Module 02 Data Engineering for AI

وحدة 02 هندسة البيانات للذكاء الاصطناعي

7 Modules + Lab + Glossary

Interactive Learning Journey

Course Modules وحدات المقرر

V06 أساسيات البيانات **Data Fundamentals**

V07 المستودعات والبحيرات ومعمارية **Warehouses, Data Lakes & Lakehouse Lakehouse**

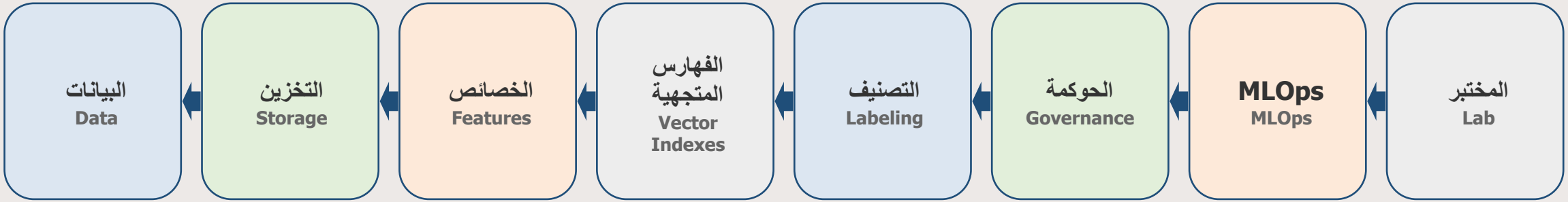
V08 هندسة الخصائص ومخازنها **Feature Engineering & Feature Stores**

V09 قواعد البيانات المتجهية **RAG** + التطبيقي **Vector Databases + Applied RAG**

V10 تصنيف البيانات على نطاق واسع **Labeling at Scale**

V11 حوكمة البيانات وسلسلة العهدة والفهرسة **Data Governance, Lineage & Cataloging**

V12 أساسيات **MLOps** + تمهيد للمختبر **MLOps Fundamentals + Lab Introduction**



V06 Data Fundamentals

Why Data is Important
أهمية البيانات

THE PROBLEM

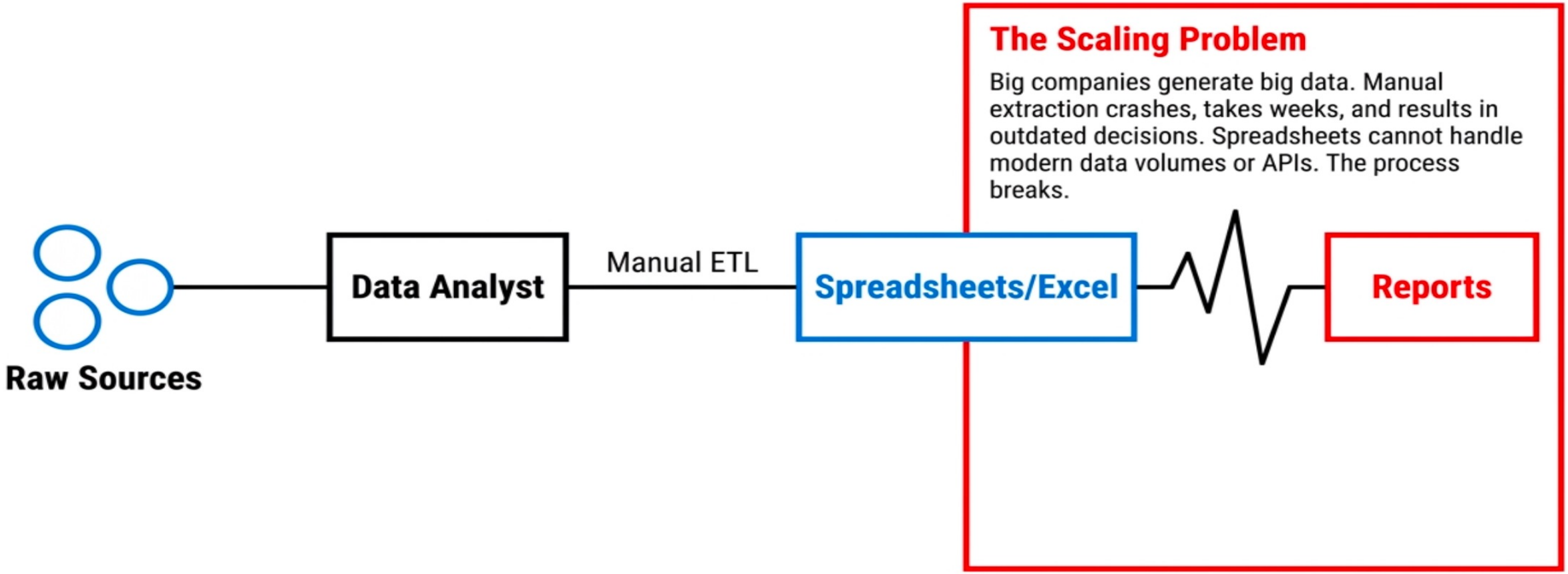
"Without data, you are just another person with an opinion." - William Deming

- Leads to confusion
- Wasted resources
- Bad decisions

THE SOLUTION

Replaces guesswork with facts for fast, smart decisions.

Big Data
البيانات الضخمة



Types of Data Platforms

أنواع قواعد و منصات البيانات

Storage Architectures

Data Warehouse

Structured, relational, optimized for BI. (Subject-oriented, integrated, non-volatile).

Data Lake

Flexible, stores structured + unstructured (images/logs). Risk: Can become a data swamp.

Lakehouse

The modern hybrid. Flexibility of a lake + structure of a warehouse.

Data Mesh

Decentralized. Departments build and share their own data products.

Database Types Matrix

Relational (SQL)

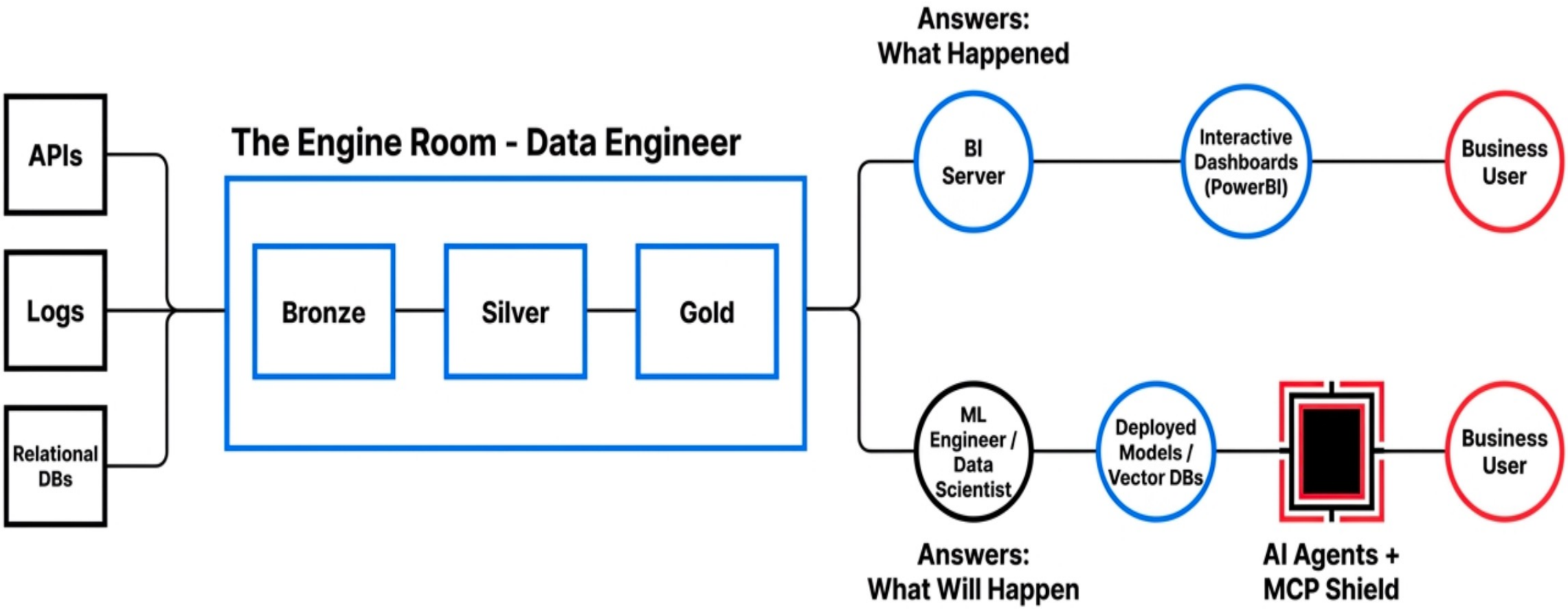
Tables/Rows/Columns.
(e.g., SQL Server, PostgreSQL).

NoSQL Variants

- Key-Value (Dictionaries)
- Column-based (Heavy searching)
- Graph (Relationships/Nodes)
- Document (JSON pages)

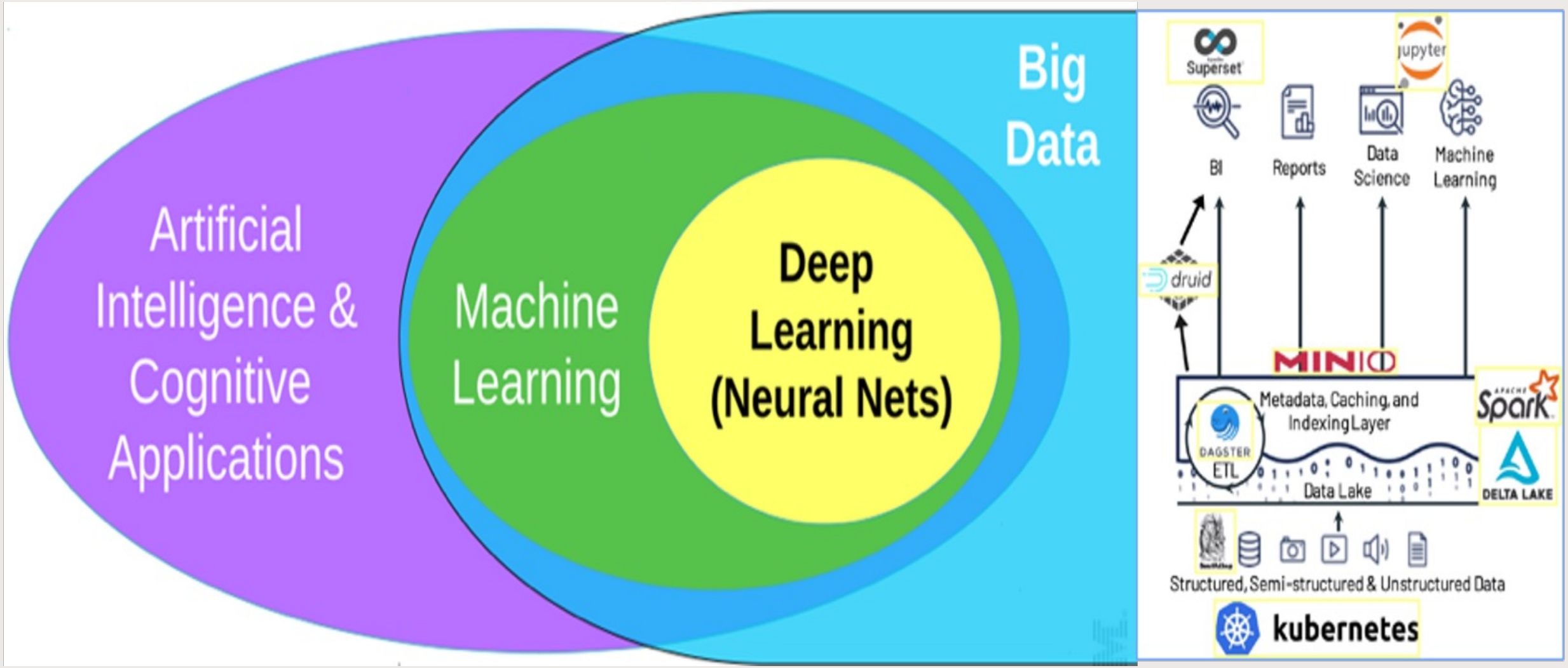
Unified Data/AI Platform

تقنيات معالجة منصة البيانات الموحدة و الذكاء الاصطناعي



Unified Data/AI Platform

تقنيات معالجة منصة البيانات الموحدة و الذكاء الاصطناعي



Unified Data/AI Platform

تقنيات معالجة منصة البيانات الموحدة و الذكاء الاصطناعي

Data Analysis and Segmentation

- جمع البيانات الأولية ، من خلال معالجة الأنشطة ، و التعرف على البيانات ، وتحديد جودة البيانات ، واكتشاف السمات الداخلية للبيانات مع اكتشاف المجموعة الفرعية المثيرة للاهتمام لتشكيل فرضية المعلومات الضمنية.
إنشاء منظومة البيانات الموحدة من البيانات غير المعالجة و تشمل اختيار الجداول والسجلات والخواص ، بالإضافة إلى تحويل وتنظيف البيانات لأدوات النمذجة.

Predictive Forecasting Models

- استخدام تقنيات الذكاء الاصطناعي في اختيار تقنيات نماذج مختلفة وتطبيقها ، وتعديل معلمات النموذج إلى أفضل القيم. بشكل عام ، بما يساعد علي إيجاد حلول فعالة لمشكلة التنقيب في البيانات .
- تحليل السلاسل الزمنية و التنبؤ بالقيمة المستقبلية للسلسل باستخدم المتوسطات المتحركة لتحليل السلاسل الزمنية إلى مكونات موسمية واتجاهية وغير منتظمة. التعامل مع المكونات الموسمية المضافة أو المضاعفة.

