Operational AI Means Database

Torsten Steinbach

VP, Chief Architect Analytics & AI

Gianni Ciolli

VP, Practice Lead High Availability

EDB™

Torsten Steinbach

Vice President
Chief Architect Analytics & AI

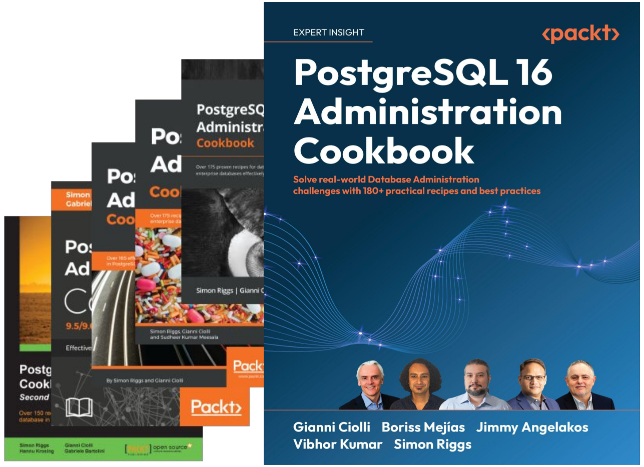
Product Architect for RDBMS, DWHs & Data Lakes

Db2, Netezza, Hadoop, Spark

20+ years speaker at DB and Analytics Conferences

Delivered Cloud Data Lake and Lakehouse aaS products

Expert for Generative AI and Vector Databases


EXPERT INSIGHT
‹packt›
PostgreSQL 16 Administration Cookbook
Solve real-world Database Administration challenges with 180+ practical recipes and best practices
Gianni Ciolli Boriss Mejías Jimmy Angelakos
Vibhor Kumar Simon Riggs
Packt›

Gianni Ciolli

Global Vice President
Practice Lead, HA

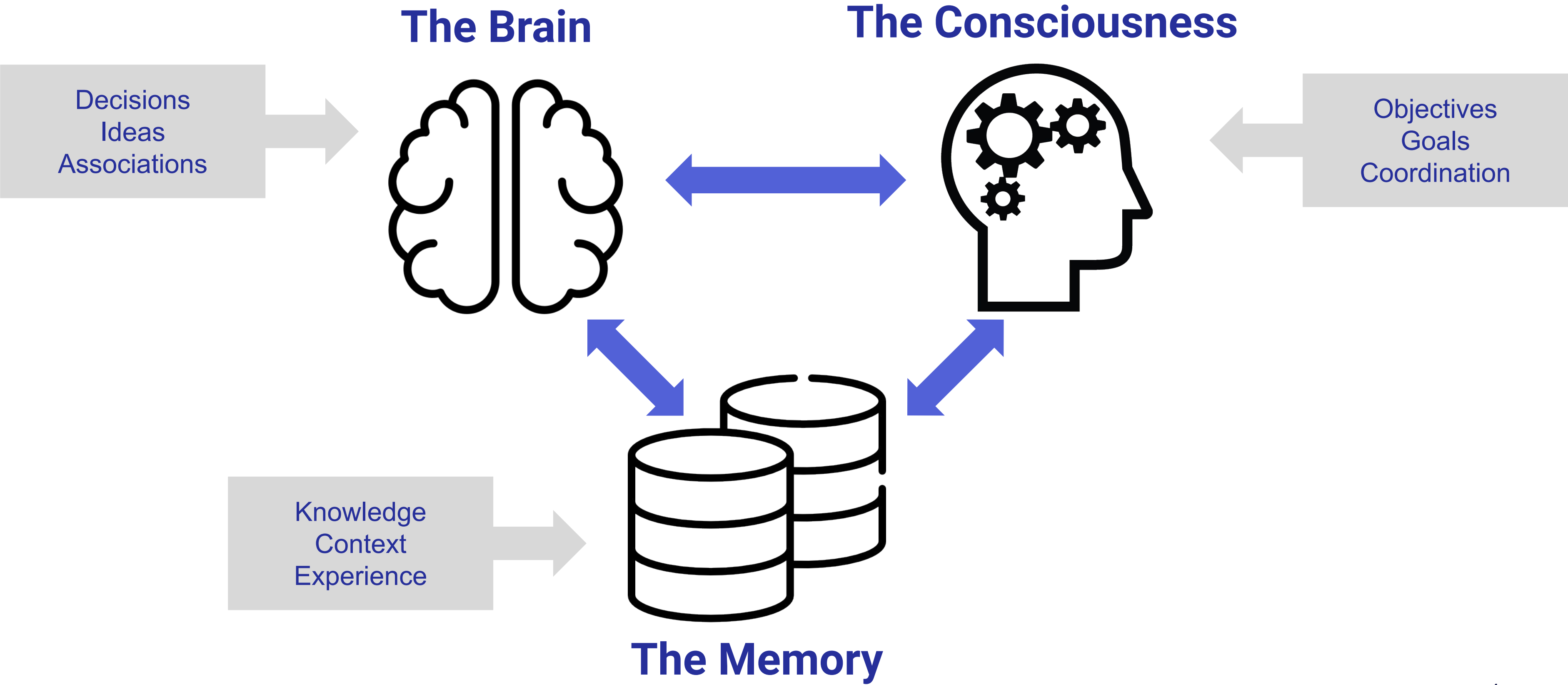
PostgreSQL consultant, developer, contributor

