**UAE Legal GraphRAG — Technical Proposal (EY)**

**0) Executive summary**

We propose an **Azure-native GraphRAG platform**—purpose-built for UAE law—that delivers **as-of correct answers**, **explainable paths**, and **auditable provenance** across federal, emirate, and free-zone sources (DIFC/ADGM). The core is **Azure AI Search** (hybrid lexical+vector with semantic rerank) for text, **Azure Cosmos DB (Gremlin API)** for graph structure, and **Azure OpenAI** for reasoning—wrapped with **function-called, parameterized Gremlin tools** and **groundedness checks** before any answer is shown. For demo velocity and analyst UX, we also include an **optional, read-only Neo4j mirror** (Text2Cypher + Bloom + GDS). This design aligns with UAE legal publication/entry-into-force rules and official sources.

**1) Objectives & success criteria**

* **Correct “as-of” law**: reflect **Official Gazette** publication/entry-into-force rules (Constitution arts. 111–112). Answers must be time-scoped and cite the Gazette.
* **Whole-system coverage**: **UAE Legislation** (federal), emirate gazettes/portals, and **free-zone** (DIFC/ADGM) judgments.
* **Explainability**: every claim backed by **paths** (e.g., Provision ←AMENDS/REPEALS—Event→Gazette).
* **Operationally safe**: private networking (Private Link), RBAC, function-called graph access, and **groundedness detection/correction** before display.
* **Cost-disciplined**: clear dials (RU/s, SU/h, $/1k semantic rerank, tokens) and reserved-capacity options.

**2) Scope & authoritative sources**

* **UAE Legislation Platform** (federal laws, implementing regs, amendments, sectors).
* **Official Gazette** (publication & entry-into-force text; “one month unless otherwise specified”).
* **DIFC & ADGM Courts** (judgments, common-law reasoning within their jurisdictions).

**3) Why GraphRAG (and which flavor)**

We adopt Microsoft’s **Local / Global / DRIFT** pattern:

* **Local** for precise, entity-centric hops;
* **Global** for corpus-wide, community-level summaries;
* **DRIFT** to aim local hops using global structure (better recall at lower cost).

On Azure, we will run GraphRAG concepts via **AI Search** + **Gremlin tools**, or optionally deploy Microsoft’s **GraphRAG Accelerator** as a managed service scaffold (for PoCs).

**4) Target architecture (Azure core, optional Neo4j mirror)**

**Azure core (production)**

* **Azure AI Search** — **hybrid** (BM25 + vectors) fused by **RRF**, with **semantic ranker** ($1 per 1k queries after 1k free) to raise precision.
* **Azure Cosmos DB (Gremlin API)** — authoritative knowledge graph; **Private Link** endpoints; regional failover; Gremlin traversals via **function calling** (tools).
* **Azure OpenAI** — reasoning/answering (regional availability per model), run **groundedness detection/correction** before display.
* **Governance & catalog** — **Microsoft Purview (Unified Catalog)** available in **UAE North** for lineage/classification.

**Optional mirror (analyst UX & demos)**

* **Neo4j AuraDB (read-only)** + **Bloom** for visual audit; **GDS** (Louvain etc.) for community analysis; **Text2Cypher** for demo NL→query.

**5) Graph model (UAE-specific, bi-temporal)**

**Key entities**: Instrument (type: Constitution, Federal Law, Decree-Law, Decree, Cabinet/Ministerial Resolution, Emirate Decision…), Provision (Article/Clause), Event/Statement (Amend/Repeal/Consolidation/Interpretation), GazetteIssue, Authority, Jurisdiction (UAE Federal, Abu Dhabi, Dubai, RAK… plus **DIFC/ADGM**), Court, Judgment, Topic/Sector.

**Key relations**: INSTRUMENT\_HAS\_PROVISION, PUBLISHED\_IN, ISSUED\_BY, APPLIES\_IN, AMENDS/REPEALS (via Event nodes), IMPLEMENTED\_BY/DELEGATES\_TO, CITES, INTERPRETED\_BY/RELIES\_ON, HAS\_TOPIC.

**Bi-temporal**: put **valid\_from/valid\_to** (real-world effect) and **tx\_from/tx\_to** (system knowledge) on **Event/Statement** nodes to answer *“what was in force on 2022-05-01?”* while keeping audit history; derive **validity** from Gazette text and instrument clauses (Constitution 111–112).

**6) Ingestion & indexing**

* **Crawl/ingest**: UAE Legislation portal (federal), emirate official gazettes/portals, **DIFC/ADGM** published judgments. Extract **article-level text** (Arabic/English) and **Gazette metadata** (no., date).
* **OCR**: **Azure AI Document Intelligence** with **Arabic support** for scans; keep PDFs & hashes for provenance.
* **Search index**: Azure AI Search with **hybrid** fields (BM25 + vector embeddings) and **semantic ranker** for reranking.
* **Graph build**: ETL to Cosmos Gremlin: normalize instruments, provisions, events, gazettes and jurisdictions; stamp bi-temporal props; index common lookups.

**7) Retrieval & orchestration**

1. **Text front-door**: AI Search (hybrid + semantic) finds high-signal provisions/judgments quickly.
2. **Structured hops (tools)**: the LLM **does not** emit free-form Gremlin. It **function-calls** whitelisted tools that run **parameterized Gremlin** (read-only, hop/time/row limits):
   * as\_of\_snapshot(article\_id, date) (filters by validity & jurisdiction),
   * amendments\_chain(provision\_id),
   * implementing\_measures(instrument\_id),
   * case\_interpretations(provision\_id, jurisdiction, since).