Anomaly & Fraud Detection

- اعداد القواعد المعرفية في النظم الخبيرة هي مزيج من الخصائص والقوانين ونتائج الخبرة المكتسبة والمعلومات الدقيقة والادراك
- استخدام برامج الذكاء الاصطناعي للوصول إلى استنباط الظواهر و الأنماط الغير الطبيعية في البيانات و احتمال مخالفتها للقوانين و المحددات من خلال المسارات داخل قاعدة المعرفة

Decision Optimization & What-If-Scenarios

- حل المشاكل الجديدة، والتخطيط، واتخاذ القرارات والقدرة على التعامل مع المشاكل غير المتوقعة واستخلاص النتائج باستخدام المعلومات المخزنة
- محاكاة النتائج المستقبلية و الوصول الي افضل البدائل لتعظيم العائد من المشروع

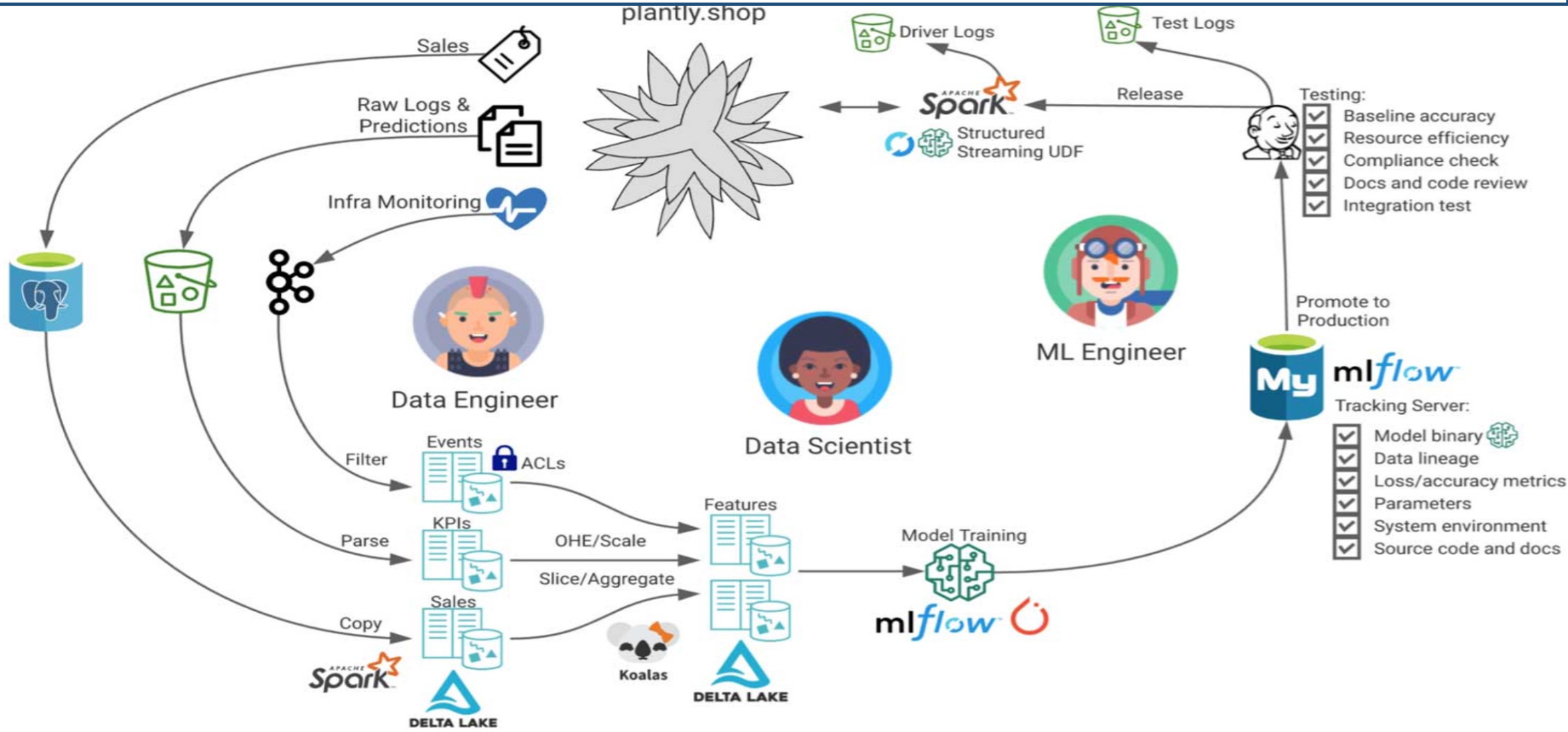
AI and Data Teams

فرق عمل الذكاء الاصطناعي و هندسة البيانات

Role	Core Question	Focus	Primary Output	System Activity
Data Analyst	What happened?	Past/Present	Visual Reports (PowerBI/Tableau	Prototyping / Discovery
BI Developer	How do we scale this?	Past/Present	24/7 Interactive Dashboards	Production / Engineering
Data Scientist	What will happen next?	Future	Predictive Models (Python)	Prototyping / Discovery
ML Engineer	How do we automate this?	Future	Deployed AI Systems / APIs	Production / Engineering
Data Architect & Engineer	How does data flow?	Foundation	Pipelines & Data Warehouses	Blueprint & Engine Room

AI and Data Teams

فرق عمل الذكاء الاصطناعي و هندسة البيانات



طيف البيانات والتنسيقات Data Spectrum & Formats

مهيكلة Structured	شبه مهيكلة Semi-structured	غير مهيكلة Unstructured
Tables جداول CSV Databases	Flexible records سجلات مرنة JSON Avro	Text, نصوص وصور وصوت images, audio PDF JPG MP3

اختر التنسيق حسب الاستخدام تبادل بسيط ← تحليلات سريعة ← محتوى غني

Choose the format for the job simple exchange ← fast analytics ← rich content

مسارات ETL & ELT

ETL التحويل قبل التحميل؛

ELT التحميل أولاً ثم التحويل داخل المنصة المستهدفة.

ETL transform before loading; ELT load first, transform inside the target platform.

ELT هو النمط السائد في بيئات **Lakehouse** الحديثة.

ELT is the dominant pattern in modern lakehouse environments.

مقارنة أنماط المعالجة Processing Patterns

المعمارية المتدرِّجة Medallion Lakehouse

Bronze خام → Silver منقَّى → Gold جاهز للتحليل؛ إدارة مركزية؛ الأنسب للتحليلات وخطوط AI ML RAG.

Medallion Lakehouse raw Bronze → cleansed Silver → analytics-ready Gold; central ownership; best for analytics and AI ML RAG pipelines.

مستودع البيانات Datawarehouse

ETL → Staging → أسواق البيانات؛ حوكمة مركزية؛ الأنسب للتقارير وذكاء الأعمال → المستودع.

Data warehouse ETL ELT → staging → warehouse → data marts; central governance; best for reporting and BI.

شبكة البيانات Data Mesh

ملكية لامركزية بالمجالات؛ منتجات بيانات بعقود SLA، مخطط، أمان وكتالوج موحد؛ الأنسب للمؤسسات الكبيرة.

Data mesh decentralized domain ownership; data products with contracts schema, security, SLA and a unified catalog; best for large organizations.

جدول المقارنة Medallion المستودع Data Mesh

المعيار	Medallion	المستودع	Data Mesh
الفكرة	طبقات Bronze→Silver→Gold	Staging→Warehouse →Marts	منتجات بالمجالات
المالك	هندسة مركزية	فرق BI	فرق المجالات
الأنسب	تحليلات و AI RAG	تقارير BI	مؤسسات كبيرة
الحوكمة	جودة بين الطبقات	مركزية	موحدة + عقود

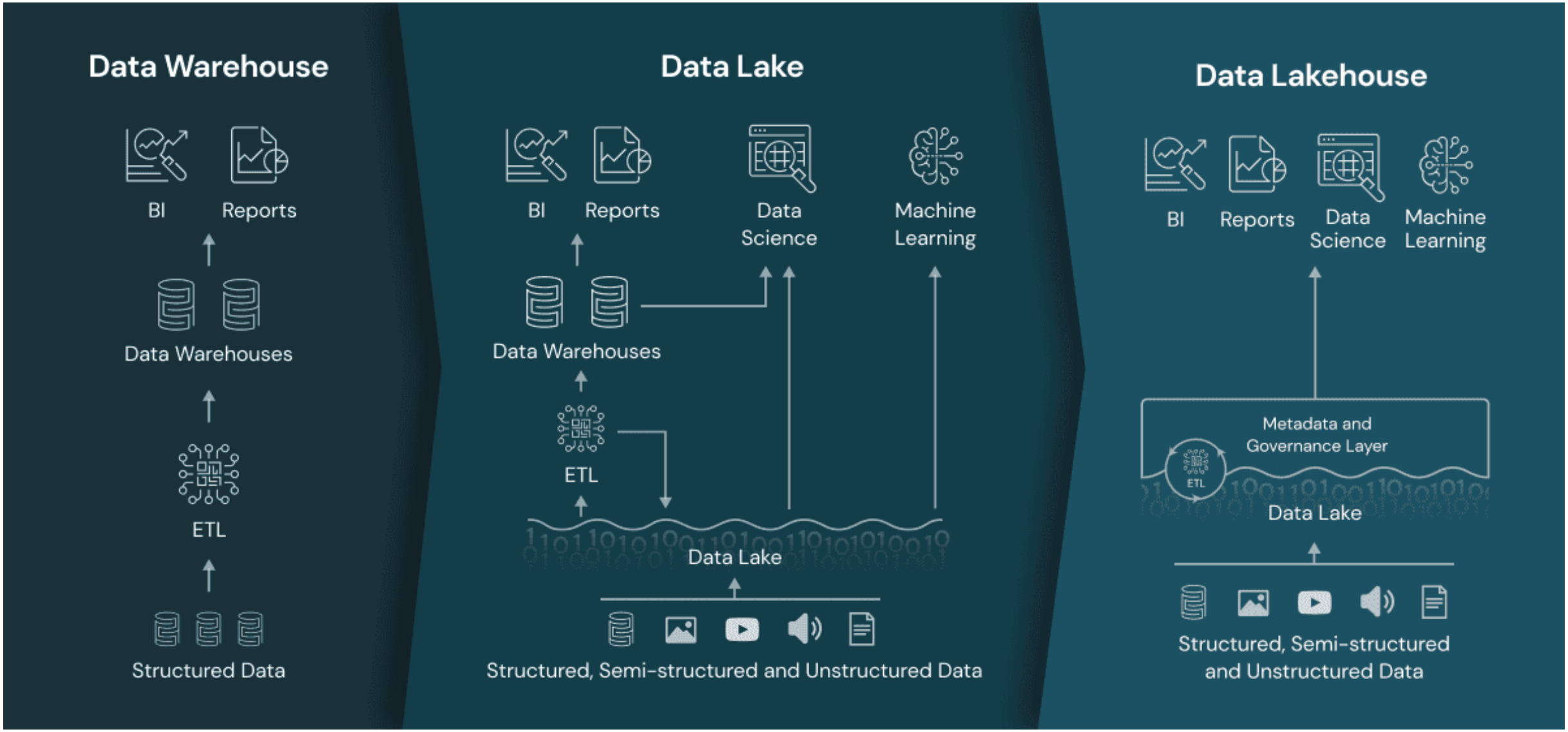
V07 Warehouses, Data Lakes & Lakehouse

أجندة العرض Agenda

V07 المستودعات والبحيرات ومعمارية Lakehouse, Data Lakes & Lakehouse

- المستودع مقابل البحيرة Warehouse vs Lake
- معمارية Lakehouse
- منتجات البيانات وعقودها Data Products & Contracts
- متى تختار كل نموذج؟ When to Choose Which

Warehouse vs Lake Decision Map المستودع مقابل البحيرة



خريطة قرار المستودع مقابل البحيرة Warehouse vs Lake Decision Map

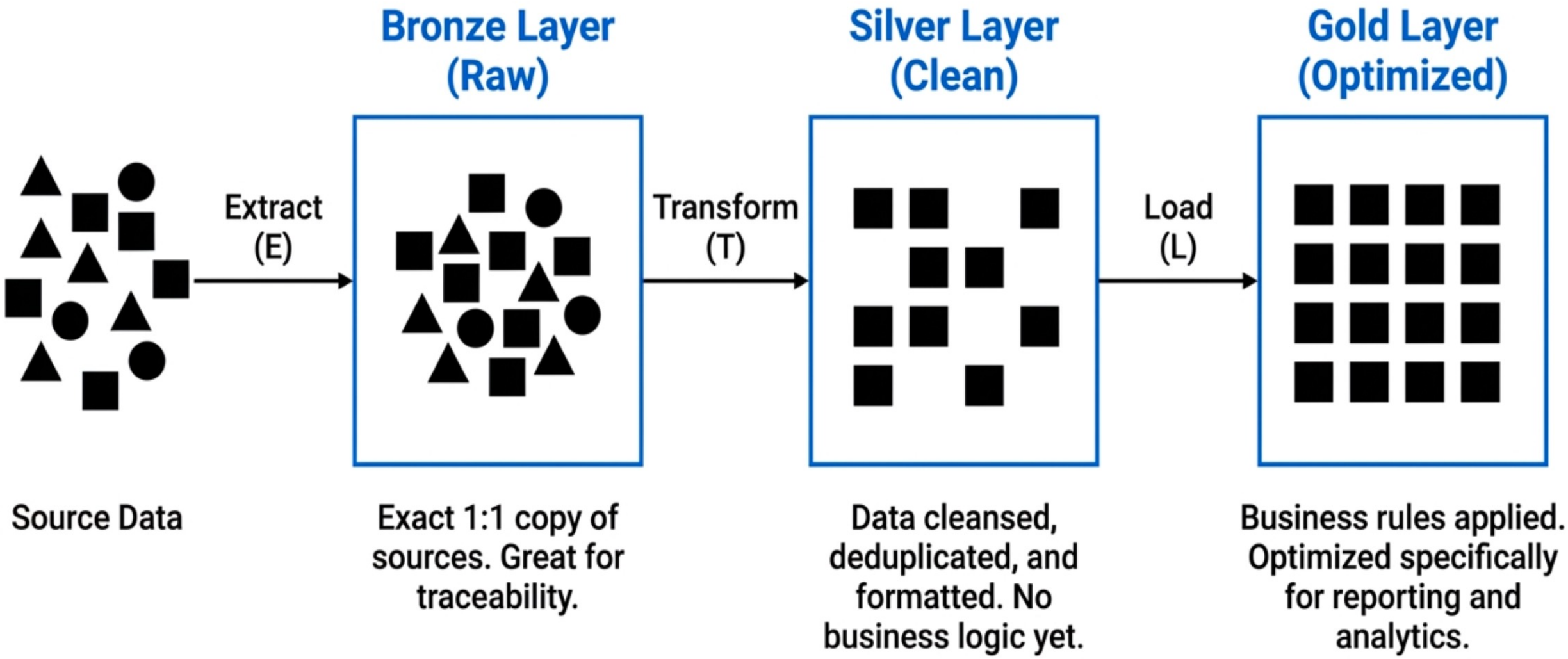
مستودع البيانات Data Warehouse	بحيرة البيانات Data Lake
مخطط عند الكتابة Schema on write	مخطط عند القراءة Schema on read
بيانات منظمة + SQL Curated data + SQL	كل التنسيقات Any format
تقارير وذكاء أعمال Reporting and BI	استكشاف و AI ML Exploration and AI ML

