ENTERPRISE AI PLAYBOOK · DEEP DIVE · PART 6 · DATA

The Data & Knowledge Foundation

The one advantage a competitor can't buy — and the layer most likely to be quietly immature. Five pillars for data that's actually ready to serve AI.

CROSS-CUTTING CAPABILITY · RUNS THROUGH EVERY STAGE

- **Sourcing & Ingestion — getting the right data in**
- **Quality & Curation — garbage in, hallucinations out**
- **Governance & Rights — who may use what, lawfully**
- **Retrieval & Serving — getting it to the model**
- **Knowledge Structure — meaning, not just bytes**

Tom Ward
Enterprise Architect · Cloud & AI

linkedin.com/in/tomward
July 2026

PILLAR 1 Sourcing & Ingestion

"Getting the right data in"

Every AI capability is downstream of the data that feeds it, and most enterprises discover too late that the data they need is trapped — in a system with no API, a vendor that owns it, a format nobody can parse. Sourcing is the unglamorous work of getting the right data in: mapping where it lives, building the connectors and pipelines to reach it, and being honest about coverage gaps before they become production surprises.

The architectural decision that matters most is batch versus streaming, and it's driven by the use case's freshness need, not by fashion. A quarterly-trends assistant can run on nightly batch; a fraud-detection agent cannot. Match the pipeline to the staleness the workload can tolerate.

KEY DECISIONS

- Which sources are authoritative, and which are convenient but unreliable?
- Batch or streaming — set by the freshness the use case actually needs.
- Where are the coverage gaps, and do they undermine the use case?

WATCH OUT FOR

- × Data you assumed was reachable but has no API or export path.
- × Building on a convenient copy that drifts from the system of record.
- × Discovering coverage gaps only after the model ships.

PILLAR 2 Quality & Curation

"Garbage in, hallucinations out"

This is the pillar Gartner puts near the top of its failure list, and for good reason: poor-quality data doesn't announce itself — it produces confident, plausible, wrong answers. Unlike a traditional analytics pipeline, where bad data yields an obviously broken report, bad data in a GenAI system yields a fluent hallucination that a user may act on. Quality is not a nice-to-have; it is the difference between a helpful assistant and a liability.

Curation is continuous, not a one-time cleanse. Accuracy, completeness, de-duplication, and enrichment have to be maintained as sources change — and for GenAI specifically, the chunking and labeling that prepare data for retrieval are quality decisions in their own right.

KEY DECISIONS

- What quality bar does each use case require, and who measures it?
- How is quality monitored continuously, not audited once?
- Who owns curation — a data team, or the source-system owners?

WATCH OUT FOR

- × Demos on a hand-cleaned sample that hide the real data's mess.
- × Chunking by page count instead of meaning — garbage chunks in.
- × Quality treated as a launch gate, not an ongoing SLA.

PILLAR 3 Governance & Rights

"Who may use what, lawfully"

Data governance is where the Data capability and the Governance capability meet: the question of who is *allowed* to use which data, for what, is both a data problem and a compliance problem. The non-negotiable rule for AI is that retrieval must enforce the same entitlements as the source systems — if a user can't open a document, the model must not surface its contents on their behalf. Access control at query time, not just at indexing time, is what separates a compliant design from a breach.

Lineage and usage rights are the other half. You need to know where each piece of data came from (to explain and audit a decision) and whether you're contractually and legally permitted to use it for AI at all — a question that trips up organizations who trained or grounded on data they only had rights to use for something else.

KEY DECISIONS

- Are source-system entitlements enforced at retrieval time?
- Is lineage captured well enough to explain and audit a decision?
- Do usage rights and privacy law permit this data for this AI use?

WATCH OUT FOR

- × Copying entitled data into an unsecured index — the classic audit finding.
- × No lineage, so you can't explain why the model said what it said.
- × Using data for AI that you only had rights to use for analytics.

PILLAR 4 Retrieval & Serving

"Getting it to the model"

Having good, governed data is necessary but not sufficient — it has to reach the model at the moment of inference, fast and relevant. This is where most of the real engineering in a grounded AI system lives: embeddings, vector and hybrid search, re-ranking, and the serving infrastructure that returns the right context within the latency the experience can tolerate. Retrieval quality, more than model choice, often determines whether an answer is right.

Freshness is the quiet killer. A retrieval index built once and never refreshed slowly diverges from reality, and the model's answers rot with it. Treat the refresh cadence as a first-class SLA tied to how fast the underlying data changes.

KEY DECISIONS

- Vector, keyword, or hybrid retrieval — and what re-ranking strategy?
- What is the freshness SLA, and how is the index refreshed?
- What retrieval latency does the experience layer actually allow?