Speaker and trainer at various PostgreSQL conferences for the past 15 years

Active member of the OS community since the 1990s

PostgreSQL Administration Cookbook author

Intelligence



The Brain

Decisions
Ideas
Associations

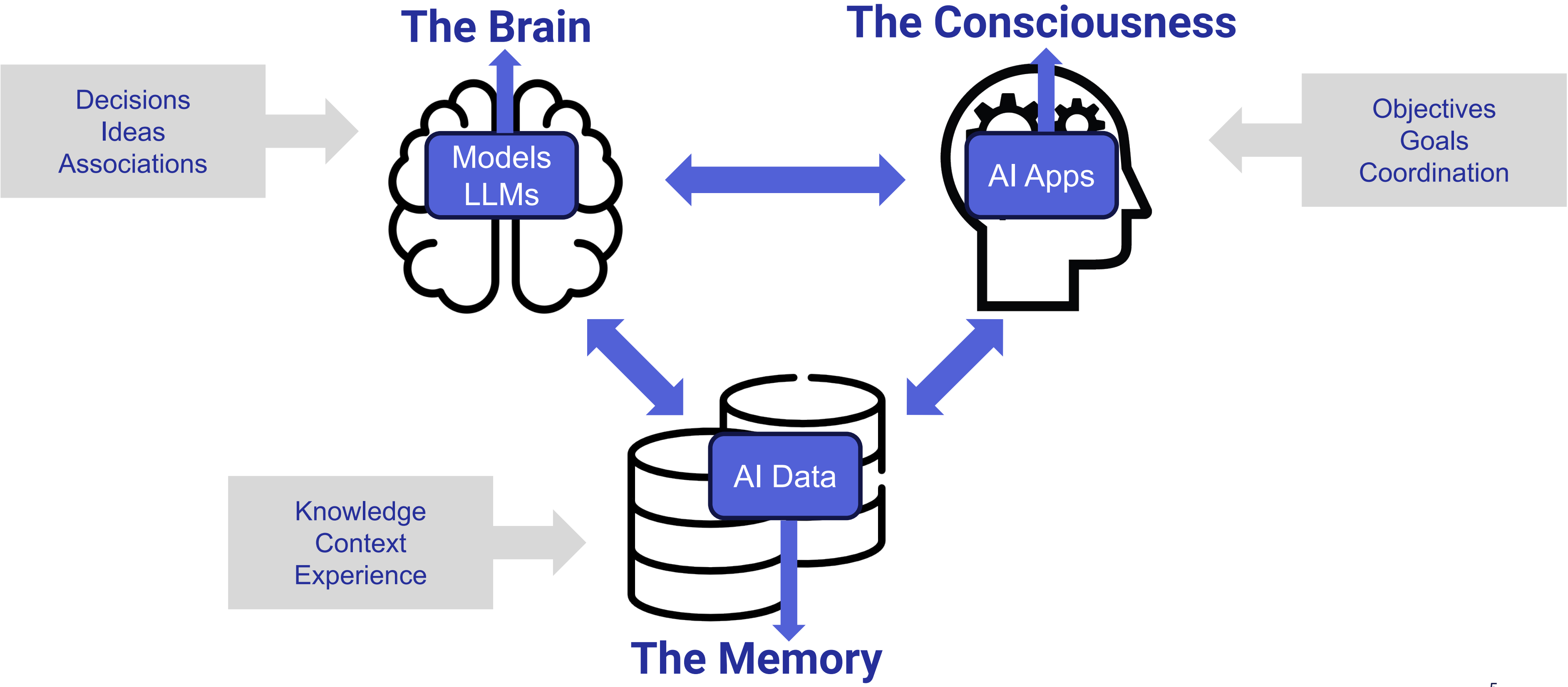
The Consciousness

Objectives
Goals
Coordination

The Memory

Knowledge
Context
Experience

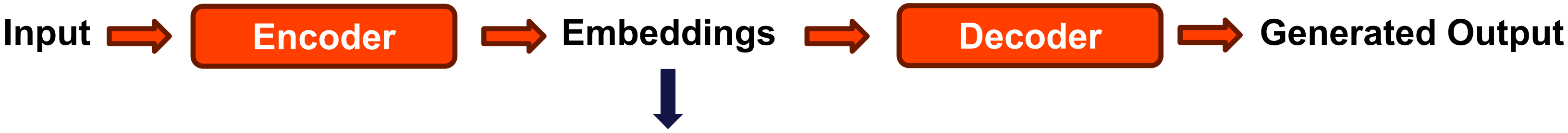
Artificial Intelligence



Vector Embeddings – The Secret Sauce of Generative AI

Generative AI relies on a **modern** form of **Deep Learning** architecture:

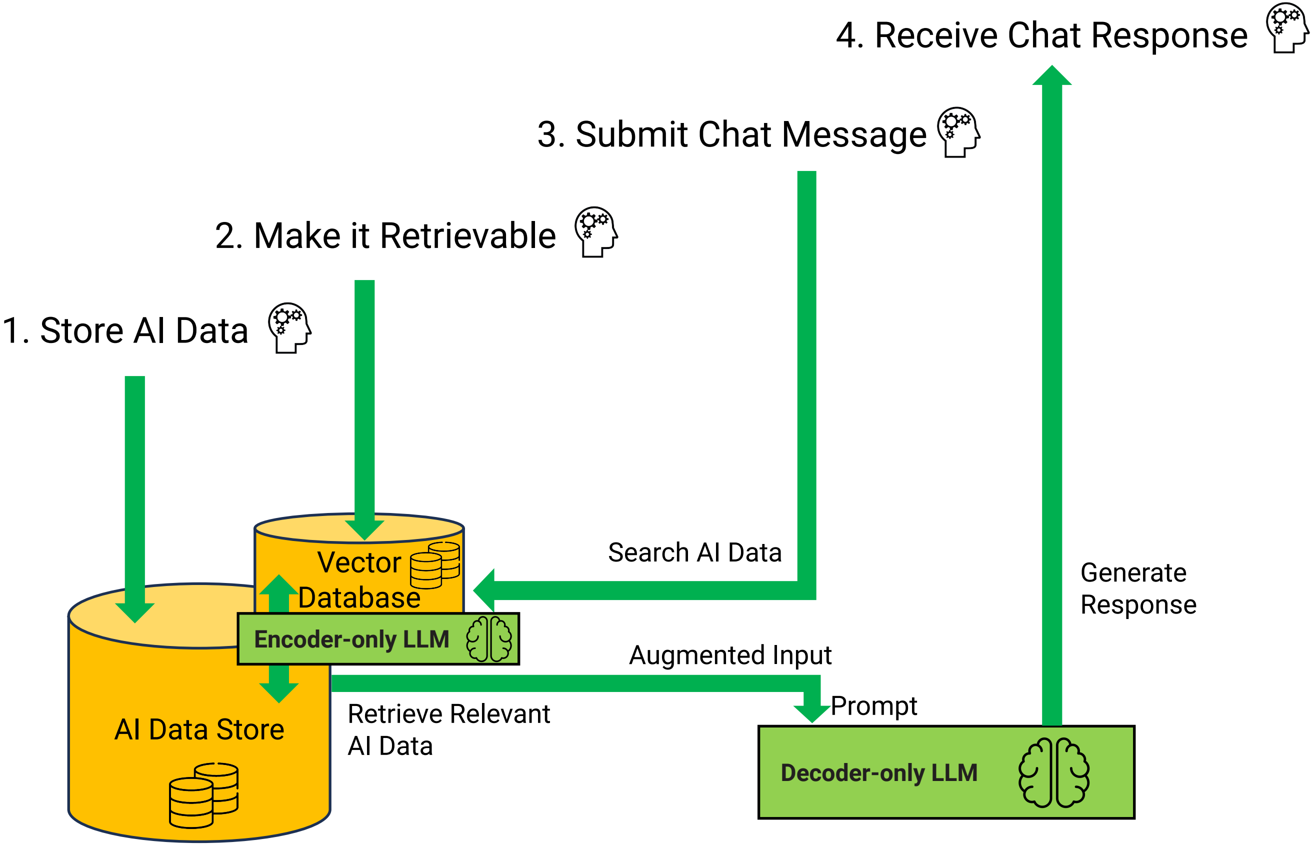
- The **Transformer Architecture**



- The **numeric representation** of the **semantics** of the input data
- Enables **cross** data **modality** encoding-decoding such as Image—2—text
- Need to be **stored and managed** just like the AI data itself

- **Popular Variants:**
  - **Encoder-Decoder** Models, e.g., Google Translate, Summarization
  - **Encoder-only** Models, e.g., Classification (sentiment, language etc.), Vector Embeddings
  - **Decoder-only** Models, e.g., Question answering, such as GPT or Llama2

Chat Bots – The John Doe of Gen AI Applications



A.k.a.:
Retrieval
Augmented
Generation
= **RAG**

OK, you've built a Chat Bot. Now What?

People love it…

You get more and more users…

People bring more data and more use cases…

People start using it routinely…

Your chat bot becomes **mission critical** for the business!

Are you **ready for that??**

Do you have a plan for **sustainable operationalization??**

Operational Qualities of Service

These are table stakes for Enterprise solutions

Always On

- No planned downtime
- In-Place Updates

Resilient

- HA & DR
- SLAs for RPO & RTO
- Ideally 0 for both

Secure

- State-of-the-art auth methods
- Fine-grained access control
- Data encryption
- Data governance

Responsive

- SLA for interactive queries
- Automatic index maintenance
- SLA for query throughput

Scalable

- Flexible scale-up and -out
- Elasticity, scale-in

Enterprise Support

- First response time and turnaround time SLAs
- Back-level SW support

Ecosystem

- Stable, well-established APIs
- Large and vital user community

Business Data

- System is not an island
- Integrates with existing data
- Hybrid Search

Business Events

- System is not frozen in time
- Business transactions are digested in real-time

Operational Qualities of Service for Postgres

These are table stakes for Enterprise solutions

Always On

- No planned downtime
- In-Place Updates

- EDB Postgres Distributed has **Rolling Upgrades**:
  - including **major versions**
  - and **in place**

Operational Qualities of Service for Postgres

These are table stakes for Enterprise solutions

Resilient

- HA & DR
- SLAs for RPO & RTO
- Ideally 0 for both

- PostgreSQL has several **HA & DR** features
  - Point-In-Time Recovery
  - Hot Standby
  - Synchronous Replication

- EDB Postgres Distributed gives **better RPO & RTO**
  - Active-Active with robust HA
  - Commit Scopes

Operational Qualities of Service for Postgres

These are table stakes for Enterprise solutions

- Postgres is widely used for **transactional workloads** in challenging scenarios
- Fully **ACID** compliant
- Highly concurrent

Responsive

- SLA for interactive queries
- Automatic index maintenance
- SLA for query throughput