Lakehouse = مرونة البحيرة + موثوقية المستودع

Lakehouse = lake flexibility + warehouse reliability

Modern Data Platform Architecture (Medalion)

المعمارية المتدرجة لمنصات البيانات الحديثة



المعمارية المتدرِّجة Medallion Layers

Bronze بيانات خام غير قابلة للتغيير
Raw immutable data

Silver بيانات منقَّاة ومنظَّمة
Cleansed & structured data

Gold تجميعات جاهزة للتحليل
Analytics-ready aggregates

Data Product & Contract منتج البيانات وعقد البيانات

Security الأمان

SLA مستوى الخدمة

عقد البيانات
Data Contract

Schema المخطط

منتج البيانات
Data Product موثوق موثّق قابل للاكتشاف

متى تختار كل نموذج؟ Decision Matrix

السبب	الاختيار	الحالة
مرونة + حوكمة طبقية	Lakehouse Medallion	فريق مركزي + تحليلات
SQL سريع + مخطط ثابت	Warehouse	تقارير صارمة
ملكية لامركزية + عقود	Data Mesh	فرق مستقلة كبيرة

Three Data Patterns أنماط معالجة البيانات الثلاثة

شبكة البيانات Data Mesh	مستودع البيانات Data Warehouse	المعمارية المتدرِّجة Medallion
فرق المجالات مؤسسات كبيرة	فرق BI تقارير وذكاء أعمال	فريق مركزي تحليلات و AI ML RAG

V08 Feature Engineering & Feature Stores

أجندة العرض Agenda

V08 هندسة الخصائص ومخازنها Feature Engineering & Feature Stores

- هندسة الخصائص Feature Engineering
- مخازن الخصائص Feature Stores
- اتساق Consistency Offline Online
- الربط بالتطبيقات الوكيلة Agentic Applications

Raw Data to Features من البيانات الخام إلى الخصائص

أحداث خام
Raw events

→

تنظيف وتحويل
Clean & transform

→

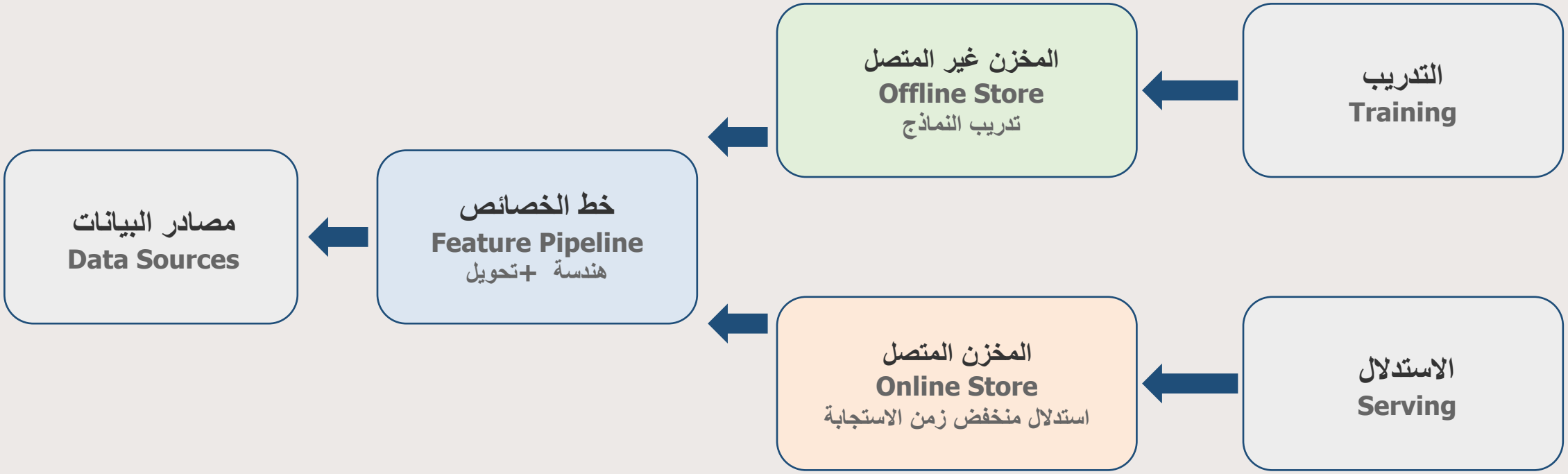
تعريف الخصائص
Feature definitions

→

تدريب + خدمة
Training + serving

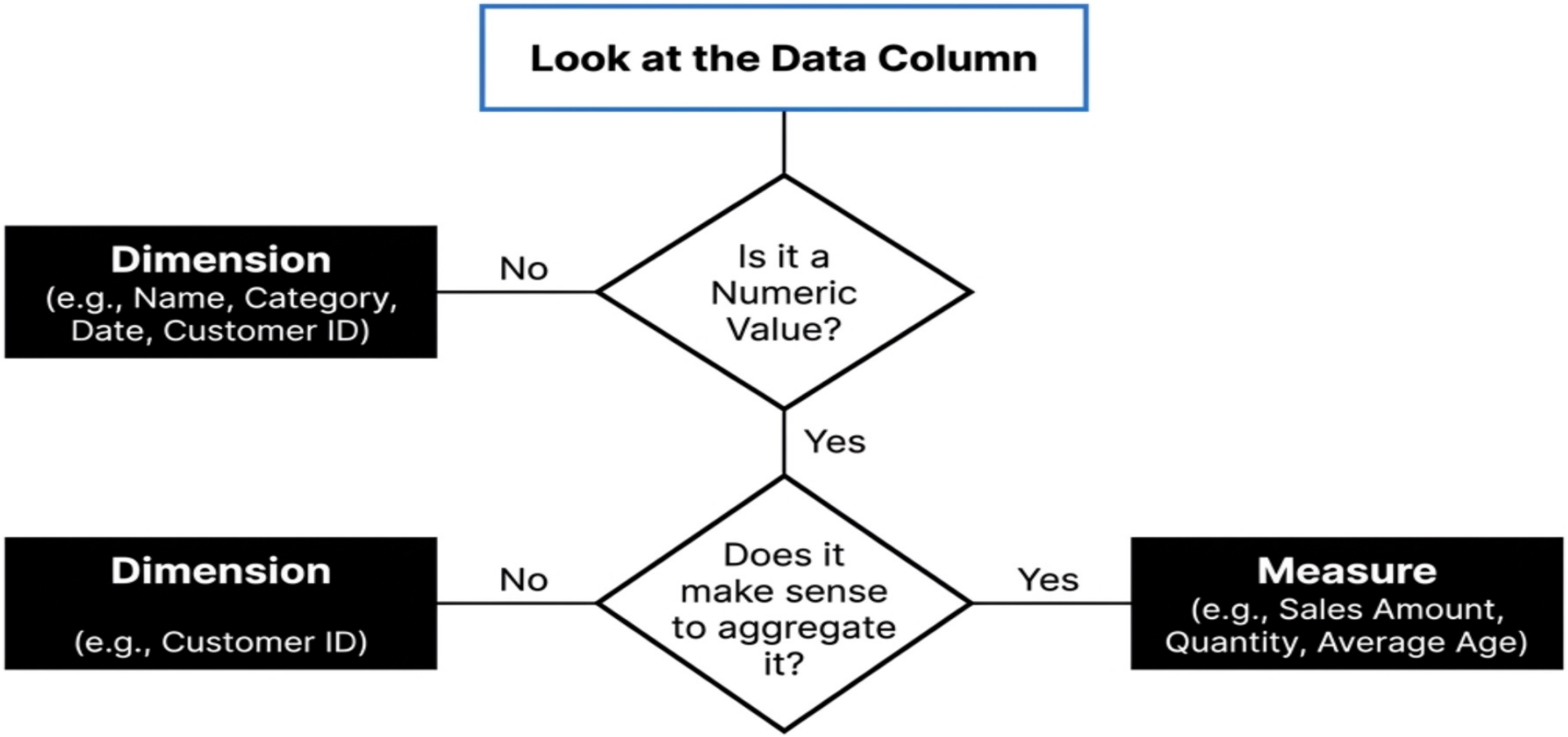
مخزن الخصائص تعريف واحد متسق للتدريب والخدمة
Feature store one consistent definition for training and serving

Feature Store Architecture معمارية مخزن الخصائص



Feature Engineering Importance

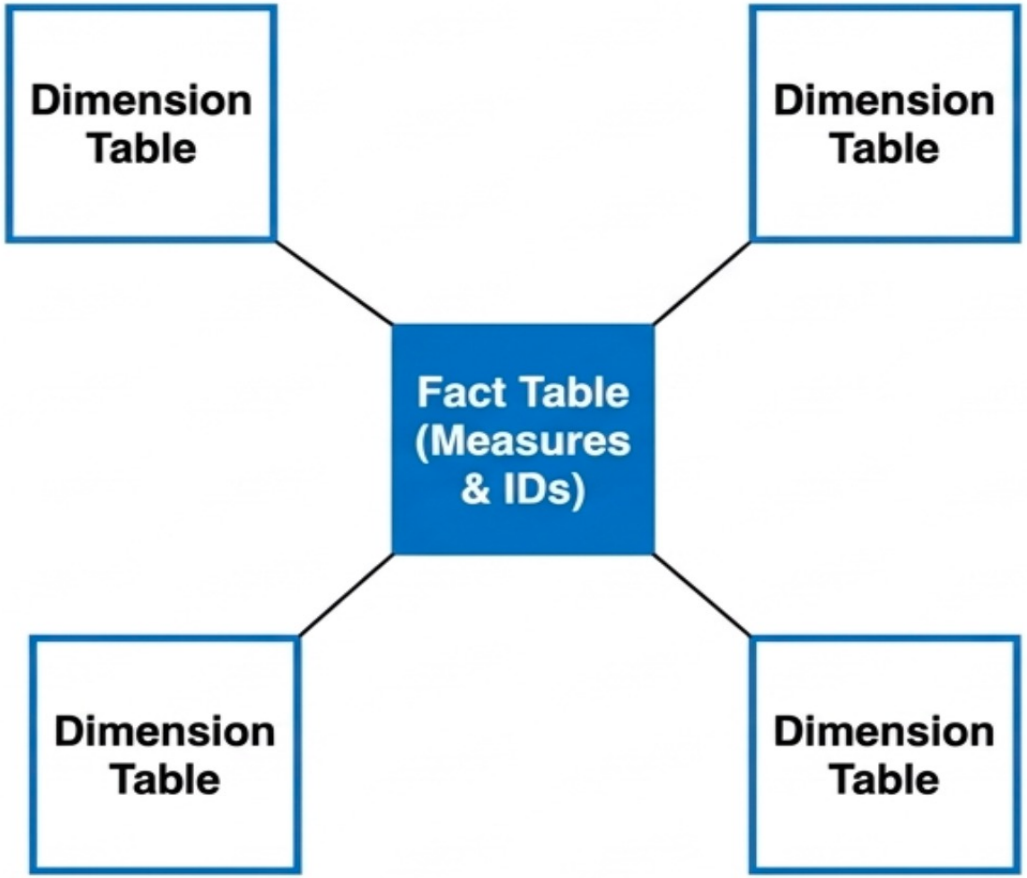
أهمية هندسة الخصائص للبيانات



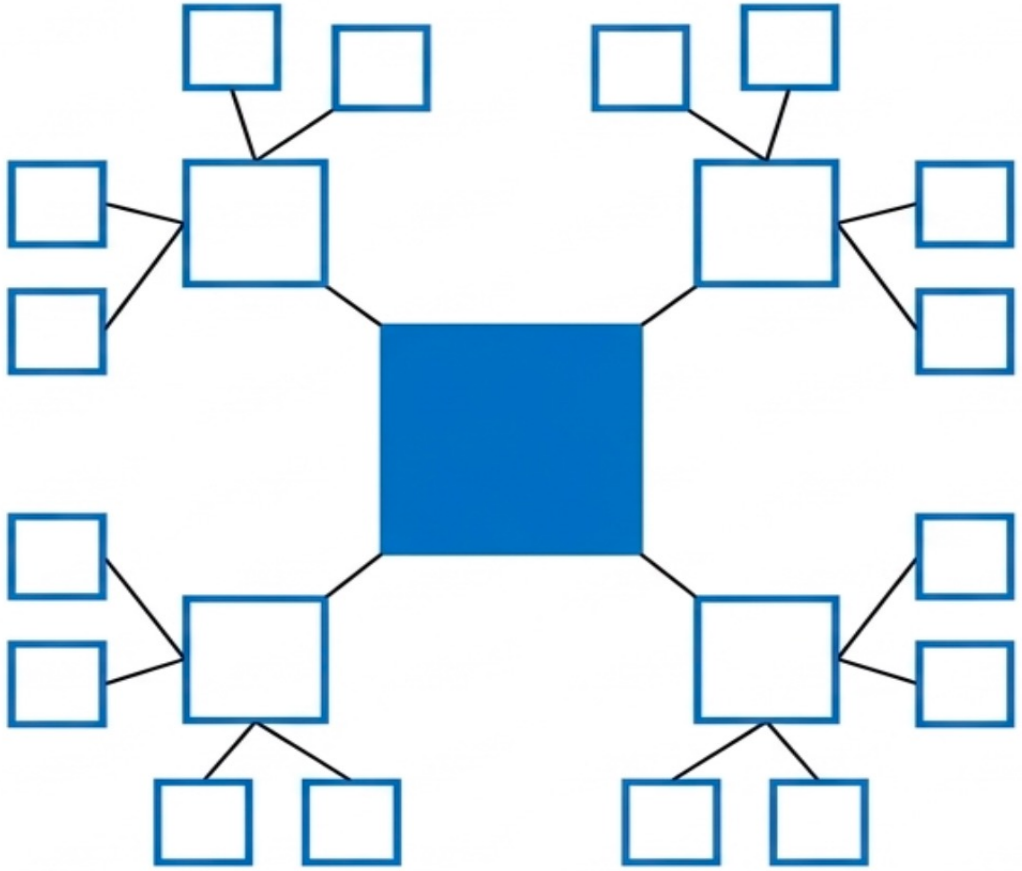
Measures answer 'How Much?' or 'How Many?'. Dimensions answer 'Who, What, Where?' and are used to group the data.

