International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5838 - 5845
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0756

Architecting AI-Ready Data Ecosystems: A Governed, Scalable and Real-Time Framework for Enterprise Intelligence

Pankaj Sahani[1]
[1]*Purdue university, West Lafayette IN*
Email id: *sahani1987@gmail.com*[1]

Abstract
The fragmented, poorly-governed and temporally inconsistent data environment is often the reason why AI initiatives don't go beyond prototypes to enterprise services that can be relied upon. This conceptual review brings together existing literature on data warehouses, data lakes, lakehouses, data mesh, streaming systems, machine-learning operations and retrieval-augmented generation to identify the features of an AI-ready data ecosystem. AI readiness is understood to be the ability to deliver discoverable, semantically consistent, quality-controlled, traceable and timely data for analytical and operational purposes. A five layer reference architecture is suggested, which includes source ingestion, unified storage, processing and quality, semantic and AI enablement, and intelligence consumption. Governance and observability are considered as cross-cutting control planes and feedback from deployed models and applications is fed back to upper layers. The framework connects architectural decisions with quantifiable metrics like ownership coverage, contract adherence, freshness, lineage completeness, quality test pass rate and training-serving consistency. The study shows that lakehouse mechanisms can be used to reliably share a substrate while data-mesh principles can enhance domain accountability with adequate platform automation and interoperability standards. Business latency budgets, not technology labels, should be used to indicate real-time capability. The suggested framework provides a succinct starting point for organisations looking to modernise their data platforms without fragmenting their data scalability, governance and AI operations into distinct programmes.
Keywords: AI-ready data; enterprise data architecture; data governance; lakehouse; data mesh; stream processing; MLOps

1. Introduction
A model that works well in the experiment may not work well in production because the data used to train the model are different from the data used in production, or because the data sources used to train the model are changed without notice, or because the prediction results cannot be traced back to reliable evidence. Studies on production machine learning reveal that the complexity of modelling code is only a fraction of the long-term complexity of data collection, data validation, feature construction, serving and monitoring [1] [2]. Field studies also demonstrate that defects that are injected upstream can ripple through an organization as data flows, resulting in expensive and hard-to-diagnose failures downstream [3]. The data ecosystem is more important than ever, as foundation models are rapidly being adopted. While foundation models provide general capabilities, enterprise applications demand access to up-to-date and domain-specific information, which must be controlled [4]. Although retrieval-augmented generation is able to link a generative model with external knowledge, its reliability relies on document quality, metadata, retrieval design and provenance [5]. As a result, enterprise AI readiness should be considered as an architectural and governance attribute, not a characteristic of a specific model or cloud service. This paper introduces a simplified conceptual framework to build AI-ready data ecosystems. It provides an overview of the evolution from warehouses to lakes, lakehouses and decentralized data products; outlines the key dimensions of AI readiness; offers a layered reference architecture; and outlines governance controls that can be assessed throughout the implementation process. The goal is not to promote a particular vendor stack, but to outline technology-agnostic design principles that

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5838 - 5845
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0756

are appropriate for organizations that have a mix of batch, streaming, machine-learning, and generative AI workloads.

2. Evolution of Enterprise Data Architecture

2.1.From warehouses to data lakes

The enterprise data warehouse was the foundation for integrated reporting in a controlled environment. Stable analytical definitions are supported by its schema-on-write approach, conformed dimensions and mature transaction mechanisms. But when an organization needs to ingest fast-changing, semi-structured and unstructured data, traditional warehouses are not as flexible. Distributed processing systems solved the scale problem by distributing storage and computation across clusters, and unified engines provided batch, interactive and iterative analysis, all through a common platform [6]. This spurred the idea of the data-lake, where raw data is stored in cheap object storage, and consumers interpret the data as needed. Data lakes increase flexibility, but can create trust issues if assets are not owned, have no data lineage and are not documented or have quality rules. Research on data-lake management reveals discovery and integration and governance as ongoing issues [7]. The issue isn't so much the absence of storage, but the lack of active metadata and enforceable controls. If users can't figure out what the files mean, where they came from, how good they are, what they can do with them, and how recent they are, then it's not an enterprise data product.

