrastructure assets, and specialized talent—can become sources of competitive advantage when they are valuable, rare, and difficult to imitate (Barney, 1991; Wade & Hulland, 2004).

Second, **Dynamic Capabilities** theory emphasizes the ability to sense opportunities, seize them, and reconfigure resources over time in response to shifting environments—an essential lens for AI systems that must adapt under drift, changing business objectives, and evolving governance requirements (Teece, 2007; Warner & Wäger, 2019).

Third, **Socio-Technical Systems (STS)** theory highlights the interdependence of social and technical subsystems, arguing that system performance depends on joint optimization rather than isolated improvements. This perspective is especially relevant for AI scaling because technical architecture and organizational structure shape each other and can either reinforce performance or amplify coordination failure (Bostrom & Heinen, 1977; Sarker et al., 2019).

While these theories are often used separately, their combined use supports a more holistic view: scaling is not a linear maturity path but a co-evolutionary process in which technical systems and organizational arrangements continuously shape each other's constraints and performance.

2.7. Synthesis and Research Gap

Taken together, the literature streams identify essential components of the scaling puzzle but do not explain how they fit into a unified causal structure. The unresolved tensions—standardization versus autonomy, centralization versus decentralization, control versus agility—appear across streams and persist because they reflect a deeper coordination-flexibility dilemma rather than isolated design flaws.

Table 2 Cross-Stream Synthesis of AI Scaling Literature

Stream	Dominant Focus	Key Contribution	Core Limitation
MLOps	Automation, deployment, monitoring	Technical scalability mechanisms	Ignores governance and organization
AI Product Lifecycle	Product uncertainty, continuous learning	Managerial insights	Ambiguous system boundaries
Data Infrastructure	Architecture, storage, pipelines	Scalable data foundations	Centralization vs decentralization tension
Governance & Alignment	Capabilities, ethics, structure	Organizational perspective	Weak integration with technical systems

This synthesis yields three integrative insights. First, contradictions within each stream remain unresolved because they arise from systemic coordination-flexibility trade-offs rather than single-variable problems. Second, these contradictions recur across both technical and managerial domains, suggesting they share a common socio-technical root cause. Third, the existing literature does not provide an overarching model that explicitly represents the relationships between technical layers (data infrastructure, model operations) and managerial layers (governance, organizational alignment) as interdependent systems.

Accordingly, a key research gap remains: the absence of a theoretical model explaining how technical and managerial systems co-evolve and how their coupling shapes scaling outcomes for AI-driven analytics products. This gap motivates the Integrated Scaling Architecture for AI Analytics (ISA-4) introduced in the next section.

3. Theorising Scaling: Integrated Scaling Architecture for AI Analytics (ISA-4)

3.1. Design and Overview

Building directly on the cross-stream synthesis in Section 2, this paper proposes the **Integrated Scaling Architecture for AI Analytics (ISA-4)**—a socio-technical framework for understanding how AI-driven analytics products scale through the **co-evolution** of technical and managerial systems. ISA-4 treats scaling not as a linear maturity progression or a purely technical optimization problem, but as an emergent outcome of how well interdependent layers are configured and coupled over time.

The central premise is that organizations do not "scale AI" by upgrading a single component (e.g., infrastructure or MLOps) in isolation. Instead, scaling performance depends on the **quality of coupling** between technical layers (data and model operations) and managerial layers (governance and organizational alignment). When coupling is weak or misaligned, scaling failures appear as familiar symptoms—deployment bottlenecks, unreliable models, inconsistent metrics, governance friction, and high coordination costs—even if individual components look strong in isolation.

ISA-4 is structured into four layers:

- Data Infrastructure
- Model Operations
- Governance
- Organizational Alignment

Each layer contains resources, processes, and decision structures. However, ISA-4 emphasizes that these layers are not independent: **scaling is a function of their configuration and coupling dynamics**—and the ability to reconfigure those relationships as environments change.

3.2. The Four ISA-4 Layers

3.2.1. Data Infrastructure

The **Data Infrastructure** layer provides the substrate for AI systems, including ingestion, storage, pipelines, feature stores, and data products. It determines the availability, quality, and consistency of data used to train, validate, and serve models. From a Resource-Based View perspective, proprietary data assets, reusable feature engineering, and lineage/observability systems can form durable sources of advantage—but their value depends on their integration with model operations rather than their existence alone.

3.2.2. Model Operations

The **Model Operations** layer captures the operational lifecycle of machine learning systems: experimentation, training, validation, deployment, monitoring, and retraining. This layer turns data into predictions (or decision support) at production quality. Dynamic Capabilities are most visible here because scaling requires continuous updating of models and operational pipelines in response to drift, changing business needs, and reliability constraints. In practice, the efficiency and reliability of this layer strongly influence scaling velocity.

3.2.3. Governance

The **Governance** layer includes policies, standards, and controls governing AI development and use—such as validation requirements, fairness and bias checks, compliance rules, risk management procedures, and incident response processes. A core principle in ISA-4 is that governance is not an external overlay; it becomes part of system design when implemented through enforceable mechanisms (e.g., policy-as-code embedded into pipelines). This shifts governance from periodic oversight to a scalable, operational capability integrated with delivery workflows.

