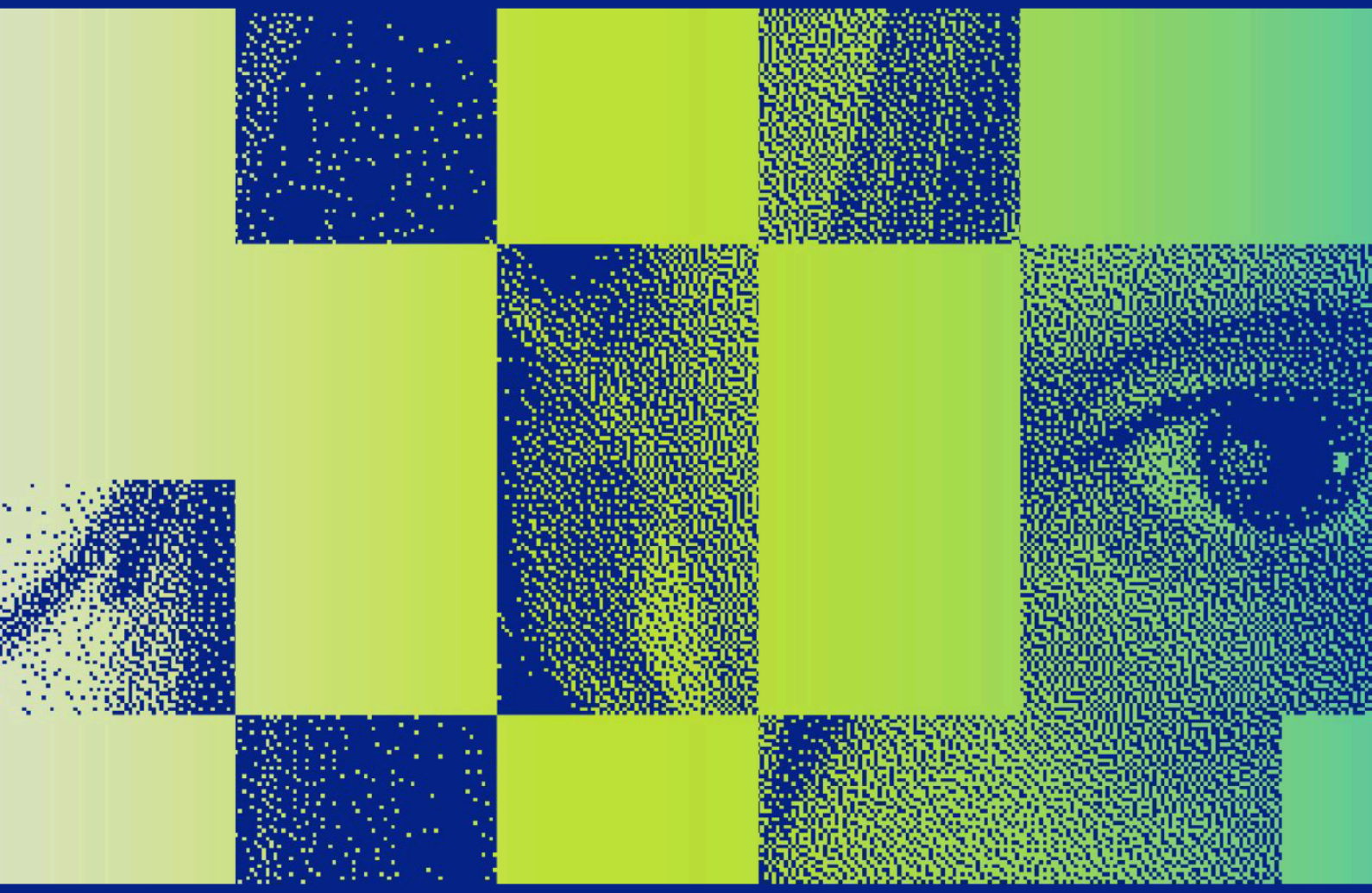

A framework for AI-ready enterprise data

Data-centric AI

DAI

ODI Research
ADVANCING TRUST IN DATA

June 2026

Contents

About

This report, authored by Neil Majithia, Thomas Carey-Wilson, Calum Inverarity, Tara Lee, Elena Simperl and Stuart Coleman, was researched and produced by the Open Data Institute (ODI) and published in June 2026.

We would like to thank all participants in this research, including interviewees and others who provided valuable input. If you would like to share feedback or get in touch, contact the research team at research@theodi.org.

Licence:

IDEA

Most enterprise data was built for human-driven processes, not AI. The result is fragmented systems, unreliable outputs, and stalled investment. Enterprises are pouring capital into AI without the data foundations to make it work, and the cost of locking in the wrong approach compounds quickly. The Open Data Institute (ODI) and SAP have launched IDEA, an **Interchange for Data and Enterprise AI**, to close that gap. IDEA is a vendor-neutral, community-led programme delivering across three integrated workstreams:

- **Applied research** producing evidence-based frameworks, learning and tools.
- **A global peer community** of data, AI, technology leaders, academics, technology partners and practitioners sharing what works, and what does not.
- **Governance:** oversight of these two workstreams will be steered by a global committee of independent thought leaders from industry, academia, and professional communities, ensuring outputs are serving the interests of the community.

Together, these workstreams give data and AI leaders a trusted reference point: cutting through vendor hype, supporting responsible adoption, and building the AI-ready foundational data infrastructure enterprises need to compete. IDEA's first output, A Framework for AI-Ready Enterprise Data, sets out actionable criteria that leaders can act upon to start making their data infrastructure AI ready. If you are leading data or AI initiatives in a global enterprise, or are building products and services that depend on AI-ready data, we want to engage with you. Please join our community at: https://theodi.org/insights/projects/idea-an-interchange-for-data-and-enterprise-ai/

The Open Data Institute

The ODI was founded in 2012 by Sir Tim Berners-Lee and Professor Sir Nigel Shadbolt and is a non-profit institute with a vision to build a world where data works for everyone. Our mission is to create an open, trustworthy data ecosystem on which AI systems and other technologies can depend.

