

U201 UNDERSTAND · PRACTITIONER

AI-Ready Data

Preparing enterprise data for retrieval, agents and models: fitness, context, lineage, permission and access.

FORMAT

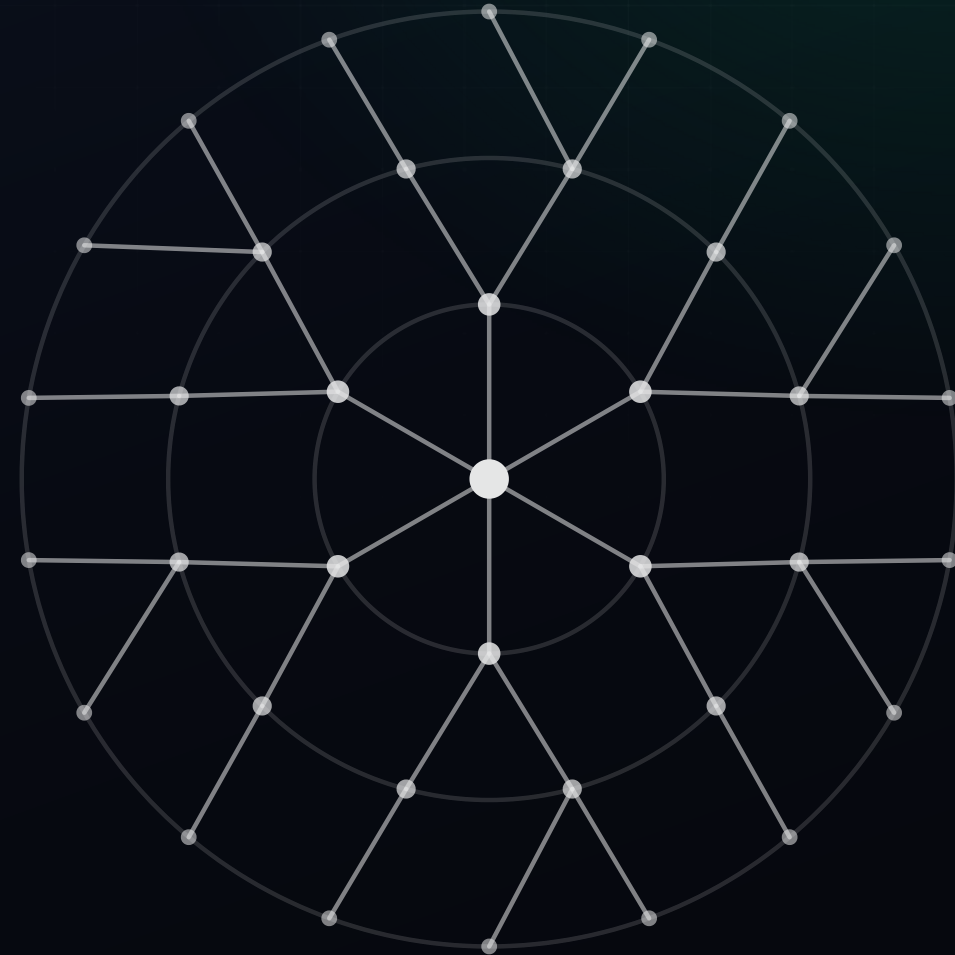
Knowledge pack · 12 slides

FOR

Data and AI leaders

LEVEL

Practitioner



THE PREMISE

Data is ready for a use case, not in general

The same customer table can be adequate for a dashboard and inadequate for an agent that issues refunds. Readiness depends on **what the AI will do**, how much error is tolerable, and who may see the result.

Decision

The action the output will drive.

Tolerance

The cost of a wrong answer.

Audience

Who is entitled to see what.

WHY IT MATTERS

Lack of AI-ready data puts AI projects at risk

60%

of AI projects unsupported by AI-ready data, Gartner predicts, will be abandoned through 2026.

Gartner prediction, February 2025

63%

of organizations either did not have, or were unsure whether they had, the right data management practices for AI.

Gartner survey, reported February 2025

Use case

In Gartner's definition, AI-ready data "must be representative of the use case, of every pattern, errors, outliers and unexpected emergence that is needed to train or run the AI model for the specific use."

Gartner, February 2025

ASSESSMENT

Five dimensions of readiness

Fitness

Accurate, complete, timely and representative enough for the decision it supports.

Context

Definitions, units and business meaning a model can use.

Lineage

Where the data came from and how it was transformed.

Permission

Who may use it, for which purpose, under which obligations.

Access

Programmatic access at suitable latency, cost and reliability.

HOW TO USE IT

Score each dimension against the use case. The weakest dimension sets the limit on what the AI may do.

TWO KINDS OF DATA

Structured and unstructured data fail differently

	STRUCTURED DATA	UNSTRUCTURED CONTENT
Typical failure	Inconsistent definitions, missing keys, stale extracts	Duplicates, outdated versions, drafts and AI output mixed with approved policy
Hidden risk	Joins that silently drop or duplicate records	Sensitive data embedded in documents and tickets
Missing metadata	Business definitions and units	Owner, version, effective date and permissions
Remedy	Semantic layer, data contracts, quality checks	Curation, versioning, permission-aware indexing

RETRIEVAL

Retrieval is a pipeline, and every stage can fail

Ingest

Parse sources with structure intact; capture owner, version, effective date and permissions.