3.2.4. Organizational Alignment

The **Organizational Alignment** layer includes team structures, incentives, accountability models, and decision rights that shape how AI systems are built and operated. It defines ownership of artifacts (data products, features, models), responsibility boundaries, escalation paths, and the way prioritization decisions are made. From a Socio-Technical Systems perspective, misfit between organizational design and technical architecture manifests as coordination failures, redundant work, and accumulated technical debt—even when individual teams perform well locally.

Table 3 ISA-4 Layer Decomposition

Layer	Core Components	Key Metrics	Theoretical Lens
Data Infrastructure	Pipelines, storage, feature stores	Data quality, lineage coverage	RBV
Model Operations	Training, deployment, monitoring	Deployment speed, reliability	Dynamic Capabilities
Governance	Policies, compliance, risk controls	Incident rate, audit coverage	STS
Organizational Alignment	Teams, incentives, decision rights	Coordination cost, alignment index	STS

The table illustrates that each layer brings unique capabilities and that scaling is not about the individual layers but how they're integrated.

3.3. Inter-layer Couplings

ISA-4 introduces coupling mechanisms as a core construct. Coupling refers to the extent to which decisions, artifacts, or changes in one layer shape and constrain another layer. By specifying coupling explicitly, ISA-4 moves beyond describing “factors” and instead models scaling as relationships that can be compared across organizations and tested empirically.

ISA-4 defines three coupling types (retained):

- Artifact coupling: shared technical artifacts across layers (e.g., feature stores, model cards, standardized metadata).
- Process coupling: interdependence of processes across layers (e.g., governance checks embedded in CI/CD or deployment pipelines).
- Normative coupling: coupling through incentives, KPIs, decision rights, and accountability structures.
- These coupling mechanisms are what transform ISA-4 from a taxonomy into a system model: they provide concrete linkages through which alignment, governance, and technical architecture jointly influence scaling outcomes.

3.4. Theoretical Integration

- ISA-4 integrates three theoretical perspectives to explain why scaling outcomes differ across organizations and over time.
- Resource-Based View (RBV): each ISA-4 layer contains valuable resources (data, infrastructure, talent, governance assets).
- Dynamic Capabilities (DC): scaling requires reconfiguring these resources continuously in response to drift and environmental change.
- Socio-Technical Systems (STS): outcomes depend on fit between technical subsystems and social subsystems; isolated optimization produces fragile systems.
- ISA-4 expresses scaling performance as multiplicative rather than additive:

Scaling Performance = f(RBV x DC x STS)

The multiplicative framing emphasizes that weaknesses in any one dimension constrain scaling even when the other dimensions are strong. Strong infrastructure cannot compensate for poor organizational alignment; strong governance cannot compensate for immature data systems; and high talent cannot compensate for fragile operational processes.

3.5. Propositions

- ISA-4 yields eight propositions describing how coupling configurations shape scaling outcomes. These propositions retain your original intent and sequence, while being phrased in a more publish-ready form.
- P1. The relationship between Data Infrastructure–Model Operations coupling and scaling speed follows an inverted U-shape: moderate coupling improves deployment speed, while excessive coupling reduces speed due to coordination and rigidity.
- P2. Embedding governance into model deployment pipelines improves deployment reliability; however, beyond a point, increasing governance requirements slows deployment.
- P3. Organizational structure and system modularity complement each other to reduce coordination costs; however, modularity–structure configurations may also increase technical debt when coupling is mismanaged or overly fragmented.
- P4. Firms with greater resource reconfiguration ability sustain scaling performance more effectively under dynamic environments than firms with rigid systems.
- P5. Governance mechanisms are positively moderated by data maturity: governance improves reliability primarily when advanced data systems and observability are present.
- P6. Greater congruence between incentives and responsible AI KPIs is negatively associated with system failure rates.
- P7. The optimal form and strength of ISA-4 coupling depends on the phase of the AI scaling lifecycle.
- P8. Outsourcing AI infrastructure components positively influences ecosystem value creation, contingent on partner absorptive capacity.

3.6. Dynamic Scaling Lifecycle

ISA-4 incorporates a dynamic lifecycle view with four phases:

- Experimentation
- Standardization
- Platformization
- Ecosystem Scaling

Each phase tends to favor different coupling configurations. In early phases, looser coupling can support rapid experimentation and learning, while later phases typically require tighter coupling to achieve consistency, reuse, and governance at scale. ISA-4 therefore frames scaling as a **temporal re-arrangement** of coupling relationships rather than a purely linear accumulation of maturity.

