Knowledge Graph for RAG — AWS & OSS Swimlane (One Pager)

Data → KG → Serving | CQs, RDF/RDFS/OWL, SKOS, SHACL, SPARQL, Federated SPARQL

YOUR LOGO

[0] Define Competency Questions (CQs)

Purpose: Scope answers; drives ontology & ETL.
Notes: capture KPIs, personas, and must-answer queries.
Tools: (AWS) QuickSight, (OSS) Markdown/ADR in Git, Confluence/Notion.

DATA

[1] Land raw sources → Bronze

Inputs: Sales DB, R&D PDFs, OCR PDFs, CAD/Images, XML/JSON
AWS: S3, Kinesis, DMS, Glue Crawlers, Textract, AppFlow, KMS/IA
OSS: Airbyte/NiFi, Tika (PDF), Tesseract (OCR), Scrapy/Playwright

[2] Clean & normalize → Silver

Parse CSV/JSON/XML, normalize dates/IDs, QA checks, light entity
AWS: Glue ETL (Spark), EMR Serverless, MWAA (Airflow), Great E
OSS: Apache Spark, dbt-core (tests), Great Expectations, Airflow.

[3] Conform → Gold

Conformed tables: dims/bridges/facts; dedupe + entity resolution.
AWS: Lake Formation, Glue Data Catalog, Athena, Iceberg on S3.
OSS: Delta / Iceberg / Hudi; dbt-core for models/tests.

KG

[4] Ontology & Vocabulary (RDF/RDFS/

Model classes/properties; labels & synonyms; ma
AWS: Store OWL/SKOS in S3; author via noteboo
OSS: Protégé; RDF/OWL/SKOS; Jena/RDF4J libs

[5] Map Gold → RDF Triples (URIs)

R2RML/RML mapping; mint URIs; serialize N-Trip
AWS: Glue/EMR PySpark jobs write triples to S3.
OSS: RMLMapper, Ontop, Jena RIOT, Python rdf

[6] Validate & Reason (SHACL + reason

SHACL shapes gate data; optional OWL reasonin
AWS: Run pySHACL/Jena on Glue/EMR/Batch; re
OSS: pySHACL, Jena SHACL; HermiT/ELK reaso

[7] Load Knowledge Graph (Graph DB)

Bulk load RDF; expose SPARQL endpoint; index/monit
AWS: Amazon Neptune (RDF/SPARQL), S3 Bulk Load
OSS: Jena TDB2 + Fuseki, Blazegraph*, Neo4j+n10s†

SERVING

[8] Retrieval indexes (RAG evidence)

8A Keyword/BM25 docs; 8B Vector embeddings for passages.
AWS: OpenSearch (BM25 & kNN), Bedrock/SageMaker embedding
OSS: OpenSearch self-managed; Qdrant/Milvus/FAISS; sentence-t

[9] Orchestrate Query → Answer (SPARQL & Federate

Entity link via SKOS; route: SPARQL (facts) | Vector RAG (open-en
AWS: API Gateway + Lambda/Fargate/EKS; Neptune SPARQL; St
OSS: FastAPI, LangChain/Haystack, Jena/RDFLib clients.

[10] Present & Observe

UI + exports; dashboards; query logs; SPARQL smoke tests (CQs).
AWS: Amplify/CloudFront/S3, CloudWatch/X-Ray, Athena on logs.
OSS: React/Vue, Grafana/Prometheus, RDFUnit, pytest.

Legend: RDF/RDFS/OWL & SKOS → [4]; SHACL → [6]; SPARQL (incl. Federated) → [7],[9]; KG (Neptune) → [7]; RAG Indexes → [8]

Generated 2025-08-31 16:51 UTC