(This is safer than raw Text2Gremlin while preserving expressiveness.)

1. **GraphRAG modes**:
   * **Local** for precise paths;
   * **Global** for community-level synthesis;
   * **DRIFT** to use global community cues to “aim” local follow-ups.
2. **Groundedness**: pass the draft to **Azure AI Content Safety – groundedness detection/correction**; only show text that aligns to retrieved sources.

For a PoC, Microsoft’s GraphRAG Accelerator gives a working Azure deployment scaffold (APIs, logging); we treat it as a reference, not a product dependency.

**8) Security, residency & governance**

* **Private networking**: **Private Link/Private Endpoints** for Cosmos and AI Search; private DNS zones; no public exposure.
* **RBAC/Key Vault** for secrets; **App Insights** for traces; **Purview (UAE North)** for catalog/lineage.
* **Data residency**: deploy in **UAE North**; confirm model availability for **Azure OpenAI** per region at provisioning time.

**9) Evaluation & QA (what we will measure)**

* **Answer quality**: groundedness pass rate; citation precision (node/edge IDs; Gazette no./date).
* **Temporal correctness**: “as-of” regression tests against known law changes (Gazette-stamped).
* **Recall on cross-jurisdiction questions**: presence of DIFC/ADGM where relevant.
* **Latency & cost**: tokens/query; RU/s & SU/h per query type; semantic rerank utilization.
* **User-visible paths**: % of answers with explorable Local paths.

**10) Cost model & projection (how you budget)**

**What drives cost most**: (1) **LLM tokens** and (2) **retrieval capacity**. Azure exposes clear dials: **Cosmos RU/s**, **Search SU/h**, **Semantic ranker $1/1k** queries (1k free), and **model tokens**.

**Azure-native lines**

* **Cosmos DB (Gremlin)** — compute billed by **RU/s**, with storage & bandwidth separate; autoscale/reserved capacity available.
* **Azure AI Search** — billed per **Search Unit**; **semantic ranker** billed by usage (first **1k free**, then **$1/1k**). Vector search has no extra fee beyond the SU.
* **Azure OpenAI** — token-metered; model & region determine rates/availability.
* **Document Intelligence (OCR)** — page-based pricing; Arabic supported.
* **Private Link** — endpoint hourly + data processed (small but non-zero).

**Neo4j (optional mirror / demo)**

* **AuraDB**: **$65/GB-mo (Professional)**; **$146/GB-mo (Business Critical)**. Excellent for demos and analyst exploration with **Bloom**, **Text2Cypher**, and **GDS**.

Projection approach: we will size (a) ingestion volume (pages/month), (b) peak QPS and mix (Local vs Global/DRIFT), (c) token budget per query path, then compute RU/s + SU/h + semantic rerank + tokens per month. This produces a defensible cost envelope you can tune by turning semantic rerank on/off or switching to smaller models for some steps.

**11) Delivery plan (phased)**

**Phase A — Pilot (Neo4j demo & Azure spine, 6–10 weeks)**

* Ingest 100–200 federal instruments + sample emirate gazettes + 50 free-zone judgments;
* **Neo4j demo**: **Text2Cypher** queries + **Bloom** views for legal reviewers;
* **Azure spine** in UAE North: AI Search index + Cosmos Gremlin schema + 3–5 **Gremlin tools** + groundedness check; show Local/Global/DRIFT behavior on the same slice.

**Phase B — Production MVP (Azure)**

* Scale ingestion to federal corpus; switch primary retrieval to **AI Search + Gremlin tools**; enforce **Private Link**, logging, Purview catalog.

**Phase C — Expansion**

* Add emirate corpora; wire **DIFC/ADGM** judgments; add conflict detectors and drafting assistants; optimize cost via semantic-ranker thresholds and reserved RU/s.

**12) Why this stack (summary of choices)**

* **UAE-fit**: models the **gazette-driven temporal lifecycle** and respects **federal/emirate/free-zone** split; sources are official.
* **Evidence-first**: GraphRAG **Local/Global/DRIFT** ensures recall and explainability at controllable cost.
* **Azure ops**: private networking, regional availability, and clear pricing dials (**RU/s, SU/h, $/1k semantic, tokens**).
* **Demo speed & audit UX**: **Neo4j Text2Cypher + Bloom + GDS** provides a great review surface while production runs in Azure.

**13) What you’ll see in the product (user experience)**

* **As-of toggle** (date picker) that re-filters graph edges by validity and shows Gazette-based entry-into-force text.
* **Explainable answers** with **path chips** (e.g., “Art. 12 → amended by Law 3/2023 (Gazette 7/2023)”) and **judgment clusters** for DIFC/ADGM.
* **Two retrieval modes**: “Direct (Local)” for specific articles; “Survey (Global/DRIFT)” for thematic/state-of-law summaries.

**Appendix — key references**

* **UAE sources**: UAE Legislation platform; Official Gazette rules (Constitution 111–112); DIFC & ADGM judgments.
* **GraphRAG**: Overview, **DRIFT**, dynamic community selection, accelerator.
* **Azure**: AI Search hybrid **RRF** & **semantic ranker** pricing; **Cosmos Gremlin** + **Private Link**; **groundedness**; **Purview UAE North**.
* **Neo4j**: **Text2Cypher** dataset; **Bloom**; **GDS Louvain**; **vector indexes**; **AuraDB pricing**.