Executive Summary

Across the world, enterprises are deploying AI to orchestrate supply chains, engage customers, and drive operational efficiency. This makes the underlying data foundations upon which AI is built more consequential than ever. When data foundations are not AI-ready, the effects on AI implementations are severe, often invisible, and costly to unwind.

The harder question is what 'AI-ready' actually looks like in practice. This report answers it by setting out 21 concrete, actionable criteria, refined through comparative analysis of 23 existing frameworks and from interviews with eight enterprise data leaders. The criteria are organised across four interdependent pillars: **dataset properties, metadata, surrounding infrastructure** and **governance**. Treated as a whole, they enable enterprises to enhance interoperability, eliminate silos and enrich data with semantic meaning, the conditions under which AI implementations can work reliably at scale.

Category	Criteria	
Dataset properties	a) Adherence to established standards and conventions	
	b) Semantic and logical consistency	
	c) Identifiable class and source imbalance	
	d) De-identification and anonymisation	
	e) Appropriate file formats	
Metadata contents	a) Machine-readable metadata formats	
	b) Attached metadata	
	c) Technical specifications and agentic instructions	
	d) Supply chain and provenance information	i) Collection
		ii) Preprocessing
		iii) Annotation
	e) Legal and sociotechnical information	i) Context and original purpose
		ii) Licence name(s), permissions and URL(s)
		iii) Intended access controls
		iv) Data protection declaration(s)

Surrounding infrastructure	a) Accessibility via a user-centric data product	
	b) Accessibility via API	
	c) Version control infrastructure	
	d) Master data management (MDM)	
	e) Semantic interlinkage and active cataloguing	i) Shared business ontology and knowledge graph
		ii) Active cataloguing infrastructure
	f) Monitoring and observability checkpoints	
Governance	a) Agent-aware access controls via policy-as-code	
	b) Documented roles and responsibilities	
	c) Publicly identifiable points of contact	
	d) Clear data access processes	
	e) An AI-data feedback loop	

Technology alone, however, is not sufficient. Sustained AI-readiness also requires an organisational shift: putting data first in how decisions are made, governed and resourced. Enterprises that combine technical data leadership with a culture of institutionalising data stewardship will be the ones positioned to absorb today's AI capabilities and rapidly convert them into tomorrow's outcomes for their organisations and their people. This framework is a starting point, not a finished position. It will evolve through application, challenge, and the experience of enterprises putting it to work. Test the criteria against your own environments, share what you learn, and consider contributing to the next iteration of our work or feeding back directly to our research team: research@theodi.org

Introduction

With the current landscape of innovation in AI, data has never been more crucial for enterprise. Consumption data can be used to train machine learning models that forecast revenues,[1] while textual documentation about products can be used to ground a chatbot that assists customers or sales teams with up-to-date, accessible information.[2] At the cutting edge, AI agents given a task can rapidly build on large-scale data sources, extracting insights and hidden patterns from them to improve the way a business works,[3] while 'embodied AI', with digital twins and robotics, interfaces with data to bring the digital world to the physical one.[4]

However, for a business to actually reap the benefits of AI, the data foundations underneath it must be optimised. If they are not 'AI-ready', AI implementations risk negative outcomes that can undermine investment and harm business activity.

At the ODI, we have defined AI-ready data as "data that is technically optimised for machine learning, high quality and adherent to standards, legally compliant, and responsibly collected". Our research has set out several criteria for datasets, metadata, surrounding infrastructure, and governance protocols to meet this definition, each published as part of our framework for AI-ready data.[5]

This framework has already provided actionable guidance to many different organisations, including the UK government. Enterprises, too, need this kind of guidance; while many organisations may have established data strategies and leadership, aligning these with the rapidly evolving world of AI is complex. We see the need to tailor our framework for the enterprise audience, iterating its contents and adding more elements to suit the needs of business at the cutting edge.

In this report, we therefore set out to build **a framework for AI-ready enterprise data**, using comparative analysis, several expert interviews, and our institutional experience at the ODI to isolate the steps businesses must take to transform their data into something that enables AI-powered enterprise.

This framework for AI-ready enterprise data provides 21 concrete, actionable criteria for businesses to take as guidance to enhance their data practices for AI. Each is unique to the context of enterprise AI-readiness, which we establish and define before going into depth on the recommendations in our framework. The framework can be used to design or evaluate data infrastructure for its

[1] Makridakis, S., Spiliotis, E. & Assimakopolous, V. (2022), "M5 accuracy competition: Results, findings, and conclusions"
[2] Jia, Z. et al. (2024), "Path to high-quality LLM-based Dasher support automation"
[3] Business Wire (2026), "Zapier Extends Enterprise AI Governance Across Every Surface Where Building Happens"
[4] Zhang, J., Wang, L. & Gao, R.X. (2025), "Embodied AI: A Foundation for Intelligent and Autonomous Manufacturing"
[5] Majithia, N., Carey-Wilson, T. & Simperl, E. (2025), "A framework for AI-ready data"

overall maturity in the context of AI, and is suitable for all audiences – technical and non-technical – that are looking to understand how the data infrastructure of today can underpin the AI-powered businesses of tomorrow.

This research was conducted as a component of our partnership with SAP on the Interchange for Data and Enterprise AI (IDEA). As with our previous framework for AI-ready data, it will be refined and updated over time, and it will also be used as a basis for further research in which we dive deep with specific enterprises to explore their AI and data ecosystems. If you would like to be involved in these next steps, please let us know by emailing research@theodi.org.

Our definition of AI-ready data

Methodology

In this research, we sought to develop a set of actionable criteria that together define the properties that data requires to facilitate and enable AI adoption in the enterprise.