Design Patterns of Data Platforms

أنماط تصميم منصات البيانات



Star Schema: Simple, fast for reporting (PowerBI), denormalized.



Snowflake Schema: Complex, highly normalized, saves storage but requires heavy SQL joins.

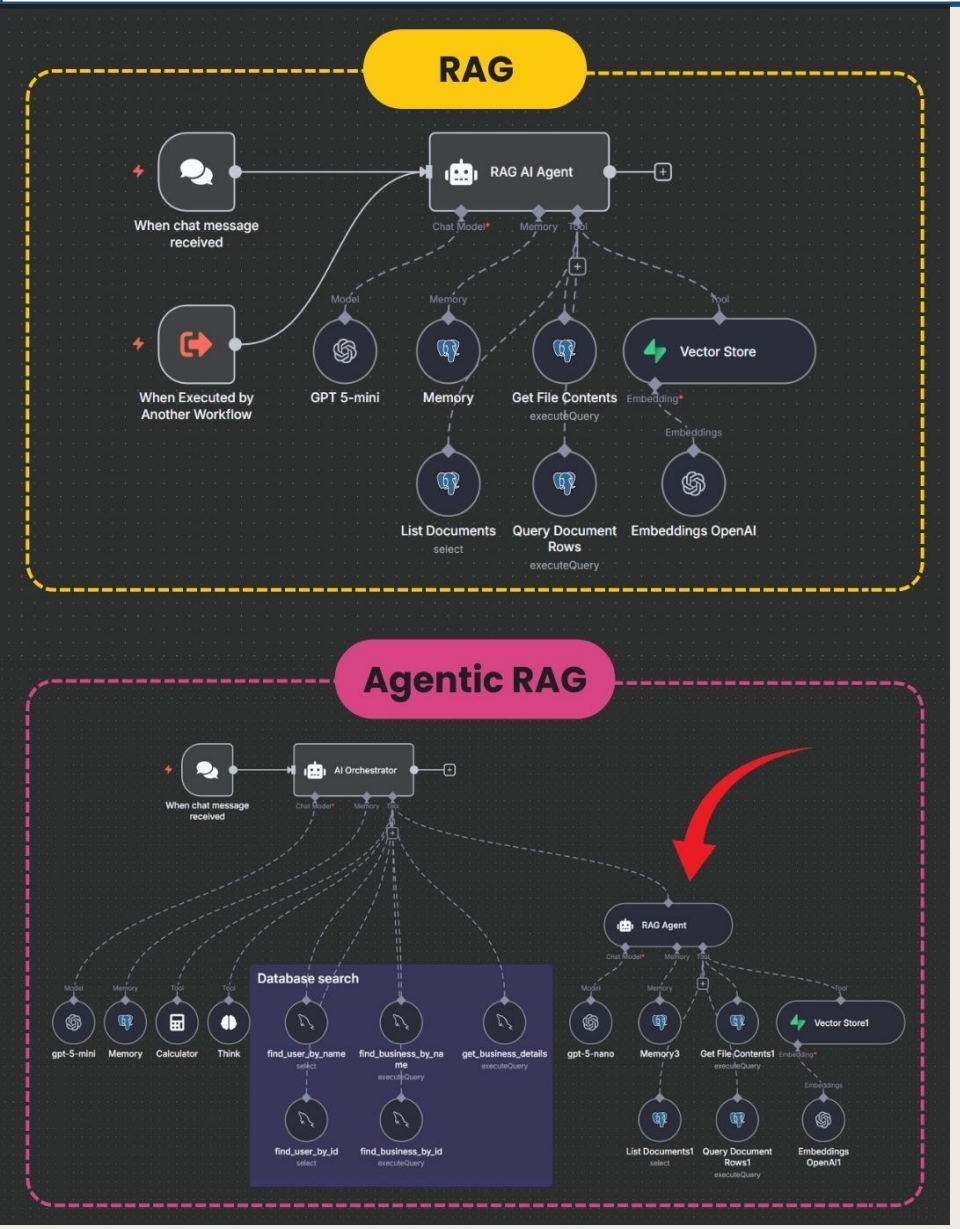
Feature Engineering is a MUST for AI

أهمية هندسة خصائص البيانات للذكاء الاصطناعي

Descriptive Analytics	Predictive Analytics
- **Question:** What happened? (e.g., How many customers left?) - **Tool:** SQL, BI Dashboards. - **Role:** Data Analyst.	- **Question:** What will happen? (e.g., Which customers will leave next month?) - **Process:** Exploratory Data Analysis (EDA) -> Feature Engineering (Deriving new data columns) -> Train/Test Data Split -> Apply Algorithm -> Deploy Model. Algorithm -> Deploy Model. - **Role:** Data Scientist & ML Engineer.

Feature Engineering is the most critical step. AI cannot find insights if the raw data isn't transformed into smart, derived features.

الربط بالتطبيقات الوكيلة Agentic Applications



الوكلاء يحتاجون ميزات لحظية متسقة لقرارات التخصيص والمخاطر والتوجيه.

Agents need consistent real-time features for personalization, risk, and routing.

مخزن الخصائص هو «مصدر الحقيقة» للميزات أمام أي وكيل ذكي.

The feature store is the single source of truth behind any agent.

من LLM إلى الأنظمة الوكيلة From LLM to Agentic AI

الأنظمة الوكيلة	الوكلاءAgents	الذكاء التوليدي	LLM
تخطيط وتنفيذ متعدد الخطوات Multi-step planning	نية + استدعاء أدوات Intent + tool calls	يولّد محتوى جديدًا Generates new content	تنبؤ بالرموز التالية Predicts next tokens

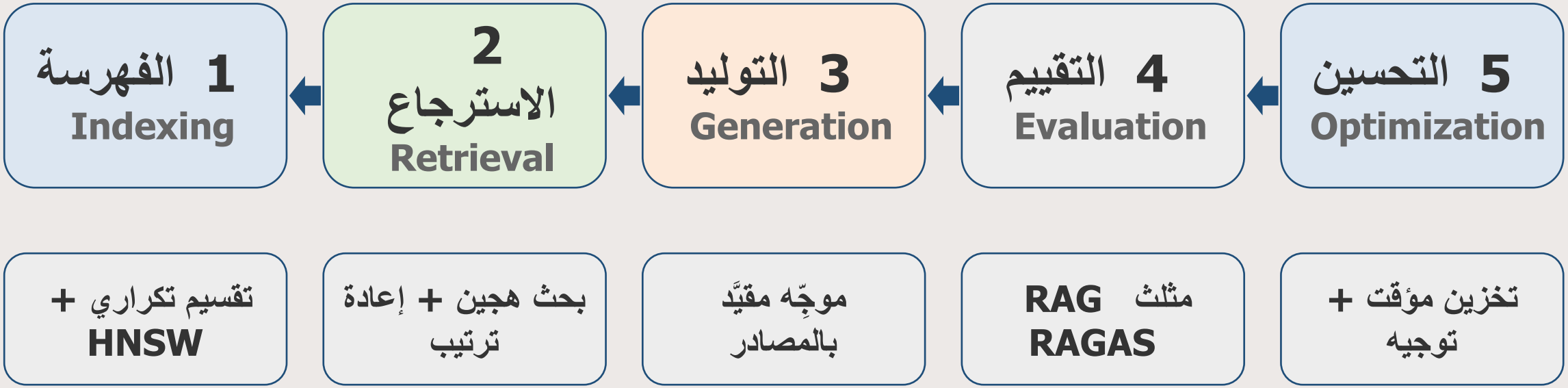
V09 Vector Databases + Applied RAG

أجندة العرض Agenda

V09 قواعد البيانات المتجهية RAG + التطبيقي Vector Databases + Applied RAG

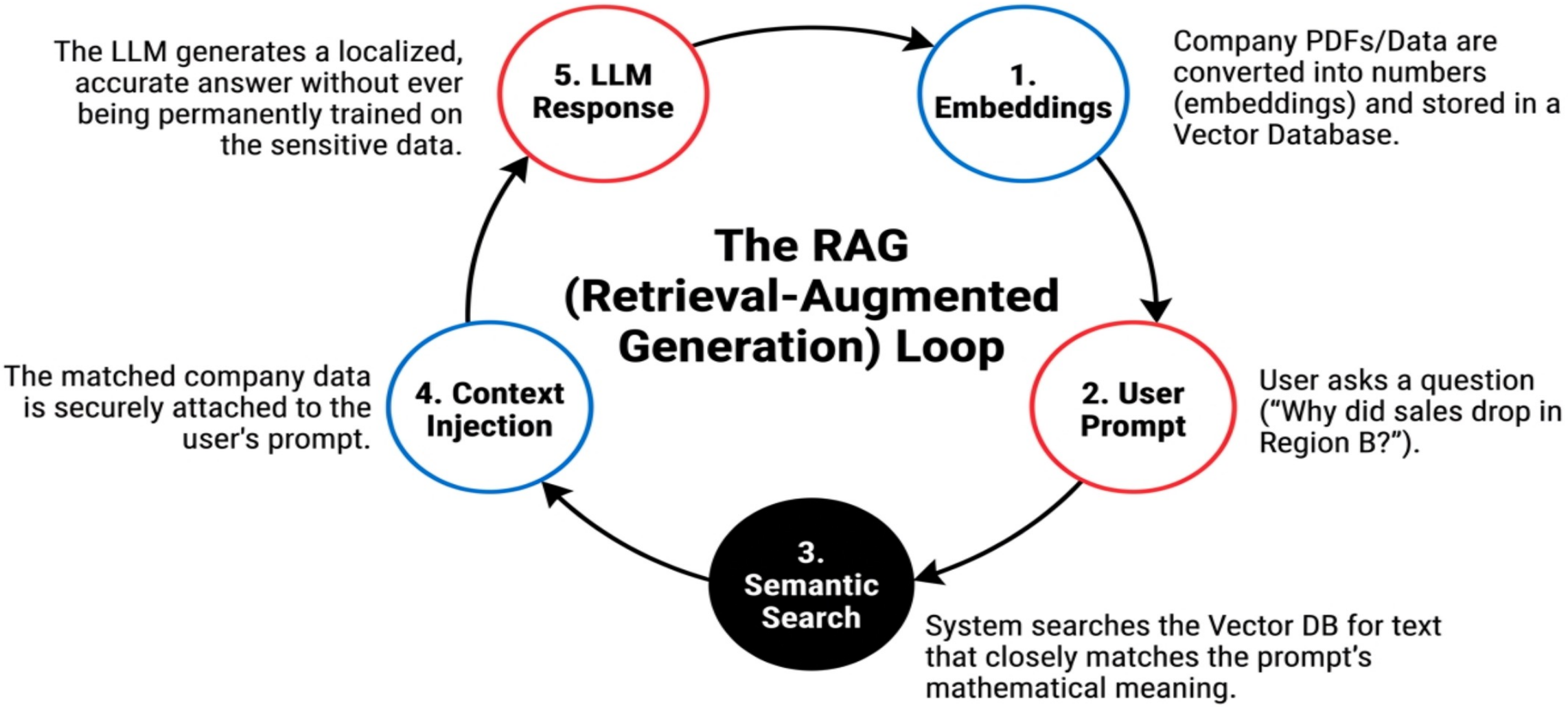
- قواعد البيانات المتجهية Vector Databases
- الفهرسة استراتيجيات التقسيم Indexing Chunking
- الاسترجاع Retrieval
- التوليدGeneration
- التقييم + التحسين Evaluation + Optimization
- الأخطاء الشائعة Common Pitfalls

خط مسارات RAG Pipeline



End-to-End RAG Pipeline

المسار المتكامل لنظام الاسترجاع التوليدي المعزز



استراتيجيات تقسيم النصوص Chunking Strategies

ثابت الحجم Fixed-size	أبسطها يقطع الجُمل
دلالي Semantic	تماسك أفضل حساس للعتبة
تكراري★ Recursive	الموصى به افتراضيًا
بحسب البنية Structure-based	عناوين وأقسام يُدمج مع التكراري
بالنموذج اللغوي LLM-based	الأدق الأغلى حسابيًا

Retrieval. الاسترجاع

بحث هجين كثيف + Dense متناثر BM25 يغطي المعنى والكلمات المفتاحية معًا.

