

Data governance for AI ready UAE enterprises

A practical control model from source systems to AI retrieval

AI readiness is a property of a specific data product and its intended use. A clean warehouse or searchable catalogue alone cannot establish that a dataset is permitted, reliable and safe for an AI workflow.

This paper gives UAE enterprise data leaders, architects, security teams and privacy specialists a practical way to turn governance decisions into testable controls. Start with one bounded business use case, assign accountable owners, and release data only when its meaning, quality, access rules and lifecycle obligations can be demonstrated.

Five decisions that make data usable

- **Establish the applicable rules.** Map the legal entity, processing purpose, data categories, recipients and jurisdictions before selecting a deployment region.
- **Give each product an accountable owner.** The owner approves permitted use and risk exceptions; a steward maintains meaning and quality; custodians implement and prove controls.
- **Bind metadata to enforcement.** Carry source identifiers, versions, classification and access policies through tables, documents, chunks and embeddings.
- **Measure fitness for the actual task.** Test important fields and bilingual retrieval behaviour against explicit acceptance criteria rather than a single estate-wide quality score.
- **Prove change and deletion propagation.** An access revocation or source deletion must reach derived copies, search indexes and caches, with evidence and an exception path.

How to use this paper

Use pages 2 to 6 as design inputs, then adopt the release checklist and 90-day plan on page 7. The architecture, thresholds and contract are illustrative engineering recommendations, not a legal opinion or a statement of any vendor's implemented capability.

Scope note. Requirements vary by entity, sector, free zone, processing activity and contract. Legal counsel and the relevant regulator should validate applicability, current implementing measures and cross-border conditions before deployment. Sources were reviewed for this edition on 1 October 2026.

Turn obligations into placement decisions

The federal Personal Data Protection Law is Federal Decree by Law No 45 of 2021. Article 2 includes exclusions for government data and government entities, specified security and judicial data, personal use, health and banking or credit data governed by specific legislation, and free-zone entities governed by special data-protection legislation. Articles 4 and 5 address consent exceptions and processing controls; Articles 22 and 23 provide cross-border transfer routes subject to conditions. They do not create a blanket rule that all UAE data must remain in the UAE. [1]

The UAE Data Office is the federal data regulator described by the government portal. DIFC has its own Data Protection Law No 5 of 2020; ADGM has its Data Protection Regulations 2021 and transfer guidance. An address in Dubai or Abu Dhabi alone is not enough to determine the regime. Map the relevant entities and activities, including any overlapping sector obligations. [2, 3, 4]

Record one decision for each processing flow

Document the purpose and legal basis or other applicable authority; the controller, processors and subprocessors; the categories and sensitivity of data; the source and destination jurisdictions; and the evidence supporting the decision. Identify sector, free-zone, government and contractual restrictions separately. Have counsel resolve uncertain scope and confirm current rules rather than treating this checklist as legal clearance.

Decision	What the engineering team records
Classification	Public, internal, confidential or restricted; personal-data and sector tags; a named owner; field-level exceptions. These are illustrative enterprise labels, not statutory categories.
Access	Allowed role, business purpose, tenant or entity, row and field scope, approval expiry, service identity and emergency-access procedure.
Retention	A rule by record category: purpose, trigger event, period, legal hold, deletion method, backup expiry and accountable approver. No universal retention period is assumed.
Placement and transfer	Storage, processing, model endpoint, disaster recovery, logs, telemetry and support access; approved countries; transfer mechanism and vendor terms where required.

Use a release gate rather than a location slogan

If applicable restrictions or the transfer route remain unresolved, keep the affected flow blocked. If approved, configure the actual service topology and test it: a regional database does not by itself establish the location of embeddings, inference, logs, backups or overseas support access.

Reassess when purpose, recipient, region, provider, subprocessor, data category or access changes. Record each exception's approver, compensating control and expiry; require reapproval rather than indefinite waivers.

Design a governed lifecycle

Separate the systems that hold and process data from the governance services that describe and control it. The following reference architecture makes release and withdrawal explicit. It is a design pattern, not a prescribed technology stack.