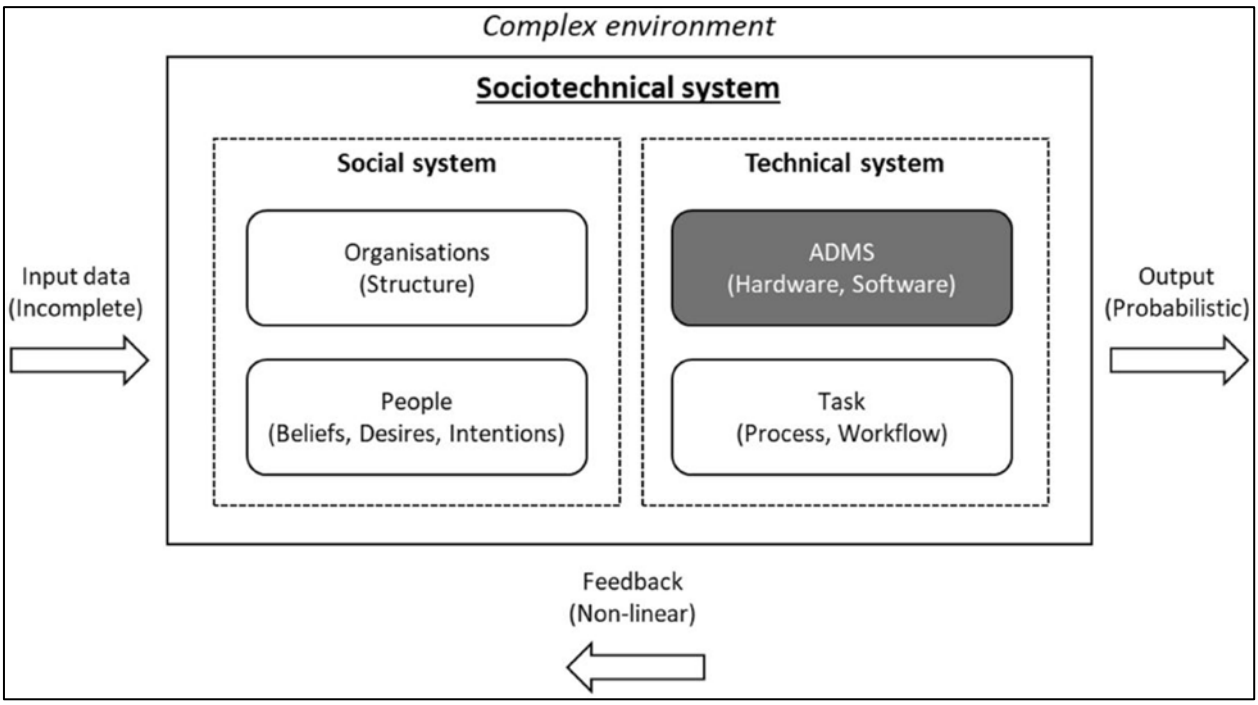

Figure 2 ISA-4 Framework and Scaling Lifecycle

The diagram shows scaling is not progressive but a temporal re-arrangement of links between layers.

3.7. Summary

This section introduced ISA-4 as an integrated socio-technical framework that models AI analytics scaling through interdependent layers, explicit coupling mechanisms, and lifecycle dynamics. The framework addresses the central limitation identified in the literature: existing work explains scaling components in isolation but under-specifies the relationships and co-evolution across technical and managerial systems. By operationalizing coupling (artifact, process, normative) and articulating testable propositions, ISA-4 provides a foundation for empirical evaluation and offers a practical diagnostic lens for identifying scaling bottlenecks across the AI product lifecycle.

The next section (Methodology) outlines the research design used to illustrate and evaluate ISA-4 through structured case analyses.

4. Methodology

4.1. Research Design

This study adopts a theory-building multiple-case research design to develop and demonstrate the Integrated Scaling Architecture for AI Analytics (ISA-4). Given the conceptual nature of the inquiry—and the limited availability of proprietary, end-to-end evidence on production AI systems in enterprises—a qualitative case-study approach is well suited to elaborate constructs, refine relationships, and strengthen theoretical coherence (Eisenhardt, 1989; Yin, 2018).

The research aims for analytical rather than statistical generalization, using empirical cases to extend and refine theory by comparing predicted patterns with observed evidence. The approach is abductive, iterating between theory and evidence to improve construct clarity and strengthen the fit between conceptual propositions and real-world scaling practices (Dubois & Gadde, 2002).

4.2. Case Selection Logic

Cases were selected using theoretical sampling to maximize variation across dimensions relevant to AI scaling. The selection logic sought organizations that (a) demonstrate scalable implementation of AI-driven analytics products, (b) provide sufficiently rich public evidence on both technical and managerial dimensions, and (c) exhibit diverse platform architectures, governance approaches, and organizational configurations.

Using these criteria, the study focuses on three cases:

- Netflix
- Amazon
- Google

These cases represent complementary scaling environments. Netflix illustrates tightly integrated, high-velocity AI systems with strong data-to-model operational coupling. Amazon represents multi-domain complexity where modular service architectures and organizational design are central. Google provides a view of ecosystem-scale platforms and governance integration in AI production environments.

Table 4 Case Selection Rationale

Case	Domain Focus	Key Scaling Characteristics	Relevance to ISA-4
Netflix	Content recommendation	High deployment velocity, centralized pipelines	Data–Model Ops coupling
Amazon	Retail, logistics, cloud	Multi-domain complexity, modular teams	Organizational alignment
Google	AI platforms, search, cloud	Platform ecosystem, governance integration	Governance coupling

4.3. Data Sources

The study relies on **multi-source publicly available evidence** to support transparency and reproducibility. Sources were selected explicitly to capture both technical architecture and managerial/governance dimensions, consistent with ISA-4's socio-technical focus.