Hybrid search dense + sparse BM25 covers meaning and keywords.

إعادة ترتيب Cross-encoder وتحويل الاستعلام Multi-Query يرفعان دقة النتائج.

Cross-encoder reranking and multi-query transformation improve precision.

Generation التوليد

موجِّه مقيَّد بالمصادر المسترجعة Grounded Prompting مع الاستشهاد «أجب فقط من السياق المرفق.»

Grounded prompting strictly on retrieved sources with citations 'answer only from the provided context'.

Grounded Answer Pattern نمط الإجابة المؤرَّضة

سؤال المتعلِّم
Learner question

ما الاستراتيجية الافتراضية للتقسيم؟
Which chunking default should I use?

أدلة مسترجعة
Retrieved evidence

الفقرات ← الجمل ← التقسيم
paragraphs → sentences → split

التكراري هو الخيار الموصى به
recursive is the recommended default

إجابة موثقة
Grounded answer

ابدأ بالتقسيم التكراري
Start with recursive chunking

No retrieved evidence? Say I don't know. لا دليل مسترجع؟ قل لا أعرف

التقييم + التحسين Evaluation + Optimization

مثلث RAG

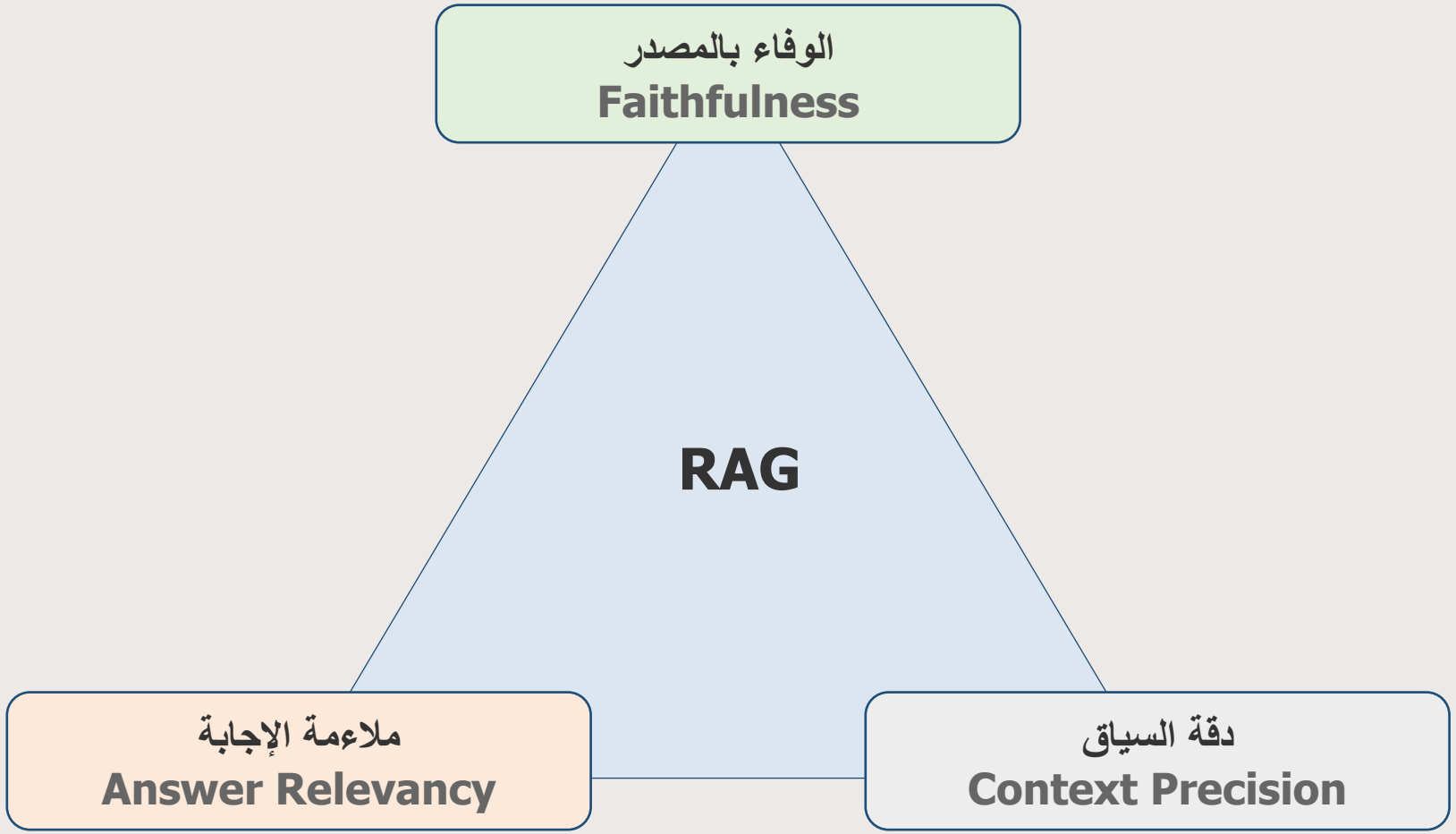
- الوفاء Faithfulness
- ملاءمة الإجابة Answer Relevancy
- ، دقة السياق Context Precision

RAG triad faithfulness, answer relevancy, context precision measured with RAGAS.

التحسين تخزين دلالي مؤقت، توجيه الطلبات بين النماذج، ومراقبة مستمرة.

Optimization semantic caching, model routing, continuous monitoring.

مثلث تقييم RAG Evaluation Triad RAG



Common Pitfalls الأخطاء الشائعة

تقسيم يكسر الجداول أو الأكواد؛ إهمال إعادة الترتيب فيضيع الأكثر ملاءمة.

Chunking that breaks tables code; skipping reranking loses the most relevant hits.

غياب تقييم الوفاء فيسمح بالهلوسة؛ استرجاع أعلى-ك دون حد أدنى للدرجة.

No faithfulness evaluation allows hallucination; top-k without a score floor.

Indexing Chunking Strategy (Recursive), Embedding Model (Domain-aligned), Vector Index (HNSW).	**Retrieval** Hybrid Search (Dense+Sparse), Reranking (Cross-encoder), Query Transformation (Multi-Query).	**Generation** Grounded Prompting (Strict Instructions), Source Citations.
	Evaluation The RAG Triad (Faithfulness, Answer Relevancy, Context Precision).	
	Optimization Semantic Caching, Model Routing, Continuous Monitoring.	

V10 Labeling at Scale

أجندة العرض Agenda

V10 تصنيف البيانات على نطاق واسع Labeling at Scale

- لماذا التصنيف؟ **Why Labeling?**
- الأدوات **Tools**
- التصنيف بمساعدة **LLM-assisted Labeling LLM**

Label Quality Loop حلقة جودة التصنيف

تسميات متسرعة أو غير متسقة
Rushed or inconsistent labels

ضوضاء ← نموذج ضعيف ← تصحيحات أكثر
Noise → weak model → more corrections

اقتراح LLM + مراجعة بشرية
LLM suggestion + human review

تسميات موثوقة ← نموذج أفضل ← تغذية راجعة
Trusted labels → better model → feedback

قِس اتفاق المقيمين قبل قبول البيانات Measure reviewer agreement before accepting data

Tools الأدوات

Label Studio للتعددية Roboflow، للرؤية الحاسوبية Scale AI، للتوسع المؤسسي.

Label Studio for versatility, Roboflow for vision, Scale AI for enterprise scale.

LLM-assisted Labeling التصنيف بمساعدة LLM

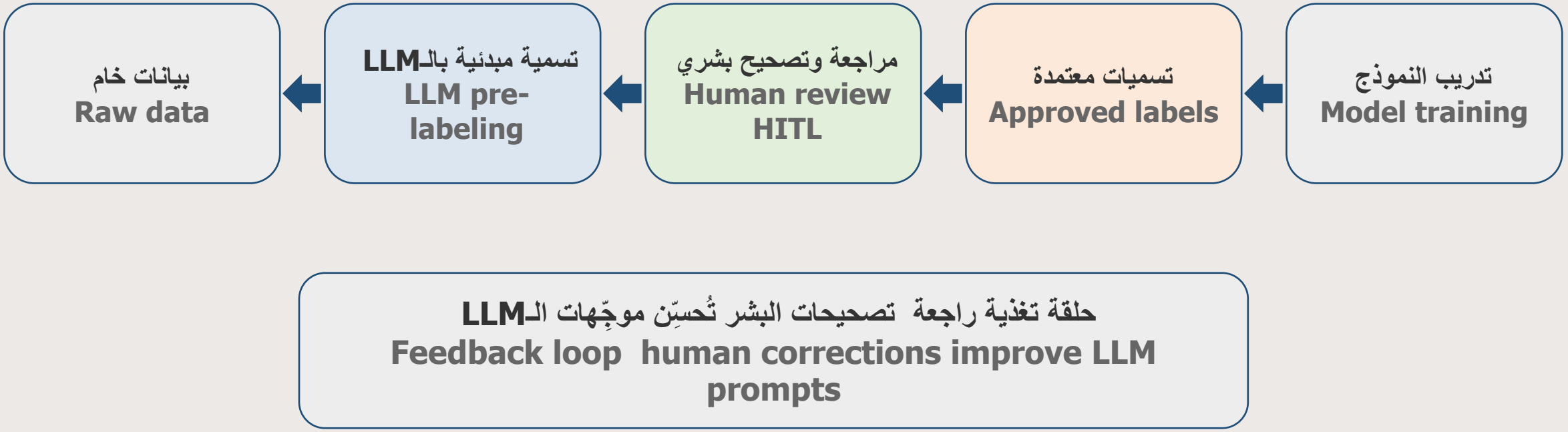
نموذج لغوي يقترح التصنيفات مبدئيًا لتسريع العملية وخفض التكلفة.

An LLM pre-labels examples to speed up the process and cut cost.

التحقق البشري Human-in-the-loop إلزامي قبل القبول، مع قياس اتفاق النموذج مع البشر.

Human-in-the-loop validation is mandatory before acceptance, measuring human–model agreement.

LLM-assisted Labeling Workflow سير عمل التصنيف بمساعدة LLM



V11 Data Governance, Lineage & Cataloging

أجندة العرض Agenda

V11 حوكمة البيانات وسلسلة العهدة والفهرسة Data Governance, Lineage & Cataloging

- الحوكمة والفهرسة Governance & Cataloging
- المعمارية المرجعية للذكاء التوليدي GenAI Reference Architecture

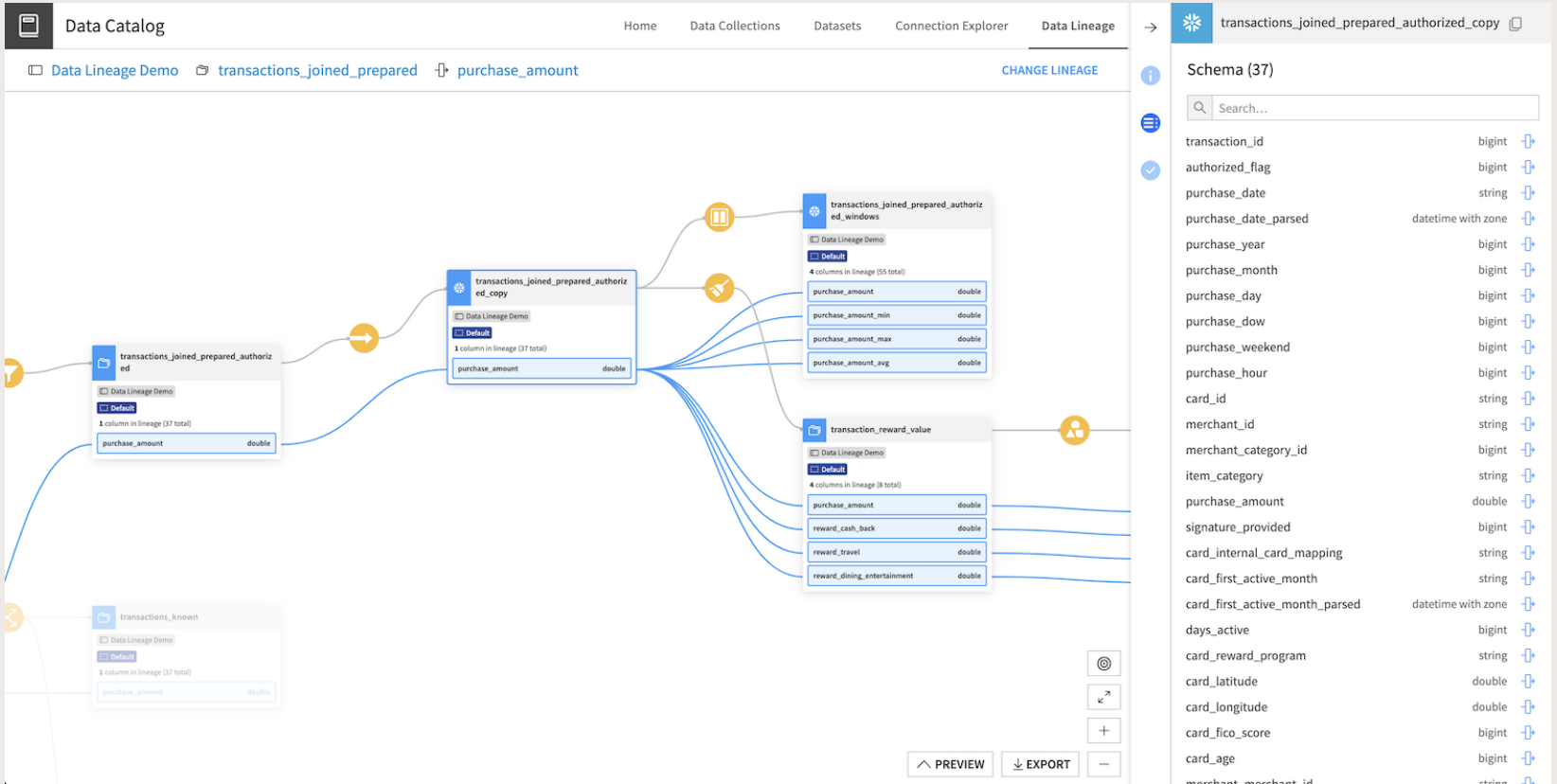
الحوكمة والفهرسة Governance & Cataloging

سلسلة العهدة **Lineage** تتبع أصل البيانات وتحولاتها؛ الفهرسة تجعلها قابلة للاكتشاف.