Business Events

- System is not frozen in time
- Business transactions are digested in real-time

Operational Qualities of Service for Postgres

These are table stakes for Enterprise solutions

- Users can purchase **first class** Postgres **support**
- Postgres skills are a safe career choice
- **All options** widely available
  - on-premises, public cloud, private cloud
  - either managed or not

Enterprise Support

- First response time and turnaround time SLAs
- Back-level SW support

Ecosystem

- Stable, well-established APIs
- Large and vital user community

Operational Qualities of Service for Postgres

These are table stakes for Enterprise solutions

- Hybrid search extracts value from your existing business data
- PostgreSQL has sophisticated and compliant relational capabilities
- Existing Full Text Search capabilities combined with relational search and similarity search via pgvector

Business Data

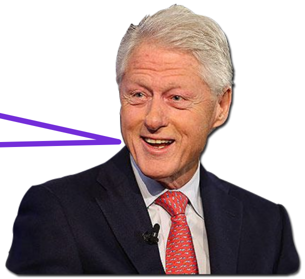
- System is not an island
- Integrates with existing data
- Hybrid Search

AI Application Frameworks

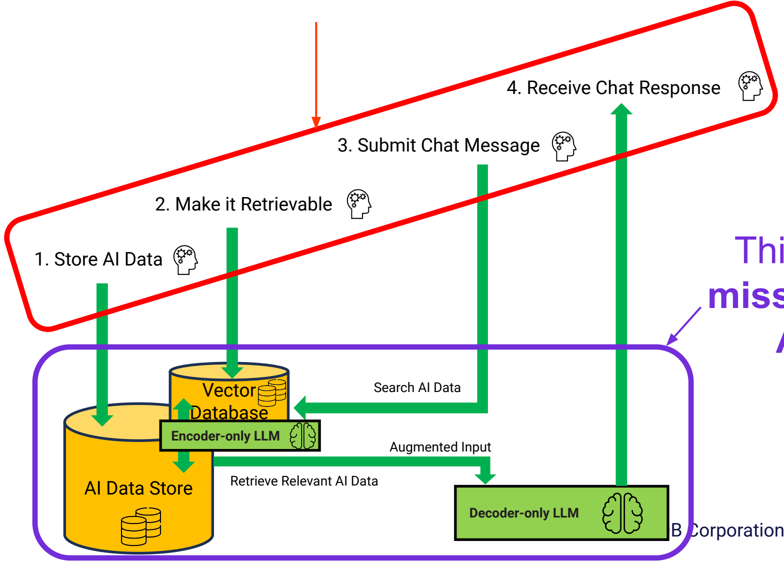
- **Current standard** method to **simplify AI** solution **development**
    - Orchestrating & automating complex AI application flows
- Growing number of such AI frameworks is emerging:
    - **LangChain** - General purpose LLM solutions
    - **LlamaIndex** - LLM-based search & retrieval solutions
    - **DSPy** - Automating LLM prompt engineering
- **Hide** a lot of AI processing **complexity** & **rapid** solution **development**
- Is **NOT** a **data management** framework:
    - Data plays a role only as connected data sources
    - You must manage own data storage, repositories and databases

Do LangChain & friends help to

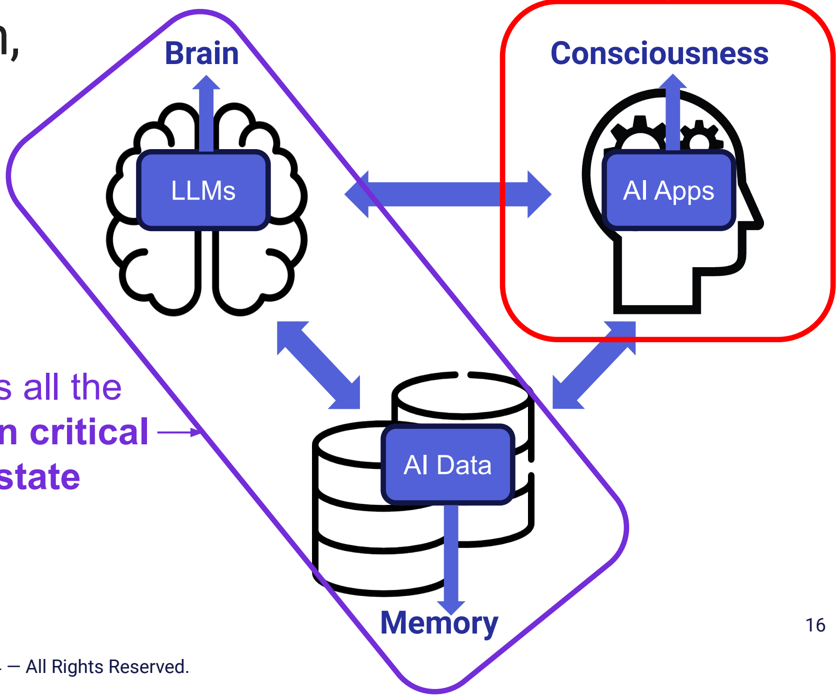
It's The Data, Stupid!