Primary sources include:

- Technical blogs and technical documentation (e.g., architecture, design, deployment practices)
- Conference talks and open-source artifacts where applicable
- Company statements (e.g., AI principles, governance documents, strategy reports)
- Organizational indicators such as staff announcements, team structures, management hires, and structural changes

Secondary sources include industry reports and credible media where they provide contextual reinforcement. Evidence was systematically recorded in an **evidence log** capturing the source, publication date, and relevance to ISA-4 layers and coupling mechanisms.

4.4. Data Analysis

Data analysis followed an **abductive, pattern-matching strategy** designed to connect theory and evidence rigorously.

First, qualitative coding was conducted to identify evidence of:

- **Layer structures** (data infrastructure, model operations, governance, organizational alignment)
- **Coupling mechanisms** (artifact, process, normative)
- **Reconfiguration events** (e.g., major platform changes, governance redesigns, organizational restructuring)

Second, pattern matching was applied to compare coded evidence against theoretical expectations implied by the propositions (e.g., changes in reliability following governance integration into deployment processes).

Third, cross-case comparison was performed to identify recurring configurations, trade-offs, and contingencies. This comparative analysis was used to refine propositions and identify boundary conditions that shape where ISA-4 is most applicable.

4.5. Analytical Strategy

The analysis proceeded through three structured steps to ensure interpretive discipline and consistency across cases:

- **Within-case analysis:** Each organization was analyzed independently to identify patterns of ISA-4 layer configuration, dominant coupling mechanisms, and scaling dynamics.
- **Cross-case comparison:** Patterns were compared across cases to surface similarities, differences, and context-dependent relationships.
- **Theoretical refinement:** Insights from the cases were used to refine the ISA-4 framework, strengthen coupling constructs, clarify proposition interpretation, and articulate boundary conditions.

This staged approach supports close alignment between conceptual development and empirical evidence, consistent with the abductive intent of the research design.

4.6. Quality Assurance

To strengthen credibility and rigor, the study applies four commonly used quality criteria in case research:

- **Construct validity:** Multiple indicators were developed for core constructs, and evidence was triangulated across different sources.
- **Internal validity:** Alternative explanations were explicitly considered and checked against available evidence during pattern matching.
- **External validity:** Findings are generalized analytically through propositions and boundary conditions rather than statistically.
- **Reliability:** The research process—including source collection, coding, and analytical procedures—is documented to support transparency and replication.

4.7. Method's Limitations

Using publicly accessible sources introduces several limitations. First, internal decision processes and informal coordination practices are only partially visible, which may constrain interpretation of tacit activities. Second, publicly disclosed materials may be selectively reported, introducing representation bias. Third, the illustrative nature of the cases supports theory building but does not constitute confirmatory empirical testing of the propositions.

These limitations are mitigated through triangulation, systematic evidence logging, and analytic procedures consistent with established qualitative case research practices.

4.8. Summary

This methodology provides a structured and transparent approach for connecting conceptual development with empirical illustrations of AI scaling. The use of theoretical sampling, multi-source public evidence, and abductive reasoning supports the development of ISA-4 as a socio-technical scaling framework and provides a disciplined basis for refining constructs and propositions.

The next section applies this approach through case analyses of Netflix, Amazon, and Google to demonstrate how ISA-4 layer configurations and coupling mechanisms manifest in real scaling environments.

5. Case Analyses

5.1. Analytical Framing

This section demonstrates the ISA-4 framework in practice through illustrative analyses of three large-scale AI-driven enterprises. The objective is not to present exhaustive case studies, but to show how **socio-technical coupling across the four ISA-4 layers** manifests in different scaling environments and produces distinct scaling outcomes.

Each case is analyzed along three common dimensions:

- **Configuration of ISA-4 layers** (data infrastructure, model operations, governance, organizational alignment)
- **Dominant coupling mechanisms** (artifact, process, normative)
- **Observed scaling dynamics**, including growth, reorganization, and trade-offs

This common framing allows cross-case comparison while respecting the contextual differences among enterprises.

5.2. Case 1: Netflix

5.2.1. Layer Configuration

Netflix operates tightly integrated, data-intensive AI systems that power recommendation and personalization at massive scale. The data infrastructure is optimized around continuous data ingestion, large-scale feature engineering, and real-time access to user interaction data. Model operations are closely integrated with product delivery, enabling frequent experimentation and rapid deployment.

Governance mechanisms are embedded directly into experimentation and monitoring workflows, rather than operating as external checkpoints. Organizational alignment is characterized by high system ownership, with teams responsible for end-to-end performance of data pipelines, models, and user-facing outcomes.

5.2.2. Coupling Mechanisms

Netflix exhibits **strong Data–Model Operations coupling**, supported by shared feature stores and standardized data pipelines that ensure consistency between model training and serving. This artifact coupling reduces friction across the lifecycle and enables fast iteration.

Process coupling is visible in automated experimentation platforms, where evaluation, deployment, and rollback are tightly linked. **Normative coupling** is reinforced through incentives and performance metrics that align individual team success with system-level outcomes.

5.2.3. Scaling Dynamics

This configuration enables **high scaling velocity**, particularly in environments requiring continuous personalization. However, tight integration also requires ongoing reconfiguration to preserve flexibility as user behavior, content diversity, and product scope evolve. The case illustrates the **integration–adaptability trade-off**, lending support to propositions related to inverted-U effects of coupling on scaling speed.