Lineage traces data origin and transformations; cataloging makes it discoverable.

أدوات. **Apache Atlas، Collibra، DataHub, OpenMetaData.**

Tools Apache Atlas, Collibra, DataHub, OpenMetaData



خريطة سلسلة العهدة Data Lineage Map

مصدر Source	تحويل Transform	منتج بيانات Data product	استهلاك Consume
CRM API	تنظيف + توحيد	مخطط + مالك	لوحة نموذج

لكل خطوة مالك + وقت + سياسة + أثر التغيير
For every step owner + time + policy + change impact

GenAI Reference Architecture 1 / 2 المعمارية المرجعية للذكاء التوليدي

أ. بوابة منصة GenAI نافذة موحدة بكتالوج حلول ومحفظة PROD → MVP → POC موثقة.

A. Portal unified pane of glass with a solution catalog and documented POC → MVP → PROD portfolio.

ب. الأتمتة والامتثال خطوط CI/CD ببوابات تحكم آلية وموافقات مرحلية وفق DevSecOps.

B. Automation & compliance CI/CD with automated control gates and stage-gate approvals DevSecOps .

GenAI Reference Architecture 2 / 2 المعمارية المرجعية للذكاء التوليدي

ج. الخدمات المشتركة بوابة AI موحدة، مكتبة موجِّهات، خدمة التغذية الراجعة، قاعدة تدقيق LLM، حواجز حماية Guardrails.

C. Shared services unified AI gateway, prompt library, feedback service, LLM audit database, guardrails.

د. الحوكمة والمراقبة أمن البيانات، الذكاء المسؤول تدقيق التحيّز والهلوسة ، وإدارة التكاليف FinOps.

D. Governance & monitoring data security, responsible AI bias hallucination audits , FinOps.

GenAI Reference Architecture المعمارية المرجعية للذكاء التوليدي

أ. بوابة منصة GenAI
A. Platform Portal POC → MVP → PROD

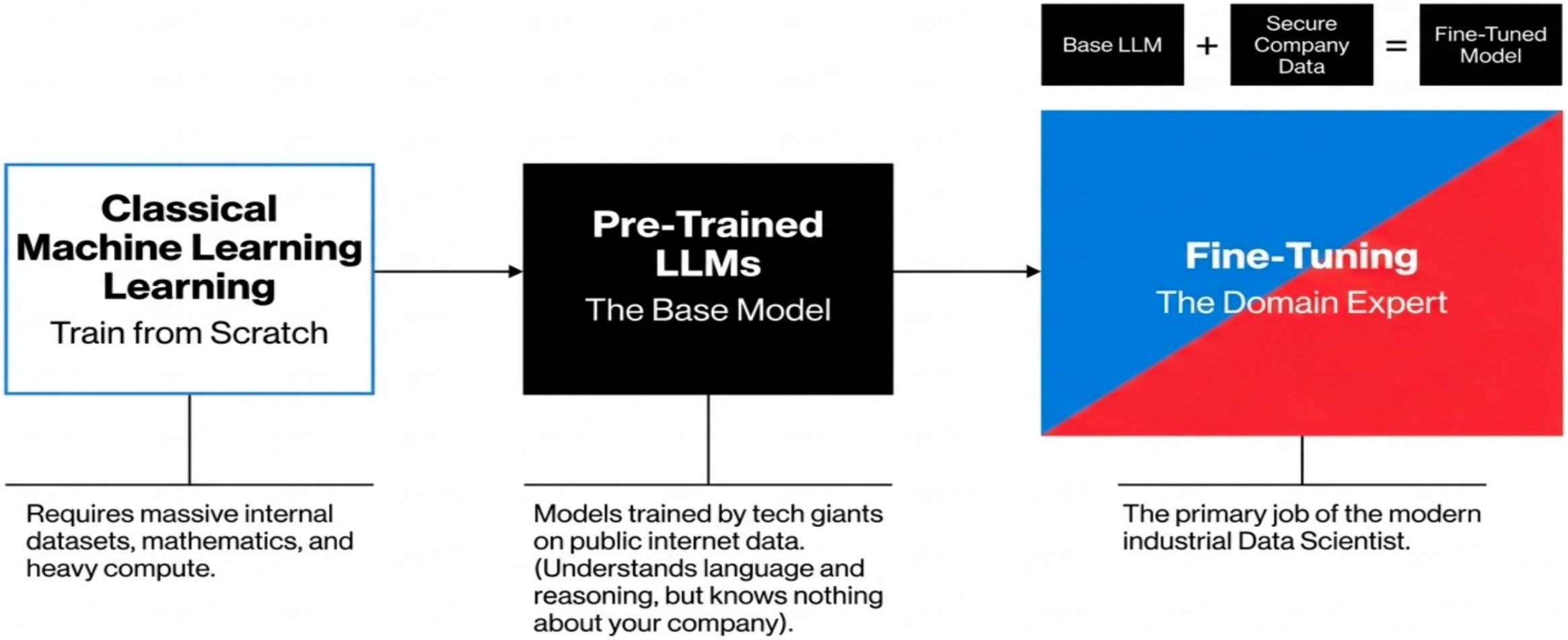
ب. الأتمتة والامتثال
B. Automation & Compliance CI CD gates

ج. الخدمات المشتركة
C. Shared Services Gateway, Prompts, Audit, Guardrails

د. الحوكمة والمراقبة
D. Governance & Monitoring Responsible AI, FinOps

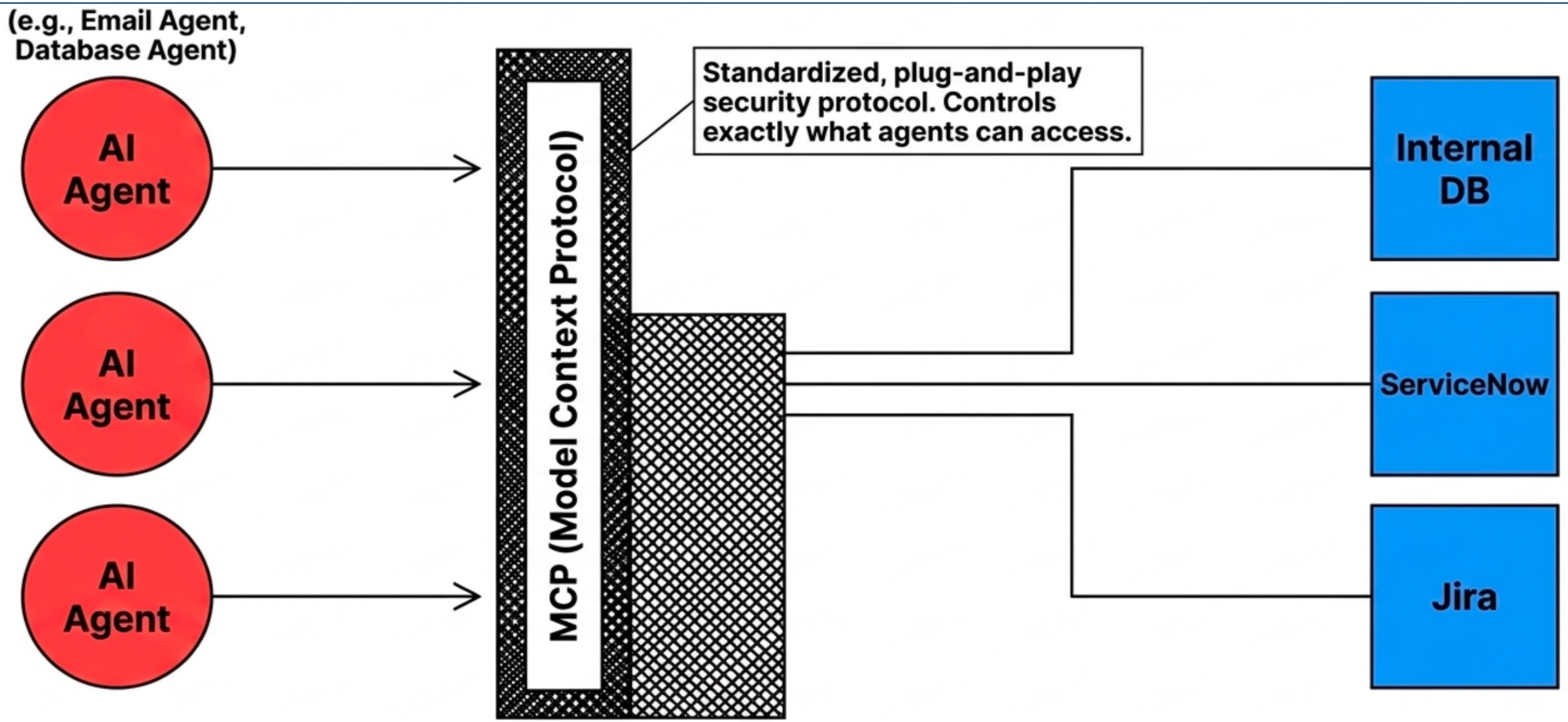
Generative AI and LLM Fine-Tuning using YOUR Data

الذكاء الاصطناعي التوليدي وتحسين نماذج التعلم الآلي باستخدام بياناتك



MCP: Model Context Protocol

بروتوكول سياق النموذج

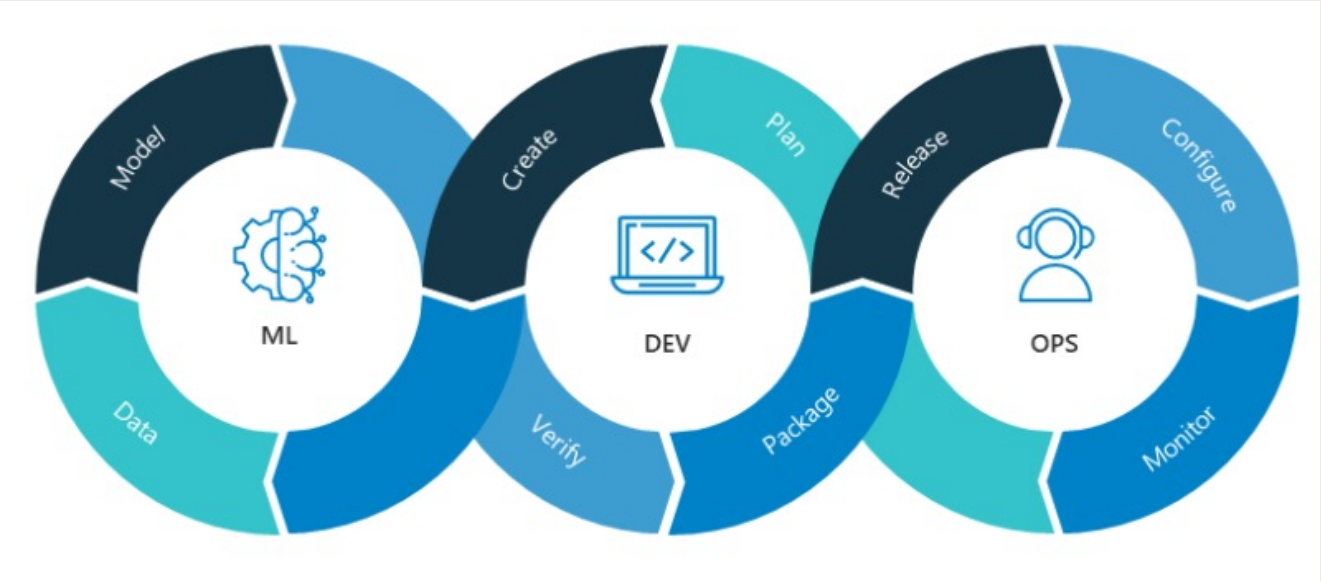


V12 MLOps Fundamentals + Lab Introduction

أجندة العرض Agenda

V12 أساسيات + MLOps تمهيد للمختبر MLOps Fundamentals + Lab Introduction

- **MLOps والأدوات MLOps & Tools**
- **تمييز المفاهيم Concept Distinctions**
- **RAG الوكيلي Agentic RAG**
- **التكلفة والأداء عمليًا Practical Cost & Performance**
- **تمهيد للمختبر Lab Introduction**



MLOps Lifecycle دورة حياة

MLflow لتتبع التجارب والنماذج **Kubeflow**، لأتمتة الخطوط **DVC**، لإصدارات البيانات.

MLflow for tracking, Kubeflow for pipeline automation, DVC for data versioning.