2.2.Lakehouse and data mesh

The lakehouse is a transactional metadata layer on top of object storage. Open table formats can support atomic commits, schema enforcement, versioned snapshots and time travel, and maintain the scalability and openness of data-lake storage [8]. These capabilities apply to AI as they allow a training run to be correlated to a snapshot of data that can be reproduced, incremental changes can be processed efficiently, and analytical and machine-learning workloads can share a common governed substrate. Data mesh is not about data storage, it's about a problem in the organization. It puts data product ownership in the hands of the business domains and provides them with a self-service platform and federated governance. Academic analysis describes data mesh as a combination of domain ownership, data as a product, platform enablement and computational governance [9]. Lakehouse and mesh are therefore not mutually exclusive options—the former can be a technical substrate, the latter a responsibility and way of operating. Only when common standards, automated policy and interoperable metadata are available is decentralization beneficial, because otherwise each domain will establish a new silo.

3. Defining AI Readiness

An AI-ready ecosystem provides data that can be safely and repeatedly utilized in training, inference, analytics and retrieval. There are six dimensions of importance. Semantic consistency means that each business concept needs to be defined and have a unit and granularity that is authoritative. Event time, processing time and point-in-time feature values must be separated to prevent future information from bleeding into training, which is temporal correctness [10]. Provenance must be reconstructable, meaning that the transformations and source versions must be able to be constructed. Enforced quality means that executable tests are needed for completeness, uniqueness, validity and distribution and not retrospective manual inspection [11]. Governed accessibility means that data is accessible, findable, reusable, but without compromising authorization. To be ready for retrieval, documents, chunks, embeddings and indexes must be versioned and associated with their sources. These dimensions are an extension of existing data stewardship principles. The FAIR principles state that digital assets should be findable, accessible, interoperable and reusable [12]. Datasheets for datasets document composition, collection, intended use and limitations, in a structured way [13]. Data-governance research also emphasizes the need for explicit assignment of data decision rights and accountability [14, 15]. These concepts need to be translated into machine verifiable controls for production AI. Documentation does not stop an invalid schema from being put into a pipeline and a policy document does not stop an unauthorized query.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5838 - 5845
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0756

Table 1 Comparison of enterprise data architecture paradigms for AI workloads

Dimension	Warehouse	Data lake	Lakehouse	Data mesh
Primary focus	Integrated reporting	Flexible low-cost storage	Unified analytics and AI	Domain-owned data products
Schema model	Schema-on-write	Schema-on-read	Enforced schema with evolution	Contract-defined product schema
Reliability	High transactional maturity	Limited without added controls	Versioned transactions and snapshots	Depends on domain substrate
AI suitability	Good for structured features	Good for raw multimodal data	Strong shared training and analytics base	Strong ownership if standards are mature
Main risk	Rigidity and slow change	Poor discovery and inconsistent quality	Operational and metadata complexity	Fragmentation across domains

4. Proposed Reference Architecture

Figure 1 presents a five-layer architecture. Data moves from operational sources through ingestion, storage, processing, semantic enablement and consumption. Governance and observability apply to every layer, and a feedback path returns model drift, user behavior, incidents and cost signals to upstream components. The separation of layers is logical; an implementation may combine several capabilities in one platform. As shown in Figure 1 Governed and observable reference architecture for an AI-ready enterprise data ecosystem.

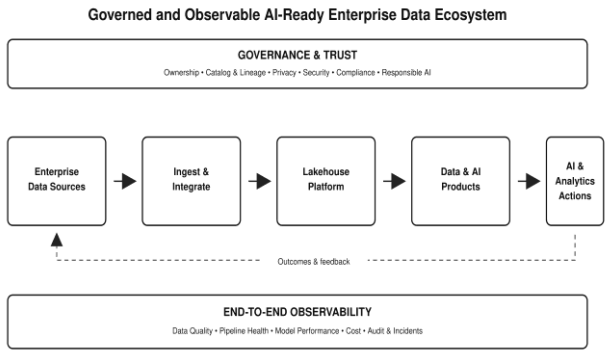


Figure 1 Governed and observable reference architecture for an AI-ready enterprise data ecosystem

4.1.Sources and ingestion

Data is ingested from operational databases, applications, files, devices and external services, into the ingestion layer. Low latency and replay: change-data capture and durable event logs. However, stream-processing research indicates that event time, windowing and late-arriving data must be represented explicitly in order to ensure that results are correct, even if data arrives out of order [10]. Therefore, each source interface should have a versioned contract which specifies structure, semantics, ownership, freshness and compatibility. Contract checks move schema failures to the producer boundary.