5.3. Case 2: Amazon

5.3.1. Layer Configuration

Amazon's AI capabilities span a diverse set of domains, including retail, logistics, and cloud services. The data infrastructure is highly distributed and domain-oriented, allowing teams to tailor data products to local needs. Model operations are modular, with services deployed across heterogeneous environments.

Governance is formalized, especially in high-impact domains such as pricing and logistics, while organizational alignment emphasizes distributed teams operating with a shared architectural and cultural foundation.

5.3.2. Coupling Mechanisms

Amazon demonstrates **strong Organizational–Architecture coupling**, where team modularity closely mirrors service and system modularity—often described as an inverse application of Conway's Law. This alignment reduces local coordination friction while enabling parallel innovation.

Process coupling is achieved through standardized deployment and review workflows, while **normative coupling** is reinforced through common leadership principles, operational metrics, and internal accountability structures.

5.3.3. Scaling Dynamics

Amazon scales through **modular expansion rather than tight integration**, allowing independent innovation across domains at the cost of increased coordination overhead. This configuration highlights trade-offs between flexibility and coordination cost and aligns with propositions concerning organizational alignment, modularity, and coordination dynamics under scale.

5.4. Case 3: Google

5.4.1. Layer Configuration

Google operates AI platforms at ecosystem scale, supporting search, advertising, and cloud-based AI services. Its data infrastructure emphasizes consistency, availability, and reuse across products. Model operations differentiate between research-oriented experimentation and production-grade deployment environments.

Governance is extensive and formalized, with explicit policies addressing fairness, transparency, compliance, and risk. Organizationally, Google combines centralized research capabilities with decentralized product teams.

5.4.2. Coupling Mechanisms

Google exhibits particularly strong **Governance–Model Operations coupling**, with policy mechanisms integrated directly into deployment workflows. Validation, monitoring, and fairness testing are implemented as enforceable process steps rather than manual oversight.

Artifact coupling is visible through structured documentation such as model cards, while **process coupling** is supported by automated validation and review systems embedded in production pipelines.

5.4.3. Scaling Dynamics

Google's scaling reflects a continuous negotiation between innovation and control. Governance mechanisms enhance reliability and trustworthiness but introduce overhead that must be dynamically managed. This case supports propositions regarding governance integration, reliability, and risk control as scaling constraints rather than purely enabling factors.

Table 5 Cross-Case Comparison of ISA-4 Dynamics

Dimension	**Netflix**	**Amazon**	**Google**
Dominant Coupling	Data–Model Ops	Org–Architecture	Governance–Model Ops
Strength	Speed	Flexibility	Reliability
Risk	Over-integration	Coordination cost	Governance overhead
Scaling Mode	Integrated	Modular	Controlled

This comparison demonstrates that there is no single optimal scaling configuration. Each firm emphasizes different coupling relationships to optimize particular performance dimensions, confirming the need for a contingent approach to AI scaling.

5.5. Cross-Case Synthesis

Three observations emerge across cases:

- Scaling outcomes depend on which layers are most tightly coupled. Organizations prioritize different coupling configurations, leading to distinct scaling paths.
- Trade-offs are unavoidable. Centralization enhances efficiency and control but limits responsiveness; decentralization increases flexibility but raises coordination costs.
- Scaling requires continuous reorganization. Static architectures or governance models become liabilities as systems and environments evolve.

5.6. Summary

The case analyses reinforce the central claim of ISA-4: scaling AI analytics products is not driven by isolated technical or organizational elements, but by the configuration and evolution of socio-technical coupling across layers. Differences across Netflix, Amazon, and Google demonstrate that scaling is a contingent, dynamic phenomenon shaped by strategic priorities, organizational structures, and governance choices.

The following section examines the theoretical and managerial implications of these findings, along with boundary conditions for applying ISA-4.

6. Discussion

6.1. Theoretical Implications

This study advances a socio-technical co-evolution view of AI analytics scaling by showing that scaling is not primarily a linear progression of technical maturity, but a dynamic process shaped by the evolving fit and coupling between technical and managerial subsystems.

First, ISA-4 bridges research streams that have traditionally been treated as separate—MLOps, data platforms, governance, and organizational alignment—by modeling them as interdependent layers whose interactions shape scaling outcomes. This shifts the unit of analysis away from isolated elements (e.g., "a feature store" or "a governance policy") toward relationships and coupling dynamics, which better matches how scaling challenges arise in enterprise practice.

Second, ISA-4 contributes by specifying coupling mechanisms as the explanatory construct. Rather than treating the relationship between technical systems and organizational context as background conditions, the framework operationalizes coupling through artifact, process, and normative dimensions. This improves theoretical precision and opens a path for measurable constructs that can be tested empirically.

Third, the study proposes a multiplicative theorization of scaling performance by defining outcomes as a product of resource endowments, reconfiguration capability, and socio-technical fit. The implication is that scaling performance is constrained by the weakest dimension—strong infrastructure cannot fully compensate for weak organizational alignment, and strong governance cannot compensate for immature data systems. This directly addresses why many organizations fail to scale despite excellence in one domain.

Finally, ISA-4's dynamic lifecycle adds a process perspective to variance-based models by making time explicit: different phases of scaling require different coupling configurations. This reinforces that scaling is not a one-time design choice but a continual re-balancing of control, flexibility, and coordination across the lifecycle.