The initial stage involved a comprehensive desk review of the commercial data landscape, during which we curated a set of 23 already-published AI-readiness frameworks from consultancies, technology providers, and consortium organisations (listed in Appendix I). These enterprise frameworks were selected because they offer practical, actionable guidance to corporate decision-makers, distinguishing them from more academic or foundational principles like the FAIR principles.

To identify a consensus set of requirements, we conducted a comparative analysis of these frameworks. Recognising that some expressed identical criteria using entirely different terminology, we anticipated heterogeneity in the language used. We therefore systematically reduced the criteria in each framework to standardised descriptions, often splitting complex requirements into constituent parts to ensure they were represented in a common, shared language. We could then perform the comparative analysis and isolate criteria shared across the enterprise data frameworks.

This analysis produced two things. First, it identified an overlap between enterprise AI-readiness frameworks and our own previous iteration that contained 17 established criteria. We were able to refine these and recontextualise them for enterprise, carrying them forward into the new framework developed in this report.

Second, it produced a longlist of consensus criteria that multiple organisations and stakeholders agreed were important for AI-readiness. We built this list by combining conceptually similar requirements and discarding those that were uncommon across the frameworks we reviewed. This resulted in an initial longlist of 10 criteria that would guide the development of the fifth pillar of our framework: low latency data access; continuous iteration via an AI ‘flywheel’; unified and multimodal storage; robust data pipelines; observability, monitoring, and quality enforcement; dynamic access control; multi-jurisdictional compliance; enterprise-level metadata; active cataloguing; and master data management (detailed further in Appendix II).

The final stage involved a series of domain expert interviews to qualitatively validate and ratify our criteria selections. We conducted interviews with eight individuals (Appendix III), using their institutional experience and guidance regarding real-life implementations to further refine the longlist to six criteria, having removed some elements and combined others. After drawing two of these into the previous 17, our research culminated in a final framework comprising 21 actionable criteria for AI-ready enterprise data, supported by a corpus of evidence outlining their importance and implementation.

Context: AI-ready data in enterprise

We define AI-ready data as “data that is technically optimised for machine learning, high quality and adherent to standards, legally compliant, and responsibly collected”.[6] By fitting this definition, data can maximally support AI implementations.

However, what does this specifically mean, especially in the context of enterprise? In our research, we explored what published literature, and our interviewees, had to say about the landscape of AI-ready data in the private sector.

Enterprise datasets do not exist in a vacuum. Each piece of data in an organisation is indelibly linked to some other. Sales data relates to customer data, which relates to marketing data, and although each can be analysed independently, their strength and depth lies in being analysed together. It is therefore important to measure not only the AI-readiness of an enterprise dataset by itself, but also how it sits within its organisation’s broader data ecosystem.[7] [8]

Relatedly, an enterprise dataset does not sit self-contained within a file on a disk or in the cloud. The reality of enterprise data is fragmented and heterogeneous, locked inside ERPs, CRMs and other operational systems, dispersed across SaaS platforms, or held in legacy data warehouses.[9] To make the data in this environment available for analysis, data is packaged, maintained and accessed through data catalogues and data products.[10] For an organisation's data to be AI-ready, the underlying operational systems in which the data sits and the data products built on top of them must be modernised together.

Secondly, AI interacts with data in enterprise in a broad spectrum of ways. A primary use case is, of course, in using statistical data to train machine learning models for analytics or forecasting. However, AI deployments now extend much further. For example, retrieval augmented generation (RAG)[11] grounds AI chatbots that are used to sales teams or other employees when they need help finding and using internal documentation.[12] [13] AI-readiness therefore includes making this documentation – itself, a form of data – easily discoverable and accessible via RAG workflows.

---

[6] Majithia, N., Carey-Wilson, T. & Simperl, E. (2025) “A framework for AI-ready data”
[7] Capgemini UK (2024), “Data powered enterprises 2024”
[8] Ransbotham, S. et al. (2019), “Winning With AI”
[9] MuleSoft (2026), “2026 MuleSoft Connectivity Benchmark Report”
[10] Krantz, T. & Jonker, A. (n.d.), “What Is a Data Product?”
[11] Lewis, P. et al. (2020), “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”
[12] Jia, Z. et al. (2024), “Path to high-quality LLM-based Dasher support automation”
[13] Deloitte UK (2026), “The State of AI in the Enterprise”

However, AI is increasingly being proposed to go beyond this use case. AI agents are autonomous LLM-based systems that can take a single user prompt, independently search for datasets to answer it, write code to analyse them, self-improve, and return in-depth reports and visualisations within minutes.[14] [15] This presents an opportunity for significant gains in efficiency and overall capability within an enterprise, facilitating faster, data-led decision-making. However, this interaction pattern between the agents and the organisation's data requires entirely unique forms of AI-readiness, necessitating not only more technical optimisation but also new guardrails and governance processes that enable these agents to be trustworthy.

These considerations guided the development of our framework. We adapted our findings from the comparative analysis with them in mind, then used them as focal points for interviewee discussions.

Our framework for AI-ready enterprise data

In our prior work, we defined the baseline requirements for AI-ready data across four pillars: **dataset properties** (the core properties of the data that sits within datasets), **metadata** (the contextual information attached to the data), **surrounding infrastructure** (the technology that data sits within), and **governance** (the processes for responsible use of the data).

To build our framework for AI-ready enterprise data, we first tailor the guidance under each of these pillars to the context of enterprise: in a global corporate environment, the baseline principles of data quality and structure remain essential, but their implementation must scale to support vast, interconnected business operations and autonomous AI agents.

Our six new criteria for enterprise, built from our comparative analysis and interviews, are appended to the infrastructure and governance pillars.