WATCH OUT FOR

- × Indexing once and never refreshing — stale retrieval poisons quality.
- × Optimizing the model while ignoring weak retrieval.
- × No evaluation of retrieval quality, only of the final answer.

PILLAR 5

Knowledge Structure

"Meaning, not just bytes"

The most mature data foundations don't just store and retrieve data — they capture *meaning*. Metadata, ontologies, and knowledge graphs turn a pile of documents into a connected model of your business: which entity relates to which, what a term means in your specific context, how a policy links to a product links to a customer. For AI, this structure is what lets retrieval move beyond keyword-similar text to genuinely relevant, reasoned context.

This is also the most-skipped pillar, because it's the hardest and its payoff is longest-term. You don't need a full enterprise ontology to ship a first use case — but the organizations pulling durably ahead with AI are the ones investing in a semantic layer that every future use case can draw on, instead of re-solving "what does this term mean" for each one.

KEY DECISIONS

- Where does a knowledge graph or ontology earn its cost vs. plain retrieval?
- Is there a shared semantic layer, or does each use case redefine terms?
- Who owns the metadata and keeps it alive as the business changes?

WATCH OUT FOR

- × Boiling the ocean on an enterprise ontology before shipping anything.
- × Every team redefining the same entities and terms from scratch.
- × Metadata created once at a project's start and never maintained.

Data at every stage

Data isn't a phase you finish — it's the capability the whole lifecycle stands on. It shows up, differently, at each of the four stages.

Assess · Part 1	The Data Platform is a layer you score — is it serving-ready, or analytics-ready?
Decide · Part 2	"Is the data for this use case ready?" is the gate that stops the wrong initiatives before they start.
Build · Part 3	RAG, grounding, and access-trimmed retrieval are the Data foundation in action.
Advance · Part 4	Agents inherit exactly the data access and quality you gave them — for better or worse.

The AI-Ready Data Test

"Is our data ready?" is the wrong question — it's unanswerable at enterprise scale and it stalls everything. Ask it per use case instead. For a specific initiative, data is AI-ready only when all of these hold; each "no" is a scoped, fundable task, not a reason to quit.

- ☐ **Reachable** — the data this use case needs is accessible through a real, supported path, not a one-off export.
- ☐ **Sufficient** — coverage and history are complete enough that the model isn't guessing in the gaps.
- ☐ **Trustworthy** — quality is measured and monitored, not assumed from a clean demo sample.
- ☐ **Governed** — retrieval enforces source-system entitlements at query time, with lineage captured.
- ☐ **Lawful** — you have the usage rights and privacy clearance to use this data for this AI purpose.
- ☐ **Fresh** — the refresh cadence matches how fast the underlying data actually changes.
- ☐ **Retrievable** — the right context comes back fast enough for the experience, and retrieval quality is evaluated on its own.
- ☐ **Meaningful** — where the use case needs it, a semantic layer or knowledge structure gives the data context, not just similarity.

The organizations winning with AI aren't the ones that "finished" their data — no one does. They're the ones who reframed an impossible enterprise-wide question into a scoped, per-use-case test, and funded the specific gap in front of them.

Going Deeper: References

Sources spanning data management, platform, retrieval, and quality.

DAMA-DMBOK — Data Management Body of Knowledge
The reference framework for data management disciplines — governance, quality, metadata, and more.
dama.org/cpages/body-of-knowledge

Microsoft — Azure Data Architecture Guide
Concrete patterns for data platforms, pipelines, and stores — translates well beyond Azure.
learn.microsoft.com/en-us/azure/architecture/data-guide

Databricks — The Data Lakehouse
The lakehouse pattern for unifying analytics and AI data on one governed platform.
databricks.com/glossary/data-lakehouse

Google Cloud — Retrieval-Augmented Generation
A clear primer on RAG — the mechanism that gets your governed data to the model.
cloud.google.com/use-cases/retrieval-augmented-generation

Martin Fowler — Data Mesh Principles
A vendor-neutral take on data ownership and domain-oriented architecture at scale.
martinfowler.com/articles/data-mesh-principles.html

Data Observability — an introduction
How to monitor data quality and freshness continuously, the way you monitor uptime.
montecarlodata.com/blog-what-is-data-observability

NIST AI Risk Management Framework
The governance vocabulary for the data-rights and privacy questions in Pillar 3.
nist.gov/itl/ai-risk-management-framework

About the Author

Tom Ward is an Enterprise Architect with 20+ years across Fortune 500 financial services, retail, and government — including Wells Fargo and Bank of America — specializing in EA governance, reference architectures, hybrid cloud, and legacy modernization, with recent hands-on GenAI and mobile delivery work.

The full series: archreviews.com/playbook · Connect: linkedin.com/in/tomward