6.2. Context and Applicability

ISA-4 is best suited to environments characterized by enterprise-scale operations, advanced data maturity, and volatile markets, where AI systems are embedded into core workflows and scaling requires coordinated decisions across infrastructure, operations, governance, and organization. In such contexts, technical and organizational complexity interact, and scaling challenges emerge from coupling dynamics rather than isolated component failures.

At the same time, ISA-4 should be adapted for certain environments. For small and medium-sized enterprises, the full four-layer decomposition may be unnecessarily complex; scaling issues may be driven more by resource constraints than by coordination and coupling across layers.

In low data-maturity settings, governance may act as a constraining force rather than an enabler. Without strong data quality, lineage, and observability, attempts to formalize governance can increase friction and reduce agility, because policies cannot be operationalized reliably into technical workflows.

In highly regulated sectors, governance may become the dominant layer shaping decisions in other layers (e.g., driving architectural choices, deployment gates, and monitoring requirements). Under such conditions, influence between layers becomes strongly bidirectional and may be more tightly constrained by external compliance demands than by internal performance optimization goals.

6.3. Managerial Implications

ISA-4 offers several actionable implications for managers scaling AI-powered analytics products in production settings.

First, scaling must be treated as a system-level problem. Organizations should avoid "local optimization" (e.g., investing heavily in MLOps tools while ignoring governance or team design). In practice, scaling bottlenecks frequently originate in misalignment between layers—such as incompatible governance processes and deployment workflows, or

team incentives that conflict with reliability requirements. ISA-4 supports viewing scaling as an integrated portfolio of technical and managerial investments.

Second, sequencing matters. Capability development often has dependencies: for example, robust governance automation typically depends on data maturity (quality controls, lineage, observability) and stable operational pipelines. Likewise, platform investments may fail if decision rights, ownership models, and accountability structures are not aligned with the architecture being built. ISA-4 reframes "what to build next" as a sequencing question across layers.

Third, ISA-4 supports diagnosis. Scaling failures that appear technical (e.g., slow deployments) often have organizational causes (e.g., unclear accountability, excessive coordination) or governance causes (e.g., manual reviews and inconsistent approval criteria). ISA-4's coupling concept enables leaders to pinpoint whether the bottleneck is rooted in artifact fragmentation, process coupling misfit, or misaligned incentives—rather than defaulting to tool adoption as a solution.

Finally, re-alignment is continuous. Because models drift, environments change, and governance requirements evolve, organizations must repeatedly reconfigure architectures, processes, and team structures. Scaling strategies that assume stable operating conditions are unlikely to remain effective; sustained success requires deliberate investment in reconfiguration capability across layers.

Table 6 Key Implications of ISA-4

Dimension	Insight	Implication
Alignment	Scaling depends on inter-layer fit	Coordinate technical and managerial systems
Sequencing	Order of capability development matters	Build data maturity before governance
Diagnosis	Bottlenecks emerge from coupling gaps	Use framework for system-level analysis
Adaptation	Scaling requires continuous change	Invest in reconfiguration capability

The table provides practical recommendations based on theoretical insights, showing that ISA-4 is both explanatory and prescriptive.

6.4. Explaining Trade-offs in AI Scaling

A key insight of ISA-4 is that scaling AI analytics products inevitably involves trade-offs, particularly between coordination and flexibility, and between control and agility. These trade-offs are not incidental; they are structural outcomes of how coupling is designed and managed across layers.

Increasing coupling often improves reliability and consistency by reducing fragmentation and eliminating mismatches between layers (e.g., aligning training/serving features, embedding governance in pipelines). However, excessive coupling can create rigidity, increase coordination overhead, and slow response to change. Conversely, loose coupling increases adaptability and local autonomy but can create divergence, duplicated effort, and difficulties in maintaining shared standards.

ISA-4 reframes these trade-offs as design decisions that should be managed rather than eliminated. Optimal coupling varies by lifecycle phase, strategic context, and regulatory environment—making contingency and dynamic adjustment central to sustainable scaling.

6.5. On the Complexity of AI Scaling

The complexity represented in ISA-4 is intended to reflect the real complexity of AI analytics scaling rather than introduce unnecessary abstraction. AI scaling involves mutually dependent technical and organizational subsystems, each evolving at different speeds and under different constraints. Reduced models that focus only on isolated technical maturity or only on managerial governance may fail to explain why scaling stalls in practice.

By emphasizing coupling and co-evolution, ISA-4 prioritizes explanatory power and system realism. This aligns with socio-technical systems theory, which argues that performance depends on joint optimization and that local improvements can inadvertently worsen system-level outcomes when interdependencies are ignored.

6.6. Summary

This discussion shows how ISA-4 offers both theoretical and practical value by providing an integrated, dynamic account of AI analytics scaling. The framework strengthens prior research by making interdependencies explicit, specifying coupling mechanisms as analyzable constructs, and clarifying how trade-offs and boundary conditions shape scaling outcomes.

Most importantly, ISA-4 reframes scaling as a system design and operating model challenge—where success depends on continuously aligning data infrastructure, model operations, governance mechanisms, and organizational structures over time.

7. Limitations and Future Research

7.1. Limitations

As with all theory-building research grounded in illustrative cases, this study has several limitations that delineate the scope of its explanatory claims.