[14] Gutowska, A. & Stryker C. (n.d.), "What are Agentic Workflows?"
[15] McKinsey (2025), "The agentic organization: Contours of the next paradigm for the AI era"

Category	Criteria	
Dataset properties	a) Adherence to established standards and conventions	
	b) Semantic and logical consistency	
	c) Identifiable class and source imbalance	
	d) De-identification and anonymisation	
	e) Appropriate file formats	
Metadata contents	a) Machine-readable metadata formats	
	b) Attached metadata	
	c) Technical specifications and agentic instructions	
	d) Supply chain and provenance information	i) Collection
		ii) Preprocessing
		iii) Annotation
	e) Legal and sociotechnical information	i) Context and original purpose
		ii) Licence name(s), permissions and URL(s)
		iii) Intended access controls
		iv) Data protection declaration(s)
Surrounding infrastructure	a) Accessibility via a user-centric data product	
	b) Accessibility via API	
	c) Version control infrastructure	
	d) Master data management (MDM) and authoritative sources of truth	
	e) Semantic interlinkage and active cataloguing	i) Shared business ontology and knowledge graph
		ii) Active cataloguing infrastructure
	f) Monitoring and observability checkpoints	
Governance	a) Agent-aware access controls via policy-as-code	
	b) Documented roles and responsibilities	
	c) Publicly identifiable points of contact	
	d) Clear data access processes	
	e) An AI-data feedback loop	

Dataset properties

Enterprise datasets are typically assembled from many sources, formats and jurisdictions. At the dataset level, inconsistencies compound as AI systems scale. For instance, ambiguous labels have downstream effects on machine learning processes, regional skew introduces biases, and weak anonymisation creates compliance exposure that only surfaces once a model is in production. The criteria below address these foundational risks, ensuring that the data feeding into corporate AI ecosystems is robust, interoperable, and safe to use at scale.

Adherence to established standards and conventions

Enterprises operate across borders and diverse software ecosystems. Adhering to recognised international standards ensures that data flows seamlessly between different departments and supply chain partners. For example, using ISO-3 codes[16] for geographical data, or ISO-8601[17] for timing information, guarantees that AI models forecasting global logistics or sales trends operate on a consistent, interoperable baseline.

Semantic and logical consistency

Data labels must be semantically clear and logically applied across the organisation. In a corporate dataset, ambiguous labelling can cause an AI agent to conflate distinct business concepts, leading to operational errors. Annotators and data stewards must adhere to a set protocol and use internationally recognised vocabularies. Implementing linked data protocols and persistent uniform resource identifiers ensures that datasets are functional and shareable across the enterprise network.

Semantic and logical consistency is fundamental for semantic linkage of datasets via a knowledge graph, another recommendation of our framework.

[16] IBAN (n.d.), "Country codes alpha-2 & alpha-3"
[17] ISO (n.d.), "ISO 8601 — Date and time format"

Identifiable class and source imbalance

Enterprise data is frequently aggregated from multiple localised systems or external vendors. If an aggregated dataset contains significantly more information from one regional market than another, it introduces a class imbalance that can severely skew an AI model's performance. To mitigate this, datasets must explicitly document provenance information for each data entry, making any source imbalances easy for data scientists and automated systems to identify.

De-identification and anonymisation

Corporate datasets frequently contain sensitive personal information regarding customers and employees. To protect subjects' privacy and safety, standard de-identification and anonymisation methods must be applied before this data is made accessible to broader AI systems. Data publishers must explicitly account for the elevated re-identification risk that occurs when data is used to train or prompt machine learning models.[18]

Appropriate file formats

The scale of enterprise data necessitates highly efficient file formats. While comma-separated value files remain useful due to their flexibility, formats like Apache Parquet offer distinct advantages for enterprise AI. Parquet's columnar storage, innate compression, and multimodal capabilities make it highly optimised for the massive scale of corporate data lakes. Highly relational enterprise datasets can use JSON-LD or the Resource Description Framework format.

[18] Curzon, J. et al. (2021), "Privacy and Artificial Intelligence"

Metadata contents

Even when enterprises hold high-quality data, the metadata describing it is often inconsistent, incomplete, or written for humans rather than machines. Without rich, machine-readable context, AI systems can silently misinterpret synthetic records as real, miss critical operational caveats, or violate licensing and privacy terms they were never told existed.

Metadata management is already well-established within enterprise ecosystems. In the financial sector, the Financial Industry Business Ontology[19] is widely used to standardise complex semantic relationships and regulatory reporting. Retail and e-commerce rely heavily on Schema.org[20] to structure product specifications and supply chain data, while the Simple Knowledge Organisation System[21] helps organisations transform unstructured content into structured, machine-readable data, fostering interoperability across disparate systems and creating a foundation for knowledge graphs. The challenge, however, is making this metadata serve AI as effectively as it currently serves humans.

Machine-readable metadata formats

Metadata must be stored and accessible so that programmatic workflows and automated systems can parse it effectively. Using standards like Croissant[22] to format metadata into machine-readable structures enhances interoperability and discoverability across an organisation's internal data catalogues. This ensures that AI agents can autonomously read and understand the context of the data they are retrieving.

Attached metadata

Datasets must be provided to users and systems with their corresponding metadata directly attached. When querying enterprise APIs or downloading from portals, the return payload should include both the raw data and its contextual information. Integrating metadata directly with the data payload guarantees that AI applications and human practitioners have immediate, frictionless access to the context required for accurate analysis.

---

[19] EDM Council (n.d.), "FIBO Interest Group"
[20] Schema.org (n.d.), "Schema.org"
[21] Miles, A. & Bechhofer, S. (2009), "SKOS Simple Knowledge Organization System Reference"
[22] Majithia, N., Carey-Wilson, T. & Simperl, E. (2024), "Transforming AI data governance with Croissant: a new standard for ML metadata"

Technical specifications and agentic instructions

Enterprise metadata must detail a dataset's modalities, dimensionality, semantics, and basic summary statistics. This allows practitioners and automated systems to evaluate a dataset's utility efficiently. Furthermore, any dataset component that has been synthetically generated or machine-annotated must be explicitly demarcated. Identifying synthetic components ensures that downstream AI models and human evaluators can appropriately adapt their use of the data during training and inference, aligning with emerging global regulations.