GOVERNANCE SERVICES

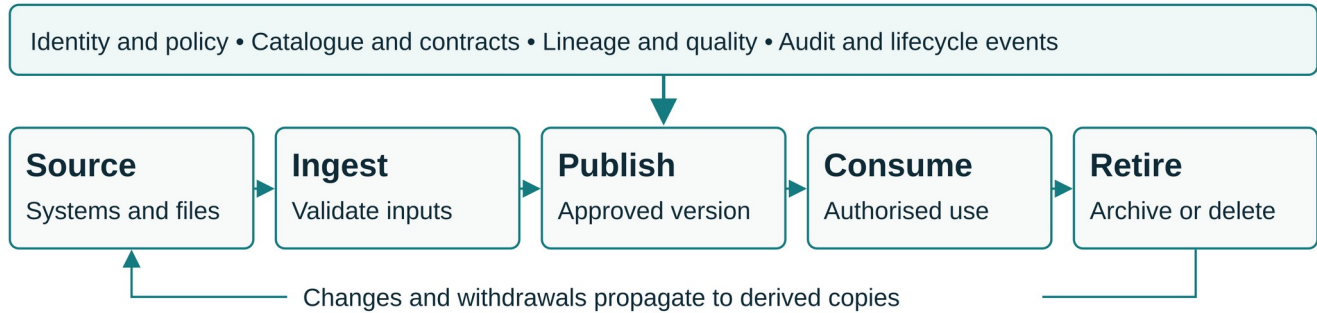


Figure 1 A data product carries policy and provenance through its lifecycle

Keep the catalogue honest

Register a stable product ID, business definition, owner and steward, authoritative source, schema version, sensitivity, approved purposes, access route, retention rule and last successful quality check. For documents, add language, effective date, document version, rights and canonical source link. Keep sensitive sample values out of metadata unless explicitly authorised.

Microsoft's Purview documentation makes a useful distinction: Data Map and Unified Catalog hold metadata, and their roles do not themselves grant access to the underlying data. Treat catalogue discovery and runtime authorisation as separate controls, regardless of platform. [6]

Capture lineage at the grain of the risk

For a critical report, trace output fields to source fields and transformation versions. For retrieval, trace each chunk and embedding to a document ID, version, section or page, parser version and index build. Store pipeline run IDs and deployment hashes so an operator can reconstruct what changed.

Automated lineage coverage varies by connector and operation. Microsoft documents this explicitly, including manual and API-supported lineage paths. Inventory coverage gaps and label manually supplied lineage rather than representing the graph as complete. [7]

Make lifecycle events actionable

An approved change should trigger compatibility checks and notify registered consumers. Quarantine failed loads before publication. Archive only where the retention policy calls for it; a legal hold suspends the relevant deletion action but does not authorise new AI uses. A withdrawal event should identify all downstream copies and record which systems acknowledged removal.

Test the reverse path. Select one source record or document and prove that its derived versions, consumers and deletion obligations can be found without relying on a single engineer's memory.

Assign decisions and measure quality

Federate responsibility to business domains while keeping enterprise policy, identity and assurance consistent. Name people or durable accountable roles for every product. “The data team” is not a sufficient owner.

Activity	Owner	Steward	Custodian	Privacy or risk	Consumer
Approve purpose and classification	A	R	C	C	C
Maintain definitions and contract	A	R	C	C	C
Implement access and lifecycle rules	A	C	R	C	I
Run quality checks and resolve defects	A	R	R	C	C
Review legal and transfer assessment	C	C	C	A/R	I
Approve release and residual risk	A	R	R	C	C

R = responsible for delivery; A = accountable decision maker; C = consulted; I = informed. This illustrative RACI does not transfer a controller's legal responsibility to an internal role. One named owner makes each release decision; specialist review remains mandatory where applicable.

Use a small set of interpretable measures

Dimension	Example measure	Response to a failure
Completeness	Required non-null values ÷ eligible records, by field and source	Block missing identifiers; quarantine optional-field failures if agreed.
Validity	Values satisfying the declared type, range and format ÷ tested values	Reject invalid timestamps, currency codes or unexpected enums.
Uniqueness	Duplicate primary keys per load, using the declared record grain	Stop publication and reconcile retries or source duplication.
Consistency	Conflicting values across defined authoritative relationships	Resolve precedence rules; do not silently overwrite.
Timeliness	Current time minus source event time and last successful load	Display freshness; suspend time-sensitive use beyond the limit.
Accuracy	Agreement with independently verified reference samples	Review representative samples; format validity is not accuracy.

Set thresholds per critical element and consequence. Report the numerator, denominator, sample method, time window and exclusions. A 99% aggregate can hide a serious failure in one tenant or Arabic-language source. Track rejection volume and downstream impact as well as pass rates.