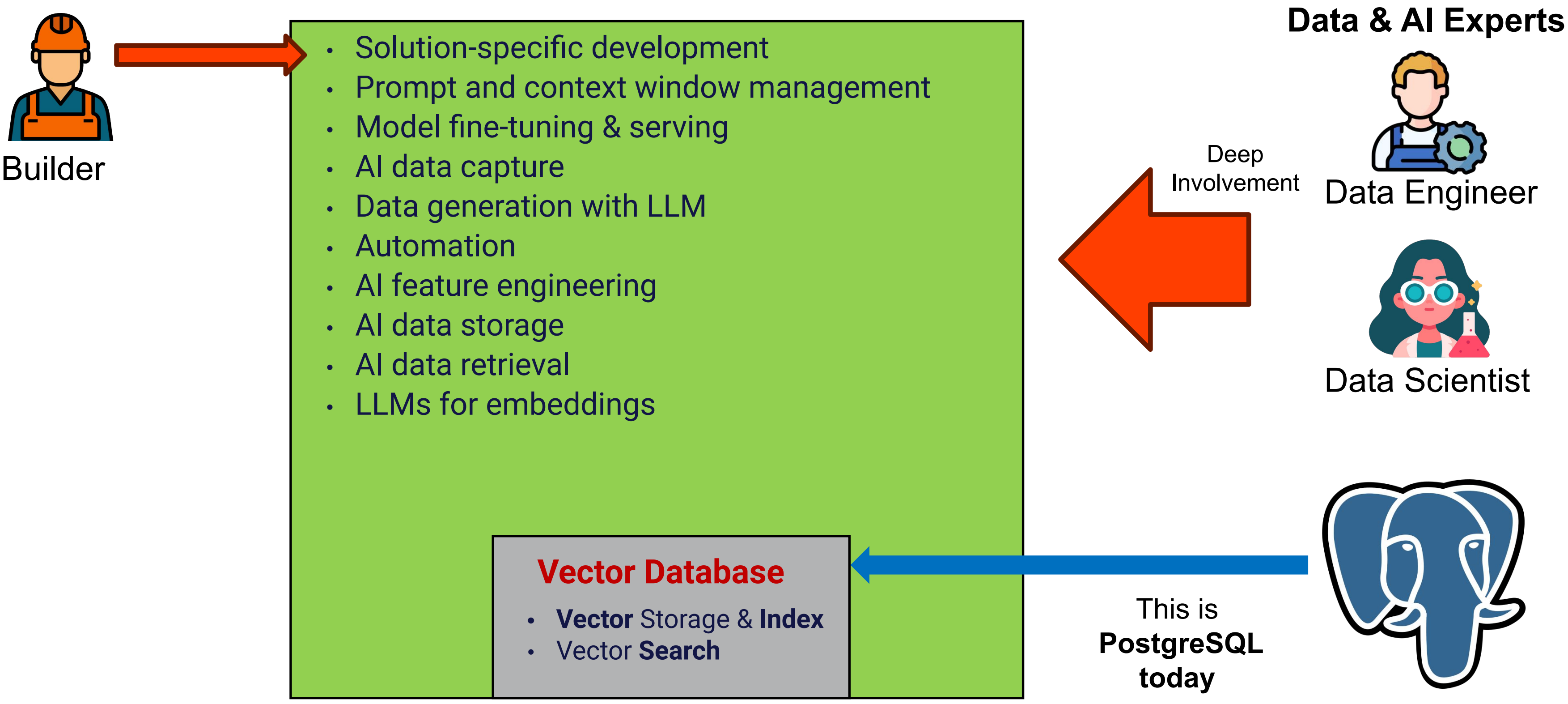
- **Not really!** They simplify building the AI application only
- They don't operationalize the solution, they only automate your app flow



This is all the **mission critical AI state**



Building AI Solutions with Vector Database



Builder

- Solution-specific development
- Prompt and context window management
- Model fine-tuning & serving
- AI data capture
- Data generation with LLM
- Automation
- AI feature engineering
- AI data storage
- AI data retrieval
- LLMs for embeddings