First, the analysis relies primarily on publicly accessible data sources, including technical blogs, documentation, public talks, and organizational disclosures. While these sources provide valuable insight into large-scale AI systems, they limit visibility into informal coordination practices, internal trade-offs, and undocumented decision processes. This constraint may lead to partial representations of how coupling mechanisms operate in practice.

Second, the case analyses are illustrative rather than confirmatory. Although Netflix, Amazon, and Google were selected to maximize theoretical variation, the findings do not constitute statistical generalization. Accordingly, the propositions generated by the ISA-4 framework require subsequent empirical testing through quantitative or mixed-method designs.

Third, ISA-4 is intentionally designed for enterprise-scale organizations with mature AI and data capabilities. The framework is therefore less suitable for small firms, low-data-volume contexts, or environments where AI adoption is exploratory rather than operational. As outlined in the boundary conditions, ISA-4 would require simplification or reconfiguration to remain applicable in such settings.

Finally, while the notion of artifact, process, and normative coupling is theoretically grounded, the study relies on proxy indicators to identify coupling in case settings. These proxies may not fully capture the richness of socio-technical interaction, underscoring the need for more refined measurement instruments in future research.

7.2. Future Research Directions

The ISA-4 framework opens several promising avenues for future research.

First, the propositions articulated in this study invite **large-scale empirical testing**. Survey-based research could be used to assess coupling strength across layers, while archival and operational data (e.g., deployment frequency, incident rates, recovery time) could be used to evaluate scaling outcomes in relation to coupling configurations.

Second, future work could focus on the **longitudinal dynamics of AI scaling**. The lifecycle perspective proposed by ISA-4 suggests that optimal coupling configurations shift over time. Longitudinal case studies or panel data analyses could examine how organizations transition between experimentation, standardization, platformization, and ecosystem scaling—and how reconfiguration capabilities moderate these transitions.

Third, the framework can be extended to the **ecosystem level**. As AI analytics increasingly depend on external platforms, APIs, foundation models, and third-party services, scaling challenges transcend organizational boundaries. Future research could investigate coupling mechanisms across inter-organizational networks and examine how ecosystem dependencies shape collective scaling dynamics.

Fourth, **sector-specific studies** may refine the boundary conditions of ISA-4. Highly regulated domains such as healthcare and finance impose different governance constraints than consumer-oriented digital platforms. Comparative industry studies could identify how regulatory pressure reshapes coupling priorities and alters optimal scaling configurations.

Finally, there is a need for **formal measurement development**. The coupling constructs introduced in this research—artifact, process, and normative coupling—would benefit from validated scales and metrics. Developing such instruments would support cumulative theory building and enable more rigorous hypothesis testing.

Table 7 Future Research Opportunities

Research Area	Key Question	Suggested Method
Empirical testing	Do coupling mechanisms predict scaling outcomes?	Surveys, archival analysis
Longitudinal studies	How do scaling configurations evolve over time?	Panel data, case studies
Ecosystem analysis	How does scaling extend across organizations?	Network analysis
Industry comparison	How do constraints vary across sectors?	Comparative studies
Measurement development	How can coupling be quantified?	Scale development

The table illustrates that ISA-4 is not a destination, but a starting point for an agenda for AI scaling.

7.3. Summary

This section positions ISA-4 not as a finalized solution, but as a foundational framework for understanding and studying AI analytics scaling. By articulating limitations and outlining a structured research agenda, the study emphasizes that the value of ISA-4 lies in its ability to guide empirical inquiry, refine socio-technical theory, and inform evidence-based scaling strategies over time.

8. Conclusion

The scaling of AI-driven analytics products remains a persistent challenge for organizations. Despite major advances in machine learning techniques and data platforms, many AI initiatives fail to transition from experimental success to sustained enterprise value. This gap reflects a fundamental misunderstanding of scaling as a purely technical endeavor rather than a socio-technical process shaped by the interaction of data infrastructure, operational practices, governance mechanisms, and organizational alignment.

This study addresses that challenge by proposing the Integrated Scaling Architecture for AI Analytics (ISA-4)—a framework that synthesizes technical and managerial perspectives into a unified model of AI scaling. By conceptualizing scaling as the co-evolution of interdependent layers, ISA-4 offers a more realistic and explanatory account of why AI analytics systems succeed or stall at scale. The framework advances prior research by operationalizing coupling mechanisms, introducing a dynamic lifecycle view of scaling, and framing performance as a multiplicative function of resources, reconfiguration capability, and socio-technical fit.

The empirical illustrations demonstrate that scalable AI performance does not depend on isolated excellence in infrastructure, MLOps, governance, or team structure. Instead, sustained scaling emerges from how these elements are configured, coupled, and continuously re-aligned over time. The cases further show that there is no single optimal configuration: organizations achieve scale through different coupling strategies, each involving trade-offs between speed, flexibility, reliability, and coordination cost.

For practitioners, the primary implication is that scaling AI analytics products requires moving beyond tool-centric or function-specific optimization. Effective scaling demands system-level design and operating discipline, including deliberate sequencing of capabilities, embedding governance into technical workflows, and aligning organizational incentives with platform architecture. ISA-4 provides both a diagnostic lens to identify scaling bottlenecks and a strategic guide for navigating trade-offs across the AI product lifecycle.