For bilingual data, preserve the original text and language tag. Test OCR, right-to-left layout extraction, numerals, dates and entity matching separately. Normalisation must not erase legally or commercially meaningful differences. These are proposed tests, not a claim that English and Arabic quality is equivalent.

Review serious defects weekly. Each exception needs an impact assessment, remediation owner, expiry and consumer notice.

Publish an explicit data product contract

A data product contract records what a producer provides and what consumers may rely on. Store it under version control, validate it in delivery pipelines and link it to the catalogue. It complements legal agreements and access controls; it does not replace them.

Illustrative contract for a policy knowledge product

The following fictional example is for an internal bilingual policy assistant. Names, targets and time limits are design assumptions, not CodeX71 service commitments or recommended legal retention periods.

Contract field	Illustrative value or rule
Identity and ownership	Product: corporate-policy-knowledge; version: 1.0.0; owner: policy governance lead; steward: knowledge operations; custodian: platform engineering.
Purpose and exclusions	Answer employee questions from approved current policy documents. Exclude individual HR cases, personal records, drafts and personalised legal advice. No model training or external sharing without separate approval.
Record grain and schema	One approved document version. Required: document_id, version_id, language, effective_from, owner_id, classification, source_uri, content_hash and policy_ref. Chunk records also carry chunk_id, parent_version_id and location.
Semantics and language	Arabic and English documents remain separately versioned with a translation relationship. The policy owner identifies the authoritative text when versions conflict; the assistant surfaces that conflict.
Access and placement	Employee role and legal-entity entitlement required. Apply the source policy to every chunk. Approved region and endpoints follow the signed placement decision; query-time authorisation is required.
Quality and freshness	Zero missing required identifiers; zero duplicate document-version IDs; 100% of published chunks linked to a parent and current policy reference. Proposed freshness target: approved source changes indexed within four hours.
Retention and withdrawal	Derived content expires when the source is withdrawn or its approved retention ends. Illustrative target: deny retrieval on authoritative revocation; invalidate entitlement caches within 15 minutes; remove active derived copies within 24 hours. Backup treatment follows the approved schedule and restore-suppression procedure.
Compatibility and operations	Breaking schema, meaning or policy changes require a major version, consumer approval and a migration window. Producer publishes status, incident contact, run ID and rollback target.

Enforce the contract at three boundaries

At ingestion, validate identity, schema and required policy metadata; quarantine missing or invalid records. At publication, run quality, lineage and access tests before promoting a version. At consumption, resolve the current contract and user entitlement; refuse stale, withdrawn or unauthorised content.

Test a revoked user, unknown classification, withdrawn document, incompatible schema and failed refresh. Keep results with the release manifest. A fallback version remains usable only while its purpose, freshness and permissions are valid.

Carry governance into AI retrieval

Retrieval augmented generation creates derived assets and disclosure paths. Govern indexes, rerankers, prompts, traces and answer caches as part of the data product.

Apply entitlements before content reaches the model

Authenticate the end user, derive tenant and purpose context, and evaluate a trusted policy decision. Security filters must constrain candidate retrieval before unauthorised text reaches reranking or generation. Recheck returned candidates against the current policy before prompt construction. An application prompt saying “do not reveal restricted content” is not an access control.

Carry IDs, tenant scope, classification and policy references through indexing. A privileged service account must not broaden the user’s access. Fail closed on absent or stale entitlements; rerun deny tests after changes.

Separate permitted retrieval from other AI uses

Source access does not automatically approve training, evaluation reuse or disclosure to new providers. Assess each purpose and recipient. Review model endpoints, retention and training settings, subprocessors, logging and contracts before sending data.

Treat retrieved text as untrusted evidence. Embedded instructions cannot override permissions or policy. Preserve source citations, flag evidence conflicts and route consequential decisions to the appropriate human process.

Test deletion as a distributed workflow

Step	Expected evidence
1 Locate	Source ID resolves to versions, chunks, vectors, caches, exports and registered consumers through a dependency manifest.
2 Deny	A tombstone or policy change blocks new retrieval immediately; entitlement caches are invalidated within the approved service objective.
3 Remove	Idempotent deletion jobs remove active copies and confirm each destination; failed acknowledgements create owned incidents.
4 Preserve only when justified	Legal holds and required records are isolated with restricted access. Backups expire on schedule; restores reapply deletion and revocation records before use.
5 Verify	Negative retrieval tests and an inventory reconciliation confirm removal. Evidence records IDs and outcomes without retaining the deleted content unnecessarily.