4.2.Unified storage

The storage layer should be isolated from the compute layer and should store durable data as well as keeping open and portable formats whenever possible. An immutable raw zone, a standardized conformed zone and a curated product zone are used in a practical organization. Versioned tables enable rollback and reproducibility and cost and metadata growth is managed through retention and compaction policies [8]. Sensitive attributes should be marked on input, and then carried with the data during subsequent transformations.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5838 - 5845
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0756

4.3.Processing and quality

The processing layer transforms data in batches and streams using versioned pipelines. A unified computational model makes it easier to minimize discrepancies between historical recomputation and continuous updates. Quality expectations should be part of the pipeline and block promotion if critical rules are not met [11]. Tests should include schema, missingness, uniqueness, referential integrity and accepted ranges and statistically meaningful distribution changes. These checks are critical as label and feature mistakes can skew training results and production behaviour.

4.4.Semantic and AI enablement

The semantic layer is used to transform technically valid data into reusable organisational meaning. Ownership, lineage, description and permitted use is captured in a catalogue; authoritative measures are stored in a semantic layer; consistent definitions between offline training and online inference are stored in a feature store. Training-serving skew is a significant cause of production failure identified by the Production-ML research [1] [16]. Vector indexes can enable efficient similarity search over embeddings and approximate nearest-neighbor methods can be used for large collections in generative AI [17, 18]. The model, chunking rules, version of the document, and authorization context in which you embed the model must be documented to ensure retrieved evidence can be reproduced and is secure.

4.5.Intelligence and consumption

The consumption layer consists of dashboards, analytical applications, APIs, and model serving and retrieval-augmented systems. Provenance should be revealed in outputs in a way that is commensurate with the decision. Model cards can be used to record performance, intended use and limitations [19] and retrieved answers should keep track of links to supporting passages [5]. Monitoring should include data freshness, feature availability, prediction distribution, retrieval quality, latency and cost. Research on dataset-shift shows that none of the detectors is always reliable, and that monitoring should be done by a combination of statistical signals and outcome and domain evidence [20].

5. Governance and Observability by Design

Governance is made operational when all policies are connected to a technical mechanism, a responsible owner and a measurable result. Table 2 converts the key areas of governance into implementation artefacts and verification indicators. This is a way of thinking about governance as a system design process, as opposed to a documentation process done after deployment. It also aids auditability as control performance can be evidenced in the form of logs, tests and ownership records.

Table 2 Governance controls and measurable verification indicators

Governance domain	Architectural mechanism	Implementation artefact	Verification indicator
Ownership	Domain accountability for every published product	Product record with owner and service objective	Percentage of active products with a verified owner
Metadata and lineage	Automatic capture from pipelines and queries	Searchable catalogue with column-level lineage	Metadata completeness and lineage coverage
Data quality	Blocking validation in deployment pipelines	Versioned quality-test suite	Pass rate and escaped incidents
Contract stability	Compatibility checking at source boundaries	Versioned schema and deprecation policy	Unannounced breaking changes
Security and privacy	Policy-based access and classification	Access rules, purpose tags and audit logs	Unauthorized attempts and review completion

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5838 - 5845
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0756

Freshness and reliability	Service-level objectives with observability	Freshness, volume and latency monitors	Compliance with service targets
AI reproducibility	Versioning of data, features, models and prompts	Linked experiment and deployment record	Reproducible runs and training-serving parity

Control of privacy and security should be maintained near the data. Differential privacy offers a formal framework for reducing disclosure in statistical reports [21] but there are practical considerations that need to be taken into account when implementing the privacy. To retrieve documents in the enterprise, permissions need to propagate to chunks and indexes; otherwise, semantic search can circumvent source-system permissions. Derived assets must also be governed, as embeddings, features and cached responses might contain sensitive information even if the original document is secured.

6. Scalability, Real-Time Processing and Reliability

Scalability is about scaling storage, computation and serving independently, but elasticity is not about getting rid of design constraints. Partitioning, file size, state management and metadata volume have a significant impact on performance. Checkpoints, late events, replay and backpressure are all issues that must be addressed by stream processors without compromising correctness. Real time should therefore be defined as a well-understood and well-stated latency budget that is related to a business decision. Different features may be needed in sub-second, minutes, or even batch. Reproducibility is also a key element in reliability throughout the AI lifecycle. Tracking experiments should link code, configuration, data snapshot, feature definitions and model artifacts. According to the ML Test Score framework, systematic tests for data, models, infrastructure and monitoring are the key to production readiness, not just offline accuracy [22]. Machine-learning studies also show that mature machine-learning teams invest a significant amount of time on data integration, testing and cross-functional coordination [16]. These findings can be used to justify investing in platform controls that can be used in other projects.