To support autonomous systems, this semantic layer can also include granular agentic instructions that detail specific business logic. This includes semantic descriptions of column relationships, specific operational caveats such as treating negative revenue numbers as customer refunds, and the exact logic required to accurately join the dataset with other systems. Providing this operational manual ensures that AI agents do not misinterpret data structures when independently executing tasks.

Supply chain and provenance information

To operationalise responsible AI principles within a corporate structure, metadata must capture the data's entire lifecycle. This includes detailing the methods of collection, the preprocessing pipelines used, and the roles of any outsourced entities or external vendors. Comprehensive provenance tracking with ontologies like PROV-O[23] enables dynamic audit processes, allowing enterprises to trace data lineage and ensure compliance requirements are integrated seamlessly throughout the supply chain.

Legal and sociotechnical information

Enterprise data often operates under strict internal governance and external licensing constraints. Data publishers must provide clear statements regarding intended access protocols, data protection declarations, and any applicable licensing terms directly within the metadata. Documenting these rules ensures that AI systems and human operators maintain strict adherence to legal, ethical and commercial boundaries when processing enterprise assets.

[23] Lebo, T., Saho, S. & McGuiness, D. (2013), "PROV-O: The PROV Ontology"

For example, the DPV (Data Privacy Vocabulary)[24] enables enterprises to codify legal bases, processing purposes, and technical security measures into machine-readable metadata, ensuring that AI systems and human operators can automatically verify compliance with privacy regulations and internal governance policies.

Surrounding infrastructure

Enterprise data infrastructure was largely built for human analysts working with curated, periodic datasets, not for AI agents that query continuously, traverse systems autonomously, and act on what they find. For a dataset to be AI-ready, it must be hosted within a surrounding infrastructure that is equally equipped for machine learning workloads. In an enterprise context, this infrastructure must reliably serve data at scale, securely bridging internal silos and organisational boundaries to feed both human practitioners and autonomous AI systems.

Accessibility via a user-centric data product

As mentioned previously, enterprise datasets are often packaged with others, maintained together, and made accessible via data catalogues and products.

This infrastructure can take the form of data lakehouses,[25] data fabrics,[26] data meshes,[27] or other unified architectural patterns, providing underlying systematised organisation methods for datasets as well as user interfaces that sit on top. In practice, they facilitate active engagement with data, supporting automated and manual workflows with interactive tooling for dataset exploration, evaluation, maintenance and governance[28]. This product-oriented infrastructure empowers practitioners to seamlessly assess and integrate massive, continuously updated datasets into AI systems, maintaining a responsive environment for rapid analytics and agentic discovery.

[24] Esteves, B. et al. (2026), "Data Privacy Vocabulary (DPV)"
[25] Databricks (n.d.), "What is a Data Lakehouse?"
[26] Jonker, A. & Krantz, T. (n.d.), "What Is a Data Fabric?"
[27] Amazon Web Services (n.d.), "What is a Data Mesh?"
[28] Corrie (2026), "One Data Catalog: AI-Powered Metadata and Search for Better Data Governance"

Accessibility via API

An AI-ready data infrastructure relies on standardised protocols suited for programmatic access. Datasets should be accessible via application programming interfaces that facilitate the querying of subsets without imposing artificial bottlenecks like forced pagination. These platforms preserve data sovereignty and integrity while supporting the reliable, real-time data integration required by large-scale machine learning pipelines.

Version control infrastructure

A robust version control infrastructure is a foundational requirement for enterprise AI applications. To ensure accuracy and relevance, generative AI and autonomous agents depend on a continuous stream of corporate data, yet versioning must extend significantly beyond raw records. To facilitate reproducibility, auditability and safe rollback, infrastructure must track granular transformations across four interlinked artefacts:

- the **datasets** (including both raw and derived records, as well as the snapshots used to train or ground specific AI deployments);
- the **schemas** that define their structural baseline;
- the **semantic models** (ontologies, taxonomies and knowledge graph mappings) that provide business context;
- and the **policies** that govern their access and deployment.

A modification to any single component can silently compromise downstream AI performance; treating these as separately versioned, jointly auditable assets enables comprehensive provenance tracking, seamless integration of new insights, and a definitive evidentiary trail for corporate risk management and regulatory compliance.

Master data management (MDM) and authoritative sources of truth

Our research indicated that enterprises frequently hold fragmented or duplicated records for core business entities across different departments and systems.[29] [30] For example, a customer might have multiple records in the CRM, billing and logistics systems, each with slightly different data as a result of their different interactions with the business.

As AI implementations like internal copilots or autonomous agents are deployed to query and act upon this data, the absence of authoritative records becomes a significant operational risk, with decisions and actions taken on the basis of inconsistent or inaccurate information. These AI tools can also cloud the landscape further by, for example, accelerating the generation of new insights, summaries and derivative data, potentially obscuring which records are authoritative and which are not.

Master data management (MDM)[31] is a means of governance that enables an organisation to establish authoritative sources of truth for its core business entities. Namely, these are secure, validated records (often called 'golden records') that act as the reference point for all downstream use. Rather than pretending the underlying data landscape is unified, MDM acknowledges that enterprise data is fragmented across systems and provides a clear designation of which records carry authority. This means that no matter how many interconnected AI systems are querying or generating data, they are operating from, and contributing back to, a verified, consistent foundation.

Implementing MDM at scale requires more than just assigning some datasets to be a single source of truth. Establishing these 'golden records' means deploying data integration pipelines to look at the disparate data across the organisation and perform entity resolution to identify, match and merge duplicate data, outputting authoritative, standardised sources of information to be stewarded effectively and used for trustworthy AI applications.