In conclusion, this research reframes AI scaling as an organizational design and system integration challenge rather than a technology adoption problem. By recognizing and actively managing the co-evolution of technical and managerial systems, organizations can improve their ability to translate AI innovation into durable, enterprise-scale analytics products. The ISA-4 framework offers a foundation for both future empirical research and practical platform strategy, contributing to a more disciplined and effective approach to AI analytics at scale.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Agrawal, A., Gans, J., & Goldfarb, A. (2019). Exploring the impact of artificial intelligence: Prediction versus judgment. *Information Economics and Policy, 47*, 1–6.

[2] Amershi, S., et al. (2019). Software engineering for machine learning: A case study. *Proceedings of the IEEE/ACM International Conference on Software Engineering*, 291–300.

[3] Armbrust, M., et al. (2021). Lakehouse: A new generation of open platforms that unify data warehousing and advanced analytics. *CIDR Conference*.

[4] Akidau, T., et al. (2015). The Dataflow model: A practical approach to balancing correctness, latency, and cost in massive-scale data processing. *Proceedings of the VLDB Endowment, 8*(12), 1792–1803.

[5] Barney, J. B. (1991). Firm resources and sustained competitive advantage. *Journal of Management, 17*(1), 99–120.

[6] Benbya, H., Davenport, T. H., & Pachidi, S. (2020). Artificial intelligence in organizations: Current state and future opportunities. *MIS Quarterly Executive, 19*(4).

[7] Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly, 45*(3), 1433–1450.

[8] Bostrom, R. P., & Heinen, J. S. (1977). MIS problems and failures: A socio-technical perspective. *MIS Quarterly, 1*(3), 17–32.

[9] Davenport, T. H., & Ronanki, R. (2018). Artificial intelligence for the real world. *Harvard Business Review, 96*(1), 108–116.

[10] Davenport, T. H., & Mittal, N. (2022). *All-in on AI: How smart companies win big with artificial intelligence*. Harvard Business Review Press.

[11] Dehghani, Z. (2022). *Data mesh: Delivering data-driven value at scale*. O'Reilly Media.

[12] Dubois, A., & Gadde, L.-E. (2002). Systematic combining: An abductive approach to case research. *Journal of Business Research, 55*(7), 553–560.

[13] Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of Management Review, 14*(4), 532–550.

[14] Enholm, I. M., Papagiannidis, E., Mikalef, P., & Krogstie, J. (2022). Artificial intelligence and business value: A literature review. *Information Systems Frontiers, 24*(5), 1709–1734.

[15] Floridi, L., et al. (2018). AI4People—An ethical framework for a good AI society. *Minds and Machines, 28*, 689–707.

[16] Grønsund, T., & Aanestad, M. (2020). Augmenting the algorithm: Emerging human-in-the-loop work configurations. *Journal of Strategic Information Systems, 29*(4).

[17] Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*, 389–399.

[18] Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. *Academy of Management Annals, 14*(1), 366–410.

[19] Kreuzberger, D., Kühl, N., & Hirschl, S. (2023). Machine learning operations (MLOps): Overview, definition, and architecture. *IEEE Access, 11*, 31866–31879.

[20] Lwakatare, L. E., et al. (2020). DevOps in practice: A multiple case study of five companies. *Information and Software Technology, 114*, 217–230.

[21] Machado, B. B., Costa, L., & Santos, M. Y. (2022). Data mesh: Concepts and considerations. *Procedia Computer Science, 196*, 160–167.

[22] Mikalef, P., & Gupta, M. (2021). Artificial intelligence capability and firm performance. *Information & Management, 58*(3), 103434.

[23] Paleyes, A., Urma, R.-G., & Lawrence, N. D. (2022). Challenges in deploying machine learning: A survey of case studies. *ACM Computing Surveys, 55*(6), 114.

[24] Polyzotis, N., et al. (2018). Data lifecycle challenges in production machine learning. *SIGMOD Record, 47*(2), 17–28.

[25] Rai, A., Constantinides, P., & Sarker, S. (2019). Next-generation digital platforms: Toward human-AI hybrids. *MIS Quarterly, 43*(1), iii–ix.

[26] Sarker, S., Chatterjee, S., Xiao, X., & Elbanna, A. (2019). The sociotechnical axis of cohesion. *MIS Quarterly, 43*(3), 695–719.

[27] Schelter, S., et al. (2018). On challenges in machine learning model management. *IEEE Data Engineering Bulletin, 41*(4), 5–15.

[28] Sculley, D., et al. (2015). Hidden technical debt in machine learning systems. *NeurIPS*.

[29] Shollo, A., Hopf, K., Thiess, T., & Müller, O. (2022). Machine learning value creation mechanisms. *Journal of Strategic Information Systems, 31*(3), 101734.

[30] Teece, D. J. (2007). Explicating dynamic capabilities. *Strategic Management Journal, 28*(13), 1319–1350.

[31] Warner, K. S. R., & Wäger, M. (2019). Building dynamic capabilities for digital transformation. *Long Range Planning, 52*(3), 326–349.

[32] Yin, R. K. (2018). *Case study research and applications* (6th ed.). Sage.

[33] Zhang, K. Z. K., Pentina, I., & Fan, Y. (2021). AI product management roles. *Journal of Business Research*.