Do not presume embeddings are anonymous. Vector deletion does not erase sources, caches or trained models. Assess training separately and do not promise erasure from model weights without a validated method.

Operational measures. Track unauthorised retrieval test failures, policy-change propagation time, orphaned chunks, overdue deletion acknowledgements and stale-source answers. Define escalation thresholds and an emergency route to disable affected retrieval.

Release deliberately and improve over 90 days

A practical governance release checklist

Require an evidence link and a named reviewer for each item. A missing control may be resolved or formally excepted within policy; an unresolved legal basis, unauthorised transfer or failed access-control test blocks release.

Control	Evidence to retain
Purpose and accountability	Approved use, exclusions, named owner, steward and custodian; applicability and risk assessment.
Inventory and contract	Versioned contract; classifications; source-to-product lineage; identified coverage gaps and consumers.
Access and privacy	Least-privilege policies; positive and negative entitlement tests; minimisation and processor review.
Quality and availability	Critical-field checks, freshness target, language-specific tests, known limits, quarantine and rollback evidence.
Lifecycle and response	Retention and hold rules; end-to-end withdrawal test; backup restore test; incident routing and notification assessment.

Build one governed path before scaling the estate

Period	Work and owner	Exit evidence
Days 1 to 30	Business owner selects one bounded use case. Steward inventories critical sources. Privacy and security teams map rules, risks, recipients and placement.	Signed scope and ownership; baseline quality and access findings; prioritised gap list; approved target controls.
Days 31 to 60	Custodians implement contracts, policy propagation, quality gates and lineage. Stewards resolve definitions and test Arabic and English sources separately.	Repeatable ingestion and quarantine; permission-denial tests; current catalogue entries; an observable pilot with no unresolved blocking risks.
Days 61 to 90	Owner runs a controlled release. Operations drills withdrawal, incident response and recovery. Consumers validate fitness and known limitations.	Release record; deletion and restore evidence; service objectives and support ownership; scale, revise or stop decision.

The sequence is illustrative and depends on access, stakeholder availability, legal review and technical debt. Failed gates should change the scope or timeline, not lower the evidence standard.

Use a risk framework to sustain the work

NIST’s voluntary Privacy Framework 1.0 offers five functions: Identify-P, Govern-P, Control-P, Communicate-P and Protect-P. Use a current and target profile to expose gaps and assign improvement work. The framework is a risk-management aid, not a UAE legal determination or a compliance certification. [5]

Sources and next steps

This paper combines primary-source references with CodeX71's proposed implementation patterns. The examples are illustrative; no customer results, production validation or certifications are asserted. Official materials and service features can change. Recheck relevant sources before a decision.

1 Federal Decree by Law No 45 of 2021 concerning the Protection of Personal Data

UAE Legislation. Articles 2, 4, 5, 22 and 23 support the scope, processing and transfer discussion. English text is a reference; verify the authoritative Arabic text and current implementing measures.

[Official source](#)

2 Data protection laws

UAE Government Portal. Federal data-protection overview and UAE Data Office responsibilities.

[Official source](#)

3 Commissioner of Data Protection

DIFC. Official overview of DIFC Law No 5 of 2020 and the Commissioner's role.

[Official source](#)

4 Data protection guidance

ADGM Office of Data Protection. Guides cover scope, processing, rights, accountability and international transfers under the 2021 Regulations.

[Official source](#)

5 NIST Privacy Framework Version 1 0

NIST, 16 January 2020. The voluntary risk framework and its five functions. This paper uses the published version 1.0.

[Official source](#)

6 Data governance with Microsoft Purview

Microsoft Learn. Distinguishes Data Map and Unified Catalog metadata from underlying data and runtime access.

[Official source](#)

7 Data lineage user guide for classic Data Catalog

Microsoft Learn. Connector-dependent scope, manual lineage and known limitations. Verify current support for the product and connector actually deployed.

[Official source](#)

Scope a Data Readiness Sprint

A scoped discovery conversation can define one priority use case, the required sources, an ownership model and the tests needed to establish readiness. Agree deliverables, assumptions and success criteria before implementation.

Falaj is a development-stage concept for a governed bilingual data layer. Its inclusion here is directional; the architecture and controls in this paper do not represent validated Falaj capabilities or availability.

[Discuss a scoped Data Readiness Sprint](#)