[29] Experian (2022), "2022 Global data management research"
[30] Duplessie, S. (2019), "Mass Data Fragmentation Is Quietly Killing Digital Transformation Efforts"
[31] Parker, S. (n.d.), "Master Data Management: Definition, Process, Framework and Template"

Semantic interlinkage and active cataloguing

The metadata accompanying a dataset in an enterprise organisation is often complete and in-depth, given the metadata management solutions that are integrated into data infrastructure.[32] Functionally, however, metadata can do more than just inform users about the context or provenance of the dataset at hand, especially for AI. In an enterprise ecosystem, AI agents and RAG applications need to autonomously navigate complex business logic and join the dots across disparate systems; metadata must therefore evolve from static, descriptive tags into a machine-readable map of the business.

This can be achieved through semantic interlinkage: mapping metadata to a shared business ontology that enables all data to be connected with a knowledge graph rather than kept in isolated silos. With this knowledge graph, an AI agent can understand the complex relationships and nuances between the data it has at its disposal.

For example, if an employee asks an AI agent to summarise all active risks associated with a specific corporate partner, the agent uses the knowledge graph to actively traverse the enterprise architecture. Because the shared ontology explicitly maps the CRM's 'Account', the legal database's 'Counterparty' and the procurement system's 'Vendor' to a single, unified business entity, the agent seamlessly retrieves the associated contracts, support tickets and supply chain delays to generate one cohesive summary. This interconnected semantic fabric therefore empowers AI to instantly traverse the data ecosystem of the organisation to execute complex tasks that would traditionally require a data engineer to manually construct custom joins.[33]

However, the practices of establishing the knowledge graph and keeping it updated over time require semantic interlinkage to be paired with 'active cataloguing'. This process leverages machine learning to continuously profile new data as it is collected and processed, tracking its lineage and annotating it with respect to the business ontology to update the knowledge graph in real-time without human intervention, ensuring the business is always operating on an up-to-date living, evolving relational database.

---

[32] Collibra (n.d.), "What is Metadata Management?"

[33] Importantly, a persistent, maintained knowledge graph means an AI agent does not have to rediscover conceptual mappings between data at runtime on every single invocation. This ultimately saves time and expenditure.

This element of AI-readiness therefore comprises three components: the shared business ontology, the knowledge graph platform, and the active cataloguing infrastructure.

Shared business ontology and knowledge graph

A shared business ontology is a universal vocabulary that defines the exact meaning of business concepts and the rules governing them. For example, organisations in the financial services sector frequently adopt the Financial Industry Business Ontology (FIBO), mapped alongside schema.org extensions, to provide a standardised semantic framework for complex financial products and relationships.

Knowledge graphs operationalise the business ontology by mapping these rules to actual data instances, creating an interconnected network of nodes and edges. Enterprises can link disparate concepts (such as ‘people’, ‘locations’ and ‘corporate events’) into a continuous web by deploying graph database platforms like Neo4j. This enables GraphRAG applications, allowing AI agents to traverse deterministic relationships rather than relying on simple vector similarity.

Active cataloguing infrastructure

Platforms such as Alation, Collibra and Atlan deploy active metadata intelligence to autonomously scan data streams, infer schemas, and classify sensitive information without manual human intervention.

Monitoring and observability checkpoints

With the adoption of (agentic) AI systems, enterprise organisations have access to high-performance, low-latency tooling that is flexible and self-improving.[34] [35] The frequency of direct human interaction with data will therefore naturally decrease, and so too will the level of oversight on data. If data collection silently goes awry, there is a chance of a ‘data cascade’, an invisible, negative impact on AI processes that can greatly harm an organisation.[36]

Purpose-built monitoring and observability architectures therefore become more important within data infrastructure. This can be somewhat automated; simple

[34] Gutowska, A. & Stryker C. (n.d.), “What are Agentic Workflows?”
[35] McKinsey (2025), “The agentic organization: Contours of the next paradigm for the AI era”
[36] Sambasivan et al. (2021), “Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI

logic gates can be used to catch anomalies, schema drift or data corruption before they propagate into interactions with AI systems. But human-in-the-loop observability checkpoints, where automatic systems flag edge cases, outliers, or low-confidence outputs for human review, not only reduce the chances of errors but also increase the level of trust people have with the system – at only a minimal cost to latency.

Implementing this 'circuit breaker' approach involves defining explicit 'data contracts' that specify expected schemas, acceptable distributions and strict quality thresholds. If ingested data violates this contract, automated observability platforms act as circuit breakers to immediately halt the pipeline or divert the data into a quarantine queue.

Data governance for enterprise

Traditional enterprise governance was designed around the assumption that humans are users of data. Analysis of data would come after governance policies have been read, access requests have been signed off, and judgement has been exercised, all at human speed. But AI agents are polar opposites: to complete analysis at superhuman speed, they query at machine speed and scale, eagerly pulling whatever data they can reach to complete a task – in the process, skipping over natural-language data policy. A new approach to data governance is therefore necessary.

Agent-aware access controls via policy-as-code

A large component of AI-readiness is making data assets usable by both humans and machines. Crucially, this must extend to policy and governance: a human employee can read a governance handbook or a privacy policy and understand the nuance of what they are allowed to share, but an AI agent cannot reliably interpret or enforce natural language guidelines.

LLM-based AI agents will always try their best to fulfil a task they are given, eagerly traversing whatever data they have available to construct a comprehensive solution. This presents a significant risk to data governance: if an agent has unfettered access to an enterprise's unified data architecture, it will inevitably pull sensitive, inappropriate or restricted information simply because it is trying to be as thorough as possible.[37]

[37] OWASP Gen AI Security Project (2025), "LLM06:2025 Excessive Agency"

Because these AI systems are so zealous, organisations need to strictly micromanage their data access protocols. It is no longer viable to rely on static database permissions where a dataset is either entirely locked away or entirely open. Access must become dynamic and context-aware; the AI agent should only ever be able to physically 'see' the specific rows, columns or documents that the human user prompting it has the explicit credentials to view at that exact moment.