Vector Database

- **Vector** Storage & **Index**
- Vector **Search**

Data & AI Experts

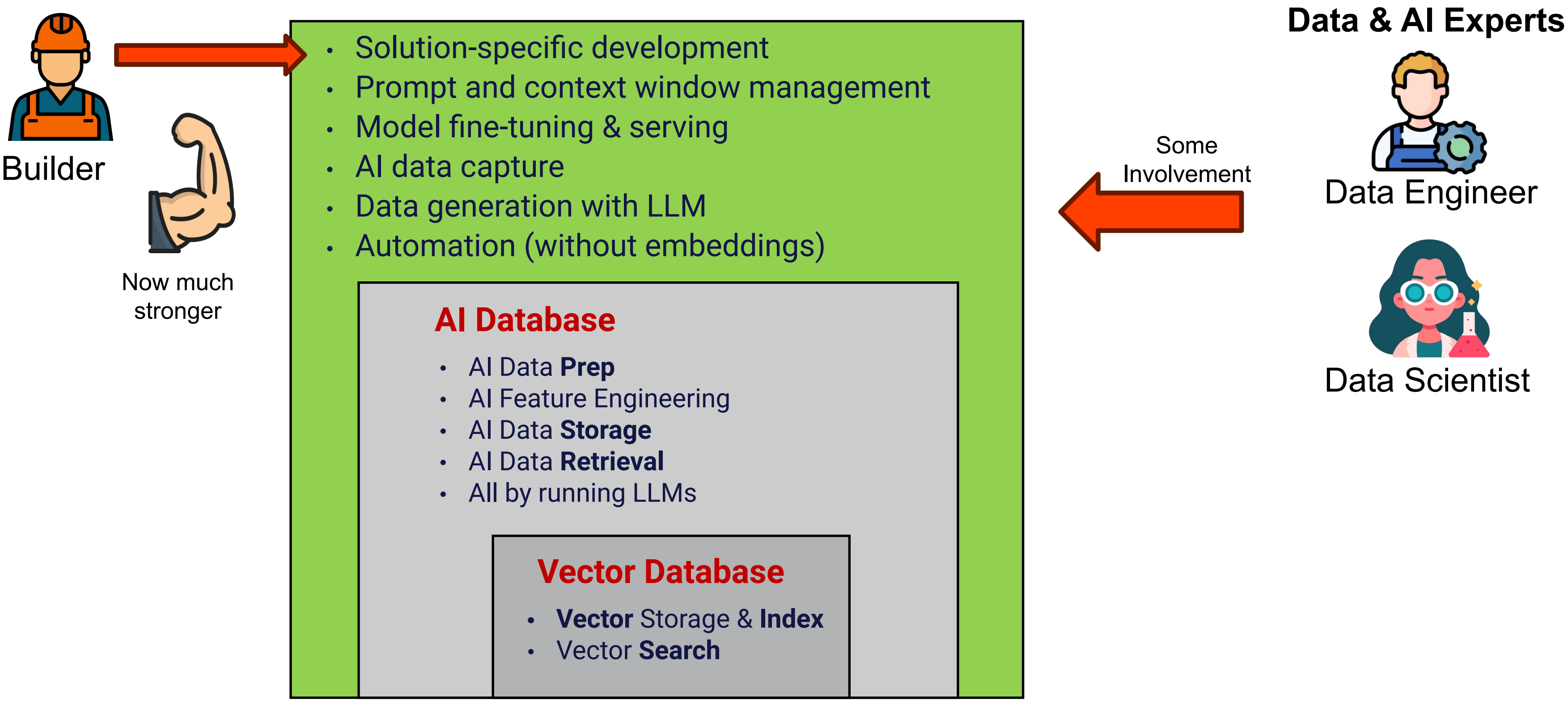
Deep Involvement

Data Engineer

Data Scientist

This is **PostgreSQL today**

Building AI Solutions with an AI Database



Builder

Now much stronger

- Solution-specific development
- Prompt and context window management
- Model fine-tuning & serving
- AI data capture
- Data generation with LLM
- Automation (without embeddings)

AI Database

- AI Data **Prep**
- AI Feature Engineering
- AI Data **Storage**
- AI Data **Retrieval**
- All by running LLMs

Vector Database

- **Vector** Storage & **Index**
- Vector **Search**

Some Involvement

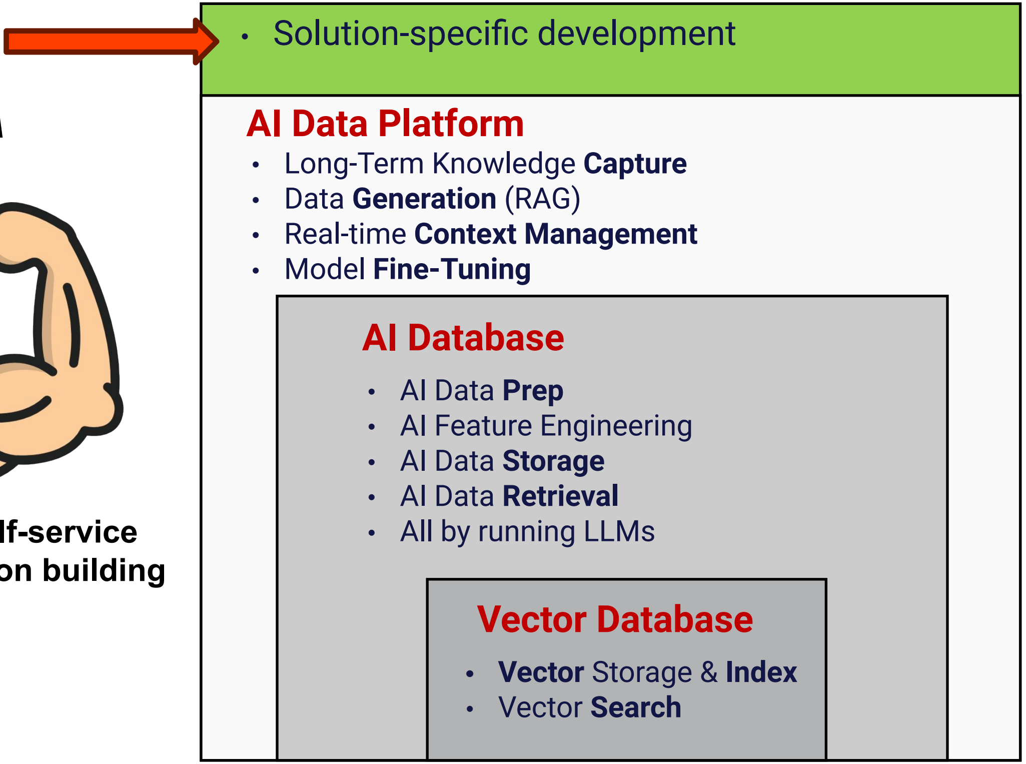
Data & AI Experts

Data Engineer

Data Scientist

Building AI Solutions with an AI Data Platform

Builder



- Solution-specific development

AI Data Platform

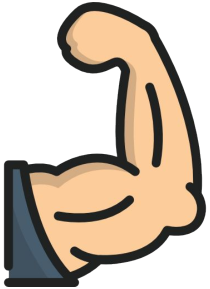
- Long-Term Knowledge **Capture**
- Data **Generation** (RAG)
- Real-time **Context Management**
- Model **Fine-Tuning**

AI Database

- AI Data **Prep**
- AI Feature Engineering
- AI Data **Storage**
- AI Data **Retrieval**
- All by running LLMs