تجربة
Experiment

→

تحقق
Validate

→

نشر
Deploy

→

مراقبة
Monitor

→

إعادة تدريب
Retrain

سجّل الإصدارات والقياسات في كل مرحلة ثم أعد التعلم من الإنتاج

Record versions and metrics at every stage then learn again from production

تمييز المفاهيم Concept Distinctions

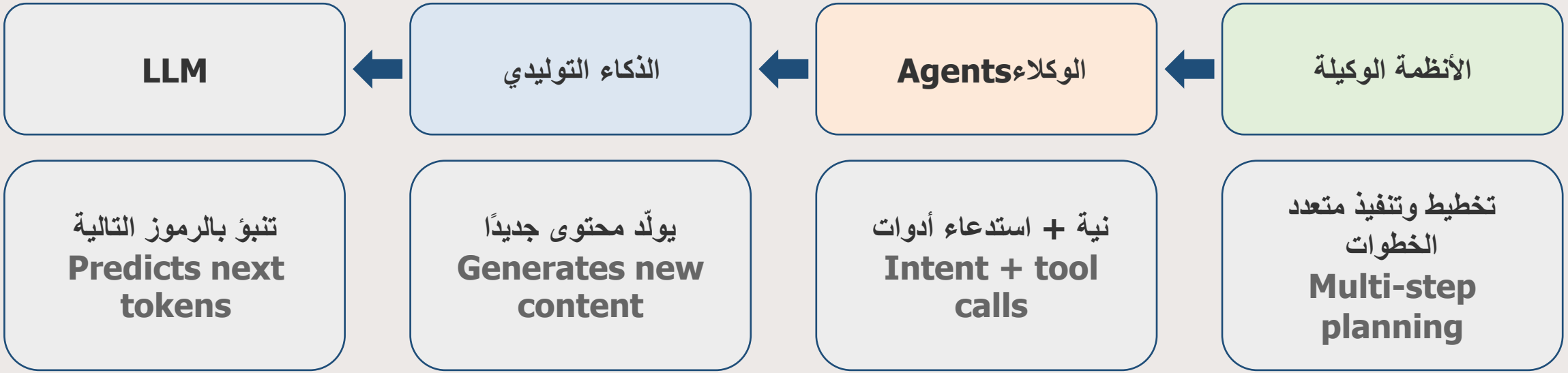
LLM يرمّز النص ويتنبأ بالرموز التالية. الذكاء التوليدي يولّد محتوى جديدًا نص صورة صوت.

LLM predicts next tokens. Generative AI generates new content text image audio.

الوكلاء يكشفون النية ويستدعون الأدوات. الأنظمة الوكيلة تخطيط ذاتي وتنفيذ متعدد الخطوات مع تقييم للنتائج.

Agents detect intent and call tools. Agentic AI self-directed multi-step planning with outcome evaluation.

From LLM to Agentic AI من LLM إلى الأنظمة الوكيلة



Practical Cost & Performance التكلفة والأداء عمليًا

التضمين **Embedding** أقل تكلفة بمرتبة كاملة من التوليد **Generation** الاسترجاع أرخص من التوليد.

Embedding costs an order of magnitude less than generation retrieval is cheaper than generation.

التخزين الدلالي المؤقت يخفض الكمون والتكلفة؛ توجيه الطلبات يرسل المهام البسيطة لنماذج أصغر أرخص.

Semantic caching cuts latency and cost; routing sends easy tasks to smaller, cheaper models.

قاعدة **FinOps** قِس التكلفة لكل طلب **per-request** قبل التوسع.

FinOps rule measure per-request cost before scaling.

Cost Embedding vs Generation التكلفة عمليًا التضمين مقابل التوليد

التضمين x1
Embedding
cheap

التوليد ~x10
Generation
costly

الحل تخزين مؤقت + توجيه
Fix caching +
routing

Lab Architecture معمارية مختبر

خط ETL etl_pipeline.py Bronze SQLite خام ← Silver 50 500 +تقسيم تنظيف ← Gold ChromaDB HNSW cosine،

ETL etl_pipeline.py Bronze raw SQLite → Silver clean + 500 50 chunk → Gold ChromaDB HNSW cosine, batches of 10

الخلفية rag_api.py 8000 استرجاع كثيف top-k=5 ←توليد مؤرَّض 0.2 حد OpenRouter free ← ←تقييم 0.92 كاش دلالي golden

Backend rag_api.py 8000 dense top-k=5 ≥0.2 → grounded OpenRouter free → semantic cache 0.92 → golden eval

الواجهة streamlit_ui.py 8501 5 تُعيد استخدام نفس عميل رفع مستندات استعلام تقييم حول تبويبات ثنائية Chroma مع حذف المجموعة

Frontend streamlit_ui.py 8501 5 bilingual tabs upload docs query eval about same Chroma client, drop & recreate

المخازن lab.db bronze_docs silver_chunks + chroma_db gold_chunks + golden_dataset.json + env. للمفتاح

Stores lab.db bronze_docs silver_chunks + chroma_db gold_chunks + golden_dataset.json + .env for keys

Running the Lab تشغيل مختبر

cd Lab + source .venv bin activate

OMP_NUM_THREADS=1 MKL_NUM_THREADS=1 TOKENIZERS_PARALLELISM=false python -u etl_pipeline.py

./run_ui.sh → http localhost 8501

uvicorn rag_api app --reload → http localhost 8000 docs

python rag_api.py evaluate

Deploy

🎓 L03 — RAG Lab

Educational Edition

Bronze → Silver → Gold → RAG Pipeline

محلي بالكامل / fully local, no Docker

📄 Docs
7

📊 Chunks
784

🧰 Pipeline Config

Generator: `openrouter`

Model: `nemotron-3-super-120`

Embedding: all-MiniLM-L6-v2 (384-dim)

Index: HNSW Cosine

Cache Entries: 0

📚 Educational Resources:

- RAGAS Evaluation
- ChromaDB Docs
- RAG Fundamentals

🎓 Educational Flow | 📤 Upload & Ingest | 📄 Documents | 💬 Query | 📊 Evaluation | ⚙️ About

🎓 RAG Pipeline — Visual Explanation

Explore each step of the Retrieval-Augmented Generation pipeline interactively.

Select a view:

- (●) ETL Overview
- ( ) Embedding Space
- ( ) RAG Pipeline
- ( ) Full Flow

📊 ETL Layers: Bronze → Silver → Gold

- **Bronze** (🗄️): Raw documents stored unmodified in SQLite
- **Silver** (🧹): Documents cleaned, normalized, recursively chunked (500 chars, 50 overlap)
- **Gold** (🥇): Chunks embedded and indexed in ChromaDB with HNSW (cosine distance)

🗄️ Bronze (Raw Documents)	🧹 Silver (Cleaned & Chunked)	🥇 Gold (Indexed Embeddings)
Documents	Chunks	Indexed Vectors
7	0	784

📝 ETL Process Details

🗄️ Bronze Layer

What happens:

- Read TXT, MD, CSV, PDF files

🧹 Silver Layer

What happens:

- Remove exact duplicates

🥇 Gold Layer

What happens:

- Convert chunks to embeddings

Deploy

🎓 L03 — RAG Lab

Educational Edition

Bronze → Silver → Gold → RAG Pipeline

محلي بالكامل / fully local, no Docker

📄 Docs
7

📊 Chunks
784

🧰 Pipeline Config

Generator: `openrouter`

Model: `nemotron-3-super-120`

Embedding: all-MiniLM-L6-v2 (384-dim)

Index: HNSW Cosine

Cache Entries: 0

📚 Educational Resources:

- RAGAS Evaluation
- ChromaDB Docs
- RAG Fundamentals

🎓 Educational Flow | 📤 Upload & Ingest | 📄 Documents | 💬 Query | 📊 Evaluation | ⚙️ About

🎓 RAG Pipeline — Visual Explanation

Explore each step of the Retrieval-Augmented Generation pipeline interactively.

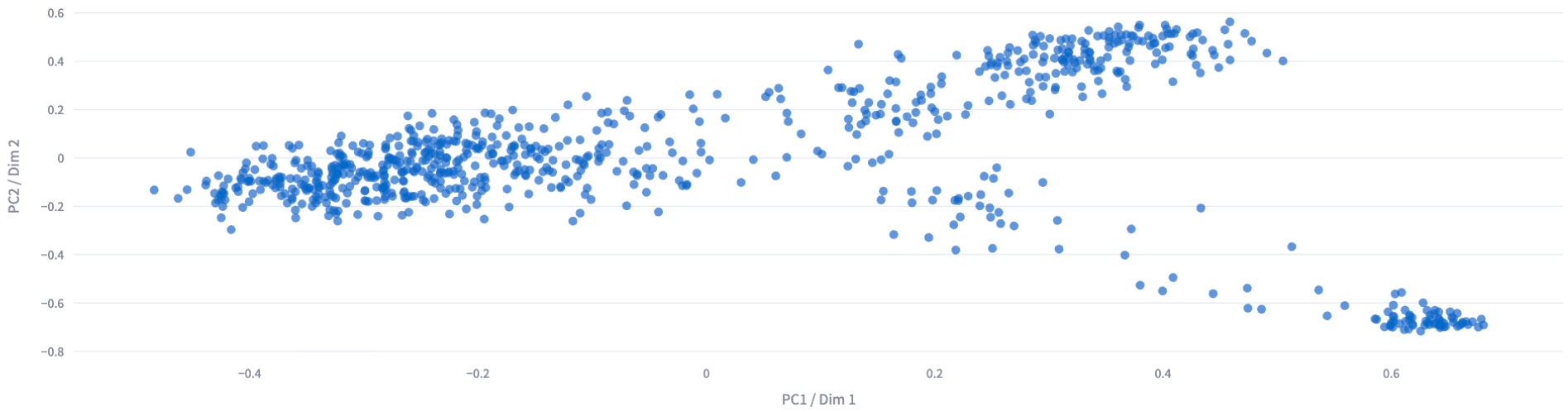
Select a view:

○ ETL Overview ◉ Embedding Space ○ RAG Pipeline ○ Full Flow

🔬 Chunk Embedding Visualization

See how chunks are positioned in embedding space. Similar chunks cluster together.

📊 Chunk Embedding Space (PCA Projection)



📌 784 chunks visualized in 2D space using PCA dimensionality reduction

Deploy ⋮

🎓 L03 — RAG Lab

Educational Edition

Bronze → Silver → Gold → RAG Pipeline

محلي بالكامل / fully local, no Docker

📄 Docs	📊 Chunks
7	784

🧰 Pipeline Config

Generator: `openrouter`

Model: `nemotron-3-super-120`

Embedding: all-MiniLM-L6-v2 (384-dim)

Index: HNSW Cosine

Cache Entries: 0

📚 Educational Resources:

- RAGAS Evaluation
- ChromaDB Docs
- RAG Fundamentals

🎓 Educational Flow | 📤 Upload & Ingest | 📄 Documents | 💬 Query | 📊 Evaluation | ⚙️ About

🎓 RAG Pipeline — Visual Explanation

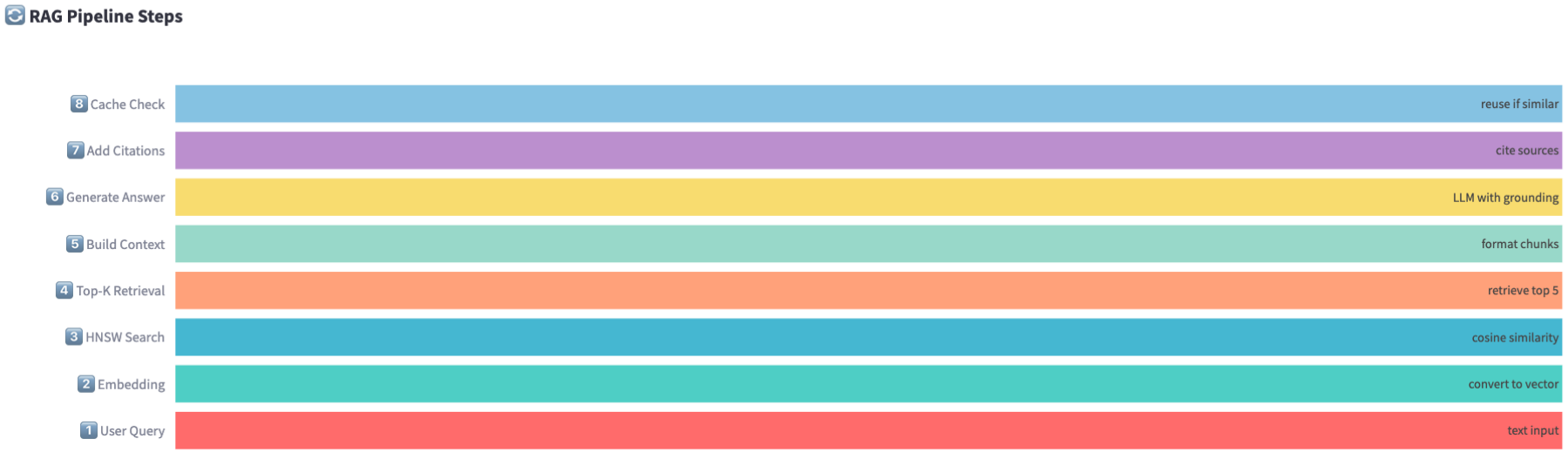
Explore each step of the Retrieval-Augmented Generation pipeline interactively.

Select a view:

○ ETL Overview ○ Embedding Space ◉ RAG Pipeline ○ Full Flow

🔄 RAG Pipeline Steps

From user question to grounded answer with citations.