Relying on system instructions written in markdown files or text-based prompts to enforce restrictions on an agent's behaviour is inherently risky. Agents can easily misinterpret these soft instructions, or users can bypass them entirely through prompt injection.[38] True governance requires 'policy as code'. By using machine-readable standards – such as the W3C's Open Digital Rights Language (ODRL),[39] implemented via Croissant 1.1,[40] or open-source policy engines like Open Policy Agent (OPA)[41] – organisations can embed access rules directly into the infrastructure. The system automatically redacts or filters data at the gateway level, long before the AI model can even access it.

Ultimately, dynamic data access powered by policy as code is about enabling accountability. If an AI system generates a problematic output, surfaces restricted personal data or breaches a compliance boundary, an organisation needs to know exactly why it happened. Relying on vague system prompts makes auditing nearly impossible; using hard-coded, machine-readable policies instead provides a clear, traceable log of exactly what data was permitted to be accessed and under what rules, ensuring that when things do go wrong, there is a definitive mechanism for investigation and remediation.

Documented roles and responsibilities

Enterprise datasets must clearly define and assign key governance roles, such as designated data owners or stewards, with explicit accountability. As AI systems autonomously query and process data across the organisation, maintaining a clear chain of responsibility ensures that the provenance and stewardship of the underlying asset are never obscured.

[38] OWASP Gen AI Security Project (2025), "LLM01:2025 Prompt Injection"
[39] W3C (2018), "ODRL Information Model 2.2"
[40] Carey-Wilson, T. & Simperl, E. (2026), "Croissant 1.1"
[41] Open Policy Agent, (n.d.), "Open Policy Agent"

Publicly identifiable points of contact

Metadata should include direct contact information for designated data stewards. In a complex corporate environment, this provides an essential feedback loop. It allows both human practitioners and automated monitoring systems to route data quality anomalies, schema drift alerts or governance issues directly to the accountable party for rapid remediation.

Clear data access processes

The procedures required for human users and machine accounts to access data must be clearly articulated. Fully describing the requirements, criteria and timelines for data access ensures that developers and AI applications can securely and efficiently onboard new datasets, preventing friction in the deployment of enterprise AI workflows.

An AI-data feedback loop

AI deployments are not purely static ‘consumers’ of data; they also contribute to the data ecosystem of the organisation within which they operate. They generate classifications, confidence scores, and user corrections that, if systematically captured and analysed, provide a foundational data asset that can be used to iterate and refine AI systems that are being implemented. Thus begins a ‘flywheel’ effect, a self-sustaining cycle that optimises the usefulness and trustworthiness of the innovations being adopted.[42]

Enabling this feedback loop requires intentionality. Organisations must purposefully design mechanisms that capture telemetry, user feedback (like thumbs up / down, or explicit corrections to AI outputs) and model outputs (like confidence intervals).

Beyond capturing this feedback, though, organisations need to act upon it. Developers of AI integrations should identify issues raised by users, run their own systematic evaluations to confirm findings, then version, test and deploy iterations with changes that reflect what users want and need.

[42] Amazon Web Services (n.d.), “Data Flywheel”

Conclusion

What happens when an organisation's data is not AI-ready? Sambasivan et al.[43] define 'data cascades' as the compounding, negative impacts on AI implementations when data quality is undervalued, noting that these downstream effects are severe, pervasive, invisible and delayed – but avoidable, if data is optimised before use.

Making a commitment to this optimisation, to making enterprise data AI-ready and treating it as critical organisational infrastructure, therefore allows a business to avert data cascades that could massively harm operations or undermine investment, stopping ambitious pilots stalling in production, ensuring agents do not act on stale or conflicting records, and mitigating the likelihood of sensitive information leaks that create compliance exposure too great for legal teams to respond to.

This research provides actionable guidance for this new frontier. Our framework's four pillars are tailored and extended to fit the enterprise context, providing the instructions to construct optimal data infrastructure for businesses looking to implement AI successfully.

But the four pillars are not independent of one another, nor the 21 criteria underneath them. They are intrinsically linked, and developing AI-ready data requires a holistic approach to designing systems and tools that ensure datasets, metadata, infrastructure and governance are considered together.

Furthermore, multiple interviewees highlighted that true AI-readiness has a sociotechnical, organisational element that sits atop our framework. Data modernisation takes time and investment, requiring buy-in from senior leadership to allocate resources; it is therefore imperative that across a business, everyone must be on the same page regarding the necessity of AI-readiness. If AI adoption is prioritised and fast-tracked but data improvement is left behind, Sambasivan et al.'s cascades are soon to follow.

Finally, it is important to note that AI-ready data practices do not solely support AI contexts. Instead, the recommendations laid out in our framework present overall best practices for enterprise data to be used to its full potential, facilitating decision-making, improving efficiency, and providing the foundations for innovation.

Conducted as a component of our partnership with SAP on the Interchange for Data and Enterprise AI (IDEA), this framework will be refined and updated over time. We will test and iterate on these principles, exploring the AI and data ecosystems of enterprises in depth.

---

[43] Sambasivan et al. (2021), "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI"

APPENDIX I

Table I.1 lists the 32 published frameworks we found that were focused on or relevant to commercial, enterprise data. We filtered this to 23 for our comparative analysis, discarding nine (marked with asterisks) due to inaccessibility or their lack of depth regarding relevant, actionable guidance for data practices.