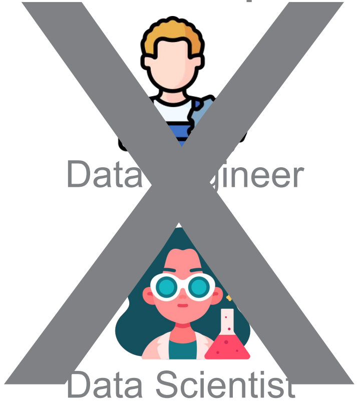
Vector Database

- **Vector** Storage & **Index**
- Vector **Search**



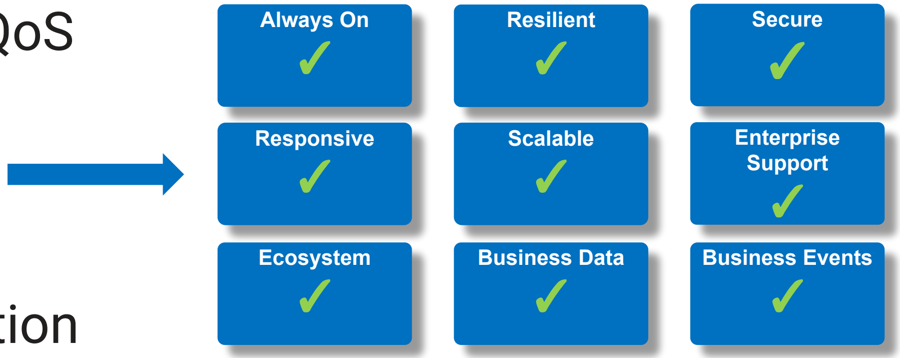
Full self-service AI solution building

Data & AI Experts



Postgres is perfectly positioned to become THE AI database

- Absolute **battle proof** Enterprise QoS
  - In **community** distro but also **very vital commercial** Enterprise option ecosystem

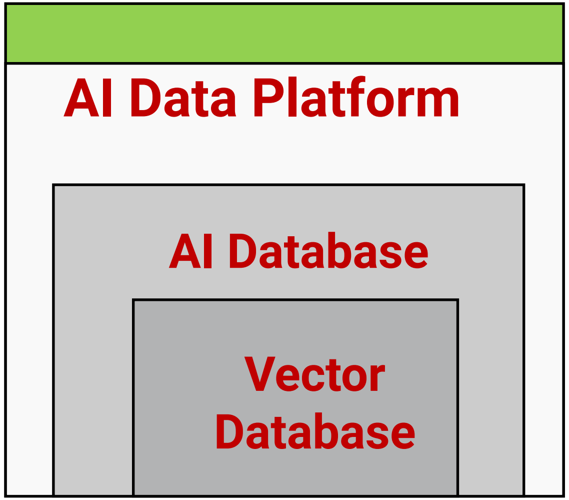


Always On ✓	Resilient ✓	Secure ✓
Responsive ✓	Scalable ✓	Enterprise Support ✓
Ecosystem ✓	Business Data ✓	Business Events ✓

- Perfect **extensibility** & customization
  - With **AI relevant** languages & ecosystems: **Python, Rust**
  - **Custom Data** Types
  - Index & Table **Access Methods**
- **Already** houses the most valuable enterprise **business data**
  - in **fully relational** manner

Postgres Graduating to AI Database and AI Platform

- PG developers becoming the **strongest AI builders**
- Integrating **new AI data storage** (Lakehouse style)
- Integrating **new AI data processing**
- **Stay tuned and watch out for EDB here!**