7. Implementation Pathway

It is important to have an implementation in controlled stages. The first phase involves creating a list of important data assets, responsible data owners and baseline data quality metrics. The second stage includes standardized ingestion, versioned storage and automated metadata capture. In the third stage, transformation and quality checks are put into reproducible pipelines and latency targets are set for operational use cases. The fourth stage involves adding semantic definitions, re-usable features, retrieval indexes and linked experiment records. The final stage allows domain teams to deploy data products via a self-service approach, with federated governance. The progression should be dependent on evidence of a control maturity, not a new architectural label. Estates should be modernized in a step wise manner. Good candidates are those high-value data flows that fail often or are repeatedly integrated. Migration risk can be minimized by operating in parallel and reconciling. Common capabilities for identity, cataloguing, contracts, lineage, quality, observability should be provided by central platform teams, and meaning and fitness for use should be the domain team's responsibility. This allocation avoids the central team being a semantic bottle-neck and avoids duplication of infrastructure in domains.

8. The discussion and research directions

The proposed framework combines concepts which are commonly used separately. It is based on the premise that AI readiness is a combination of architecture, metadata, governance and operations. Even an unowned lakehouse can become chaotic, and an unstandardized data mesh can compound inconsistencies. Likewise, enterprise knowledge retrieval is not possible without current, authorized, traceable and evaluated contents in a vector

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5838 - 5845
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0756

database. The framework is conceptual and has not been tested by comparison in controlled conditions between organizations. Future research should aim to create standardized readiness measures and explore the association between readiness measures and deployment time, incident frequency, audit effort and model stability. Interoperability of semantics between independently managed domains is a challenging issue. Machine-generated metadata might alleviate the manual work, but still needs to be checked by humans. Further work is needed in the evaluation of retrieval-augmented systems as retrieval relevance, evidence faithfulness and response usefulness are related but distinct outcomes [5, 23]. Platform design will also be influenced by regulatory and ethical considerations. The potential dangers of LLM are bias, opacity, misinformation and resource concentration [24]. So, enterprise architectures need to maintain documentation and access evidence and human accountability throughout the data and model life cycle. AI risk-management guidance is becoming more and more about ongoing governance, not a single approval [25]. In the future, platforms will probably have automated agents for quality remediation, metadata creation, cost optimisation, which will also need to be given limited permissions, logs and review.

Conclusion

Being enterprise AI ready is largely a data ecosystem characteristic. To be reliable, intelligence must be based on data that is semantically clear, temporally correct, traceable, quality-controlled, governed and suitable for both analytical and retrieval workloads. The five-layer proposed architecture links ingestion, unified storage, processing, semantic enablement and consumption via cross-cutting governance and observability. It also provides operational feedback to upstream components to enable the platform to respond to incidents or changing usage and model drift. Lakehouse mechanisms can enhance reliability and reproducibility, and data-mesh practices can enhance accountability, with the help of shared standards and platform automation. Readiness should be assessed by indicators of ownership, metadata, quality, contract, security, freshness and reproducibility, and not just technology adoption. It is an integrated approach that gives a practical base for scalable, real-time and responsible enterprise intelligence.

References

[1]. D. Sculley et al., “Hidden technical debt in machine learning systems,” in Advances in Neural Information Processing Systems, vol. 28, 2015, pp. 2503–2511, doi: 10.48550/arXiv.1504.06520.

[2]. N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich, “Data lifecycle challenges in production machine learning: A survey,” ACM SIGMOD Record, vol. 47, no. 2, pp. 17–28, 2018, doi: 10.1145/3299887.3299891.

[3]. N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo, “Everyone wants to do the model work, not the data work: Data cascades in high-stakes AI,” in Proc. CHI Conf. Human Factors in Computing Systems, 2021, pp. 1–15, doi: 10.1145/3411764.3445518.

[4]. R. Bommasani et al., “On the opportunities and risks of foundation models,” 2021, doi: 10.48550/arXiv.2108.07258.

[5]. P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 9459–9474, doi: 10.48550/arXiv.2005.11401.

[6]. M. Zaharia et al., “Apache Spark: A unified engine for big data processing,” Communications of the ACM, vol. 59, no. 11, pp. 56–65, 2016, doi: 10.1145/2934664.