Name	Organisation	URL
A framework for AI-ready data	Open Data Institute	https://theodi.org/insights/reports/a-framework-for-ai-ready-data/
AI-ready data essentials	Gartner	https://www.gartner.com/en/articles/ai-ready-data
The 8-Pillar Framework for AI-Ready Data	Astreya	https://www.linkedin.com/pulse/8-pillar-framework-ai-ready-data-astreya-ntree
What is AI-ready data? Definition and architecture	Dremio	https://www.dremio.com/blog/ai-ready-data/
AI-Ready Data: What It Is and Why It Matters	Dryviq	https://dryviq.com/ai-ready-data/
Prevent Bad AI Checklist	Actian	https://www.actian.com/wp-content/uploads/2025/12/The-Bad-AI-Prevention-Checklist.pdf
The Step-By-Step Guide To Making Your Data AI-Ready	Orases	https://orases.com/blog/the-step-by-step-guide-to-making-your-data-ai-ready/
Data 4.0: making your data AI-ready	Ernst and Young (EY)	https://www.ey.com/content/dam/ey-unified-site/ey-com/en-in/insights/ai/documents/ey-data-4-0-making-your-data-ai-ready.pdf
The AI-Ready Data Framework	AI-Ready Data Initiative	https://ai-ready-data.dev/
Unlocking AI Potential with AI-Ready Data Products	Nexla	https://nexla.com/blog/unlocking-ai-potential-with-ai-ready-data-products/
What is AI-Ready data?	Snowplow	https://snowplow.io/blog/what-is-ai-ready-data
Is your customer data AI-ready?	Deloitte	https://www.deloittedigital.com/us/en/insights/perspective/ai-ready-data.html

AI Starts with Data: Are you Ready?	The Modern Data Company	https://cdn.prod.website-files.com/66b054177abf98df89ee6edc/68db7a3f41843b1e0c07571e_AI%20Readiness%20Scorecard%20Whitepaper.pdf
Is Your Data AI Ready? Are You?	Snowflake	https://www.snowflake.com/en/blog/ai-ready-data-enterprise/
Data-Centric AI / AI-Ready Data Foundation	Accenture	https://www.accenture.com/content/dam/accenture/final/accenture-com/document-3/Accenture-Is-Your-Data-Ready-To-Power-AI-Reinvention-Report.pdf
Data for AI Operating Model	McKinsey (QuantumBlack)	https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-data-driven-enterprise-of-2025
TACO Framework (Tasker, Automator, Collaborator, Orchestrator)	KPMG	https://kpmg.com/xx/en/what-we-do/services/ai/agentic-ai-advantage.html
Data 4.0 / AI-Ready Data Flywheel	EY	https://www.ey.com/content/dam/ey-unified-site/ey-com/en-in/insights/ai/documents/ey-data-4-0-making-your-data-ai-ready.pdf
AI Adoption Framework (Data Pillar)	Google Cloud	https://cloud.google.com/resources/cloud-ai-adoption-framework-whitepaper
Cloud Adoption Framework for AI	Microsoft Azure	https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/scenarios/ai/
AI-Ready Data Framework	Informatica	https://www.informatica.com/blogs/how-to-get-your-data-ai-ready.html
Data Products for AI / Active Data Governance	Alation	https://www.alation.com/resource-center/whitepapers/data-ai-readiness-strategy-guide
AI-Ready Data Stack / Active Metadata	Atlan	https://atlan.com/know/ai-readiness/ai-ready-data/
Five steps to ensure your data is AI-Ready*	Gartner	https://www.gartner.com/en/documents/5848047
A Practical Framework for the AI-Ready Enterprise*	Boomi	https://boomi.com/content/ebook/ai-ready-enterprise-framework/

Ensuring Data + AI Reliability Through Observability*	MonteCarlo Data and O'Reilly Data	https://info.montecarlodata.com/get-resources/oreilly-report-ensuring-data-and-ai-reliability-through-observability
IDC Spotlight: Increasing AI Adoption with AI-Ready Data*	IBM	https://www.ibm.com/forms/mkt-53321
E-book: The Enterprise Guide to AI-Ready Modernization on Azure*	Rackspace Technology and Microsoft	https://www.rackspace.com/lp/enterprise-guide-ai-ready-modernization-azure
Unlocking enterprise AI*	Databricks	https://www.databricks.com/resources/analyst-research/unlocking-enterprise-ai-opportunities-and-strategies
7 Steps to Build AI-Ready Data Infrastructure*	Actian	https://www.actian.com/blog/data-governance/7-steps-to-build-ai-ready-data-infrastructure/
AI data readiness*	Deloitte	https://www.deloitte.com/content/dam/assets-zone3/us/en/docs/services/risk-advisory/2024/us-advisory-ai-data-readiness.pdf
Before you invest in AI, assess your AI-readiness*	Hyland	https://www.hyland.com/en/resources/articles/ai-readiness

APPENDIX II

Table II.2 shows our original longlist of 10 criteria for AI-ready data and outlines how each one was refined over the course of interviews into our final list.

Longlist of requirements found in comparative analysis	Final list of criteria for AI-ready enterprise data
Low latency data access	Removed – this criterion is not relevant for all AI applications, only some that require real-time data
Continuous iteration via an AI 'flywheel'	An AI-data feedback loop (Governance)
Unified and multimodal storage	Removed – simultaneity with other framework criteria
Robust data pipelines	Removed – generic and heterogeneous
Observability, monitoring, and quality enforcement	Monitoring and observability checkpoints (Infrastructure)
Dynamic access control	Agent-aware access controls via policy-as-code (Governance)
Multi-jurisdictional compliance	Combined with Agent-aware access controls via policy-as-code (Governance)
Enterprise-level metadata	Adapted to Semantic interlinkage (Infrastructure)
Active cataloguing	Active Cataloguing (Infrastructure)
Master data management (MDM)	Master data management (Infrastructure)

APPENDIX III

We thank our interviewees, who we selected and invited to build a representative sample of different businesses, large and small, that are exploring the ways in which data can enable AI.

- Pascal Weis, Global Head of IT at NTT Data Inc.
- Anonymous, business consultant at a large defence organisation
- Anonymous, data architect at a large defence organisation
- Anonymous, data engineer at a small RegTech organisation
- Anonymous, data engineer at a university
- Yashwanth Chandaka, Director of Enterprise Data Architecture, Platform and Engineering at The Clorox Company
- Javier Caycho, Operational Data Foundation & Integration Head at ZEISS Group
- Oli Bage, Advisor at the Open Data Institute (previously Head of Architecture, Data & Analytics at London Stock Exchange Group)