📋 Step Breakdown

RUNNING... Stop Deploy

🎓 L03 — RAG Lab

Educational Edition

Bronze → Silver → Gold → RAG Pipeline

محلي بالكامل / fully local, no Docker

📄 Docs
7

📊 Chunks
784

🧰 Pipeline Config

Generator: `openrouter`

Model: `nemotron-3-super-120`

Embedding: all-MiniLM-L6-v2 (384-dim)

Index: HNSW Cosine

Cache Entries: 0

📚 Educational Resources:

- RAGAS Evaluation
- ChromaDB Docs
- RAG Fundamentals

🎓 Educational Flow | 📤 Upload & Ingest | 📄 Documents | 💬 Query | 📊 Evaluation | ⚙️ About

📤 Upload & Index a New Document

Upload a document (TXT, MD, CSV, or PDF) to ingest it through the ETL pipeline. The system will:

1. 🗄️ Store raw content in **Bronze**
2. 🧹 Clean and chunk in **Silver**
3. 🥇 Embed and index in **Gold**

Choose a file (TXT/MD/CSV/PDF)

Drag and drop file here
Limit 200MB per file • TXT, MD, CSV, PDF

Browse files

-Trust-Layer-AI-Governance-Platform-BRD-SDD-v0_2.md 27.3KB ×

📄 Selected: `-Trust-Layer-AI-Governance-Platform-BRD-SDD-v0_2.md` (27258 bytes)

⬆️ Upload & Ingest Through Pipeline

🔄 Running ETL pipeline…

Gold: 110/865 مقطعًا / chunks…

Deploy ⋮

🎓 L03 — RAG Lab

Educational Edition

Bronze → Silver → Gold → RAG Pipeline

محلي بالكامل / fully local, no Docker

📄 Docs
7

📊 Chunks
784

🧰 Pipeline Config

Generator: `openrouter`

Model: `nemotron-3-super-120`

Embedding: all-MiniLM-L6-v2 (384-dim)

Index: HNSW Cosine

Cache Entries: 0

📚 Educational Resources:

- RAGAS Evaluation
- ChromaDB Docs
- RAG Fundamentals

🎓 Educational Flow | 📤 Upload & Ingest | 📄 Documents | 💬 Query | 📊 Evaluation | ⚙️ About

📤 Upload & Index a New Document

Upload a document (TXT, MD, CSV, or PDF) to ingest it through the ETL pipeline. The system will:

1. 🗄️ Store raw content in **Bronze**
2. 🧹 Clean and chunk in **Silver**
3. 🥇 Embed and index in **Gold**

Choose a file (TXT/MD/CSV/PDF)

Drag and drop file here
Limit 200MB per file • TXT, MD, CSV, PDF

Browse files

-Trust-Layer-AI-Governance-Platform-BRD-SDD-v0_2.md 27.3KB ×

📄 Selected: `-Trust-Layer-AI-Governance-Platform-BRD-SDD-v0_2.md` (27258 bytes)

⬆️ Upload & Ingest Through Pipeline

✅ **Success!** Document indexed

- 📝 Document: `-Trust-Layer-AI-Governance-Platform-BRD-SDD-v0_2.md`
- 📊 Chunks created: `865`
- 🥇 Indexed in Gold layer

View it in the 📄 **Documents** tab.

Deploy

🎓 L03 — RAG Lab

Educational Edition

Bronze → Silver → Gold → RAG Pipeline

محلي بالكامل / fully local, no Docker

📄 Docs
7

📊 Chunks
784

🧰 Pipeline Config

Generator: `openrouter`

Model: `nemotron-3-super-120`

Embedding: all-MiniLM-L6-v2 (384-dim)

Index: HNSW Cosine

Cache Entries: 0

📚 Educational Resources:

- RAGAS Evaluation
- ChromaDB Docs
- RAG Fundamentals

🎓 Educational Flow | 📤 Upload & Ingest | 📄 Documents | 💬 Query | 📊 Evaluation | ⚙️ About

📄 Indexed Documents

📚 8 Document(s) Indexed

📄 Document	📊 Chunks	Avg Size
-Global-Data-Privacy-Program-BRD-SDD-v1_3.md	165	~500 chars
-Trust-Layer-AI-Governance-Platform-BRD-SDD-v0_2.md	81	~500 chars
AEGIS_AI_ConceptPaper.pdf	197	~500 chars
AEGIS_AI_Presentation.pdf	22	~500 chars
Egypt-PDPL-self-assessment-v2_14.pdf	290	~500 chars
Introduction to Quantum mechanics.pdf	107	~500 chars
etl_notes.txt	1	~500 chars
rag_basics.md	2	~500 chars

🔍 Inspect Chunks

Select a document:

-Global-Data-Privacy-Program-BRD-SDD-v1_3.md

📋 **165 chunks** from `-Global-Data-Privacy-Program-BRD-SDD-v1_3.md`

📍 Chunk #0 — 0d7f13bc-1c8…

Global Data Privacy Program

Length
364 chars

ID: 0d7f13bc-1c8c-429e-9…

Deploy

🎓 L03 — RAG Lab

Educational Edition

Bronze → Silver → Gold → RAG Pipeline

محلي بالكامل / fully local, no Docker

📄 Docs
8

📊 Chunks
865

🧰 Pipeline Config

Generator: `openrouter`

Model: `nemotron-3-super-120`

Embedding: all-MiniLM-L6-v2 (384-dim)

Index: HNSW Cosine

Cache Entries: 0

📚 Educational Resources:

- RAGAS Evaluation
- ChromaDB Docs
- RAG Fundamentals

🔍 Inspect Chunks

Select a document:

rag_basics.md

📋 **2 chunks** from `rag_basics.md`

📍 Chunk #0 — 2da02d7f-aee…

Length
449 chars

ID: 2da02d7f-aeed-47f4-a…

RAG Basics

Retrieval-Augmented Generation (RAG) combines a retrieval step with a generative LLM. Instead of relying only on what the model memorized during training, RAG fetches relevant documents from a vector index and injects them into the prompt.

Chunking

Recursive chunking is the recommended default: split by paragraphs first, then split any oversized chunk by sentences. Aim for 300-800 characters with 10-20% overlap.

Evaluation

📍 Chunk #1 — cc03aa26-ca9…

Length
173 chars

ID: cc03aa26-ca98-4988-9…

0 characters with 10-20% overlap.

Evaluation

The RAG triad measures faithfulness, answer relevancy, and context precision. The RAGAS framework implements these metrics.

Deploy

🎓 L03 — RAG Lab

Educational Edition

Bronze → Silver → Gold → RAG Pipeline

محلي بالكامل / fully local, no Docker

📄 Docs
8

📊 Chunks
865

🧰 Pipeline Config

Generator: `openrouter`

Model: `nemotron-3-super-120`

Embedding: all-MiniLM-L6-v2 (384-dim)

Index: HNSW Cosine

Cache Entries: 0

📚 Educational Resources:

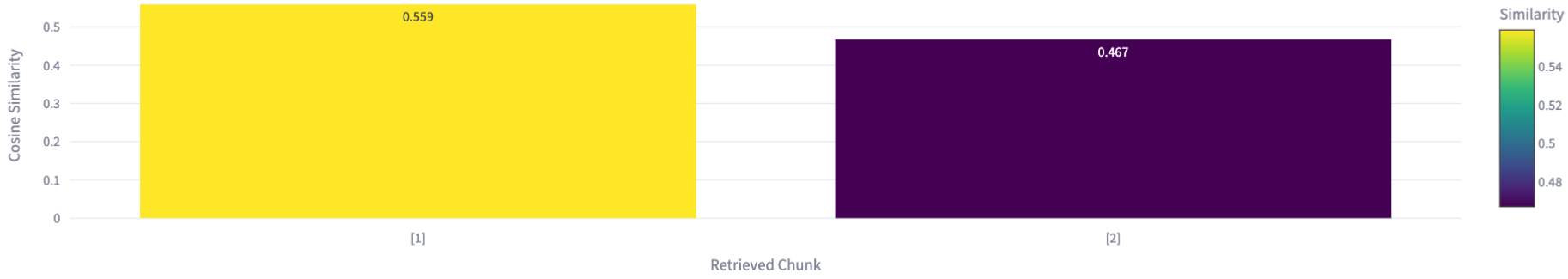
- RAGAS Evaluation
- ChromaDB Docs
- RAG Fundamentals

📍 **Step 1: Retrieval** — Finding relevant chunks…

✅ Retrieved 2 chunks

📊 Retrieval Similarity Scores

🔍 **Retrieval Similarity Scores**



🤖 **Step 2: Generation** — Creating grounded answer…

✅ Answer generated

🎯 Answer

Retrieval-Augmented Generation (RAG) combines a retrieval step with a generative LLM, fetching relevant documents from a vector index and injecting them into the prompt to supplement the model's knowledge. [2]

💾 From semantic cache (92%+ similarity)

📚 Retrieved Sources

Deploy

🎓 L03 — RAG Lab

Educational Edition

Bronze → Silver → Gold → RAG Pipeline

محلي بالكامل / fully local, no Docker

📄 Docs
8

📊 Chunks
865

🧰 Pipeline Config

Generator: `openrouter`

Model: `nemotron-3-super-120`

Embedding: all-MiniLM-L6-v2 (384-dim)

Index: HNSW Cosine

Cache Entries: 1

📚 Educational Resources:

- RAGAS Evaluation
- ChromaDB Docs
- RAG Fundamentals

🎯 Answer

Retrieval-Augmented Generation (RAG) combines a retrieval step with a generative LLM. [2]

💾 From semantic cache (92%+ similarity)

📚 Retrieved Sources

[1] from `rag_basics.md`

Similarity
0.559

View chunk text

0 characters with 10-20% overlap.

Evaluation

The RAG triad measures faithfulness, answer relevancy, and context precision. The RAGAS framework implements these metrics.

[2] from `rag_basics.md`

Similarity
0.467

View chunk text

RAG Basics

Retrieval-Augmented Generation (RAG) combines a retrieval step with a generative LLM. Instead of relying only on what the model memorized during training, RAG fetches relevant documents from a vector index and injects them into the prompt.

Chunking

Recursive chunking is the recommended default: split by paragraphs first, then split any oversized chunk by sentences. Aim for 300-800 characters with 10-20% overlap.

Evaluation

Deploy

🎓 L03 — RAG Lab

Educational Edition

Bronze → Silver → Gold → RAG Pipeline

محلي بالكامل / fully local, no Docker

📄 Docs
8

📊 Chunks
865

🧰 Pipeline Config

Generator: `openrouter`

Model: `nemotron-3-super-120`

Embedding: all-MiniLM-L6-v2 (384-dim)

Index: HNSW Cosine

Cache Entries: 1

📚 Educational Resources:

- RAGAS Evaluation
- ChromaDB Docs
- RAG Fundamentals

🎓 Educational Flow | 📤 Upload & Ingest | 📄 Documents | 💬 Query | 📊 Evaluation | ⚙️ About

📊 Evaluation Against Golden Dataset

Benchmark your RAG system against 12 reference questions.

Metrics:

- **Context Precision** ≈ average similarity of retrieved chunks
- **Expected Source Found** = whether the benchmark's expected source appeared in results

Golden Dataset:

- 12 bilingual Q&A pairs
- AR/EN questions and references
- Expected source document name

▶️ Run Evaluation

📈 Results

#	سؤال / Question	Precision	Source Found
1	ما الفرق بين ETL وELT؟...	0.3	❌
2	ما طبقات المعمارية المتدرِّجة Medallion؟...	0.47	✅
3	ما مكونات عقد البيانات الثلاثة؟...	0.47	❌
4	ما الاستراتيجية الموصى بها افتراضيًا لتق...	0.36	✅
5	لماذا نستخدم البحث الهجين في الاسترجاع؟...	0.34	❌
6	ما وظيفة إعادة الترتيب Cross-encoder؟...	0.34	❌
7	ما مكونات مثلث تقييم RAG الثلاثة؟...	0.42	✅
8	كيف يقلل التخزين الدلالي المؤقت التكلفة؟...	0.23	❌
9	ما الفرق بين AI Agent وAgentic AI؟...	0.46	❌
10	ما طبقات المعمارية المرجعية للذكاء التول...	0.48	❌

Glossary

خريطة المصطلحات Glossary Map

البيانات والتخزين
Data & storage

الخصائص
Features

RAGوالاسترجاع
RAG & retrieval

تعلم المصطلحات كمجموعات مترابطة
Learn terms as connected groups

الحوكمة
Governance

MLOps
MLOps

Glossary1 المصطلحات

Data Pipeline	خط مسارات البيانات
ETL Extract, Transform, Load	الاستخراج والتحويل والتحميل
ELT	الاستخراج والتحميل والتحويل
Medallion Architecture	المعمارية المتدرِّجة
Data Warehouse	مستودع البيانات
Data Lake	بحيرة البيانات

Glossary2 المصطلحات

Lakehouse	معمارية بحيرة البيانات الموحدة
Data Mesh	شبكة البيانات
Data Product	منتج البيانات
Data Contract	عقد البيانات
Data Lineage	سلسلة العهدة
Data Catalog	فهرس البيانات الكتالوج

Glossary3 المصطلحات

Feature Engineering	هندسة الخصائص
Feature Store	مخزن الخصائص
Feature Skew	انحراف الميزات
Vector Database	قاعدة البيانات المتجهية
Embedding	التمثيل المتجهي
ANN Approximate Nearest Neighbor	البحث التقريبي لأقرب الجيران

Glossary4 المصطلحات

HNSW Index	فهرس HNSW
Recursive Chunking	التقسيم التكراري
Semantic Chunking	التقسيم الدلالي
Hybrid Search	البحث الهجين
Dense Sparse Retrieval	البحث الكثيف المتناثر
Reranking	إعادة الترتيب

Glossary5 المصطلحات

Grounded Prompting	الموجِّه المقيَّد بالمصادر
RAG Retrieval-Augmented Generation	التوليد المعزَّز بالاسترجاع
Agentic RAG	RAGالوكيلي
Faithfulness	الوفاء بالمصدر
Answer Relevancy	ملاءمة الإجابة
Context Precision	دقة السياق

Glossary6 المصطلحات

Semantic Caching	التخزين الدلالي المؤقت
Single ChromaDB client no SQLite lock	عميل ChromaDB واحد لا قفل SQLite
Drop & recreate collection	حذف المجموعة وإعادة إنشائها
Batched encoding 10	الترميز بدفعات 10
Model Routing	توجيه الطلبات بين النماذج
LLM-assisted Labeling	التصنيف بمساعدة النموذج اللغوي

Glossary7 المصطلحات

Human-in-the-loop	التحقق البشري
AI Gateway	بوابة الذكاء الاصطناعي
Prompt Library	مكتبة الموجِّهات
Guardrails	حواجز الحماية
Responsible AI	الذكاء الاصطناعي المسؤول
FinOps	إدارة التكاليف السحابية

Glossary8 المصطلحات

Experiment Tracking	تتبع التجارب
Data Versioning	إصدارات البيانات