[7]. F. Nargesian, E. Zhu, R. J. Miller, K. Q. Pu, and P. C. Arocena, “Data lake management: Challenges and opportunities,” Proceedings of the VLDB Endowment, vol. 12, no. 12, pp. 1986–1989, 2019, doi: 10.14778/3352063.3352116.

[8]. M. Armbrust et al., “Delta Lake: High-performance ACID table storage over cloud object stores,” Proceedings of the VLDB Endowment, vol. 13, no. 12, pp. 3411–3424, 2020, doi:

10.14778/3415478.3415560.

[9]. I. A. Machado, C. Costa, and M. Y. Santos, "Data mesh: Concepts and principles of a paradigm shift in data architectures," Procedia Computer Science, vol. 196, pp. 263–271, 2022, doi: 10.1016/j.procs.2021.12.013.

[10]. T. Akidau et al., "The dataflow model: A practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing," Proceedings of the VLDB Endowment, vol. 8, no. 12, pp. 1792–1803, 2015, doi: 10.14778/2824032.2824076.

[11]. S. Schelter, D. Lange, P. Schmidt, M. Celikel, F. Biessmann, and A. Grafberger, "Automating large-scale data quality verification," Proceedings of the VLDB Endowment, vol. 11, no. 12, pp. 1781–1794, 2018, doi: 10.14778/3229863.3236237.

[12]. M. D. Wilkinson et al., "The FAIR guiding principles for scientific data management and stewardship," Scientific Data, vol. 3, Art. no. 160018, 2016, doi: 10.1038/sdata.2016.18.

[13]. T. Gebru et al., "Datasheets for datasets," Communications of the ACM, vol. 64, no. 12, pp. 86–92, 2021, doi: 10.1145/3458723.

[14]. R. Abraham, J. Schneider, and J. vom Brocke, "Data governance: A conceptual framework, structured review, and research agenda," International Journal of Information Management, vol. 49, pp. 424–438, 2019, doi: 10.1016/j.ijinfomgt.2019.07.008.

[15]. V. Khatri and C. V. Brown, "Designing data governance," Communications of the ACM, vol. 53, no. 1, pp. 148–152, 2010, doi: 10.1145/1629175.1629210.

[16]. S. Amershi et al., "Software engineering for machine learning: A case study," in Proc. IEEE/ACM 41st Int. Conf. Software Engineering: Software Engineering in Practice, 2019, pp. 291–300, doi: 10.1109/ICSE-SEIP.2019.00042.

[17]. J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," IEEE Transactions on Big Data, vol. 7, no. 3, pp. 535–547, 2021, doi: 10.1109/TBDATA.2019.2921572.

[18]. Y. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 4, pp. 824–836, 2020, doi: 10.1109/TPAMI.2018.2889473.

[19]. M. Mitchell et al., "Model cards for model reporting," in Proc. Conf. Fairness, Accountability, and Transparency, 2019, pp. 220–229, doi: 10.1145/3287560.3287596.

[20]. S. Rabanser, S. Günnemann, and Z. C. Lipton, "Failing loudly: An empirical study of methods for detecting dataset shift," in Advances in Neural Information Processing Systems, vol. 32, 2019, doi: 10.48550/arXiv.1810.11953.

[21]. C. Dwork, "Differential privacy," in Automata, Languages and Programming, Lecture Notes in Computer Science, vol. 4052, 2006, pp. 1–12, doi: 10.1007/11787006_1.

[22]. E. Breck, S. Cai, E. Nielsen, M. Salib, and D. Sculley, "The ML test score: A rubric for ML production readiness and technical debt reduction," in Proc. IEEE Int. Conf. Big Data, 2017, pp. 1123–1132, doi: 10.1109/BigData.2017.8258038.

[23]. A. Paleyes, R.-G. Urma, and N. D. Lawrence, "Challenges in deploying machine learning: A survey of case studies," ACM Computing Surveys, vol. 55, no. 6, pp. 1–29, 2023, doi: 10.1145/3533378.

[24]. E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in Proc. ACM Conf. Fairness, Accountability, and Transparency, 2021, pp. 610–623, doi: 10.1145/3442188.3445922.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 10 October 2026
Page No: 5838 - 5845
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0756

[25]. E. Tabassi, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," National Institute of Standards and Technology, NIST AI 100-1, 2023, doi: 10.6028/NIST.AI.100-1.