Index

Combine vector and keyword indexes; add a graph index for corpus-wide questions.

Retrieve

Hybrid search and re-ranking, pre-filtered to what both the user and the agent may see.

Generate

Answer only from retrieved sources, cite them, or decline to answer.

Evaluate

Test retrieval recall, then groundedness and completeness, on real questions.

THE RULE

Filter by permission at retrieval. Never rely on a model to withhold what it has already been given.

FAILURE MODES

Why retrieval answers go wrong

✗ Stale documents outrank current ones

✓ Index effective dates; exclude superseded versions

✗ Chunking breaks tables and procedures

✓ Split on headings; keep tables and steps whole

✗ Duplicates crowd out the right answer

✓ Deduplicate and mark the authoritative version

✗ Policies from other regions are mixed in

✓ Filter by jurisdiction and business unit

✗ Users see content they are not entitled to

✓ Enforce current source permissions; fix oversharing

✗ Answer quality is not measured

✓ Rerun a known-answer evaluation set after every change

DATA PRODUCTS

Data products give agents a contract they can rely on

Publish the terms below as a machine-readable data contract, such as the Open Data Contract Standard, so pipelines and agents can check schema, quality and service levels before use.

Owner

A domain team accountable for quality and change.

Schema and meaning

Fields, definitions and units, published and versioned.

Quality objectives

Freshness, completeness and accuracy targets, monitored.

Access policy

Who may use it, and for which purposes.

Lineage

Upstream sources and transformations, captured automatically.

Change process

Versioning and deprecation consumers can plan for.

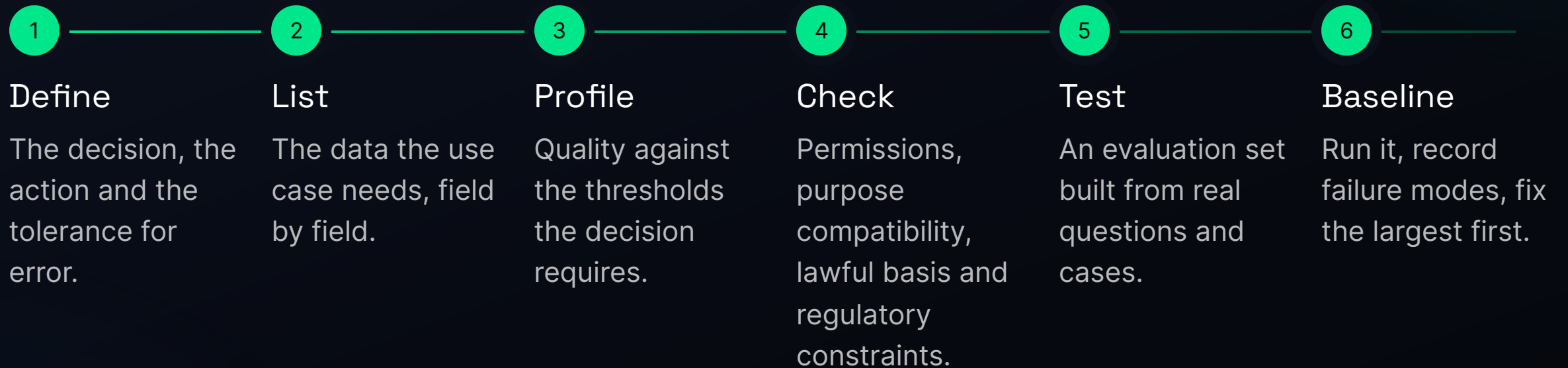
GOVERNANCE

Govern the data that AI uses

- ✔ **Classify**
Label sensitive and regulated data before it is indexed.
- ✔ **Propagate deletion**
Deletion and retention must reach indexes, embeddings, caches, logs and fine-tuning data.
- ✔ **Control providers**
Confirm whether providers train on your data, how long they retain it, and where.
- ✔ **Review access**
Include agent identities and indexing connectors in every access review.
- ✔ **Limit purpose**
Record the purposes for which each data set may be used.
- ✔ **Log**
Keep prompts, sources and outputs for audit; restrict and expire logs as the sources require.
- ✔ **Minimize**
Retrieve only the fields a task needs.
- ✔ **Monitor**
Alert on unusual retrieval volumes and patterns.

METHOD

Assess one use case in a two-week time box



MEASUREMENT

Readiness measures for each use case

Freshness

Lag between a change at source and its availability, against the requirement.

Critical completeness

Share of records with every field the use case needs.

Definition coverage

Share of fields with an agreed business definition.

Permission coverage

Share of indexed documents whose permissions match the source.

Groundedness

Share of answers fully supported by the retrieved sources.

Exposure incidents

Count of answers that exposed content the user was not entitled to see.

SUMMARY

Key takeaways

- 01 Assess readiness against a specific use case, its decision and its tolerance for error.
- 02 Check five dimensions: fitness, context, lineage, permission and access.
- 03 Treat retrieval as a pipeline and evaluate every stage with real questions.
- 04 Enforce permissions at retrieval, and propagate deletion to indexes, embeddings and caches.
- 05 Publish data products with contracts for schema, quality, access and change.

NEXT COURSE · T201

Application Rationalization: From Evidence to Switch-Off

Transform & Optimize · Practitioner