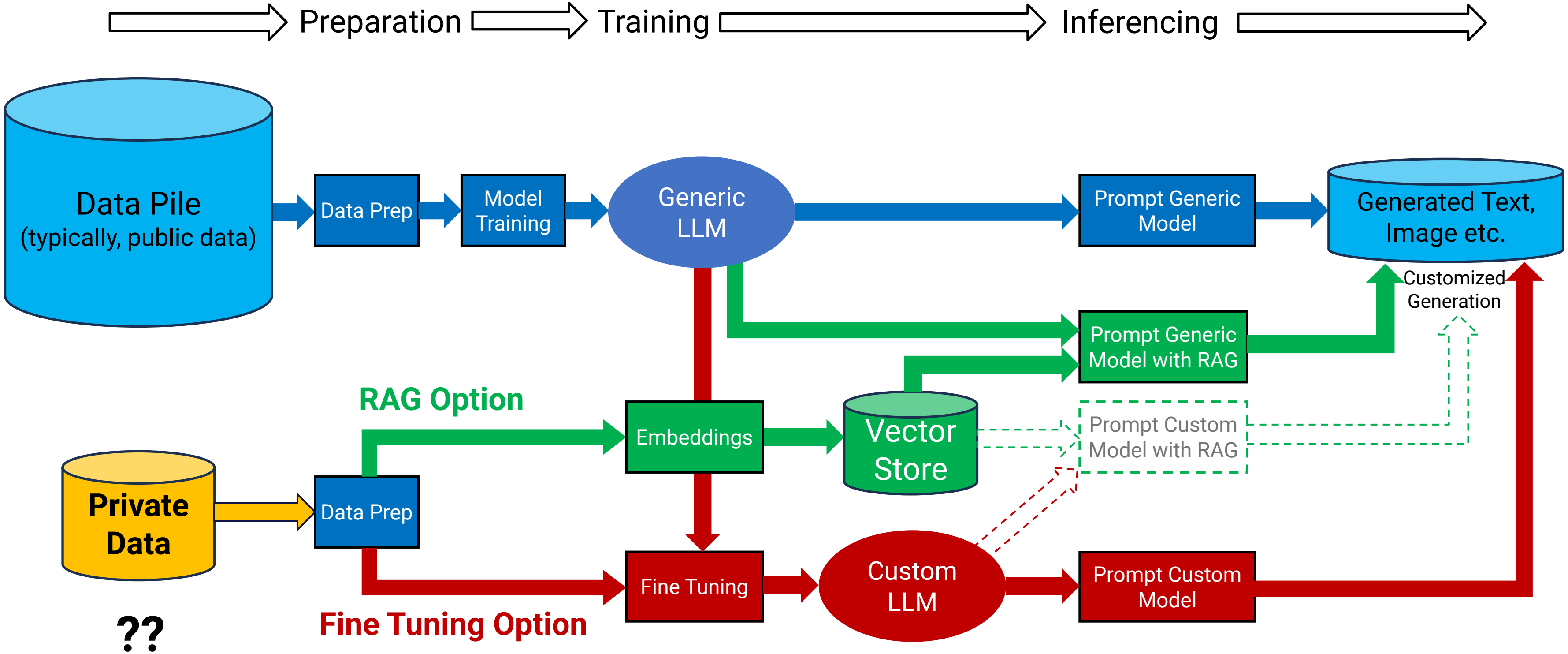


Thank You

Backup

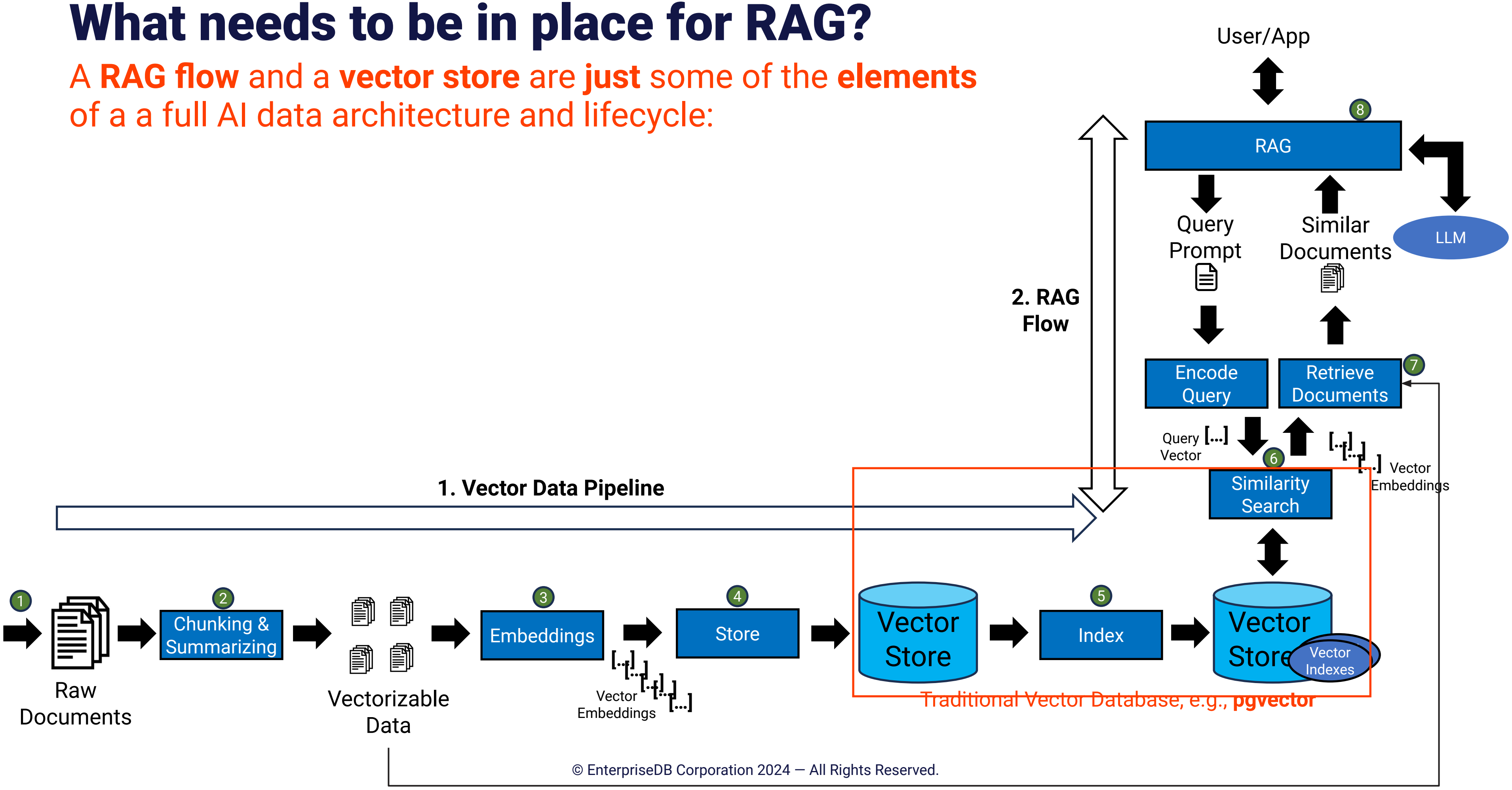
Generative AI – Big Picture

And: How to apply GenAI to your private data? – **Fine-tuning vs. RAG**

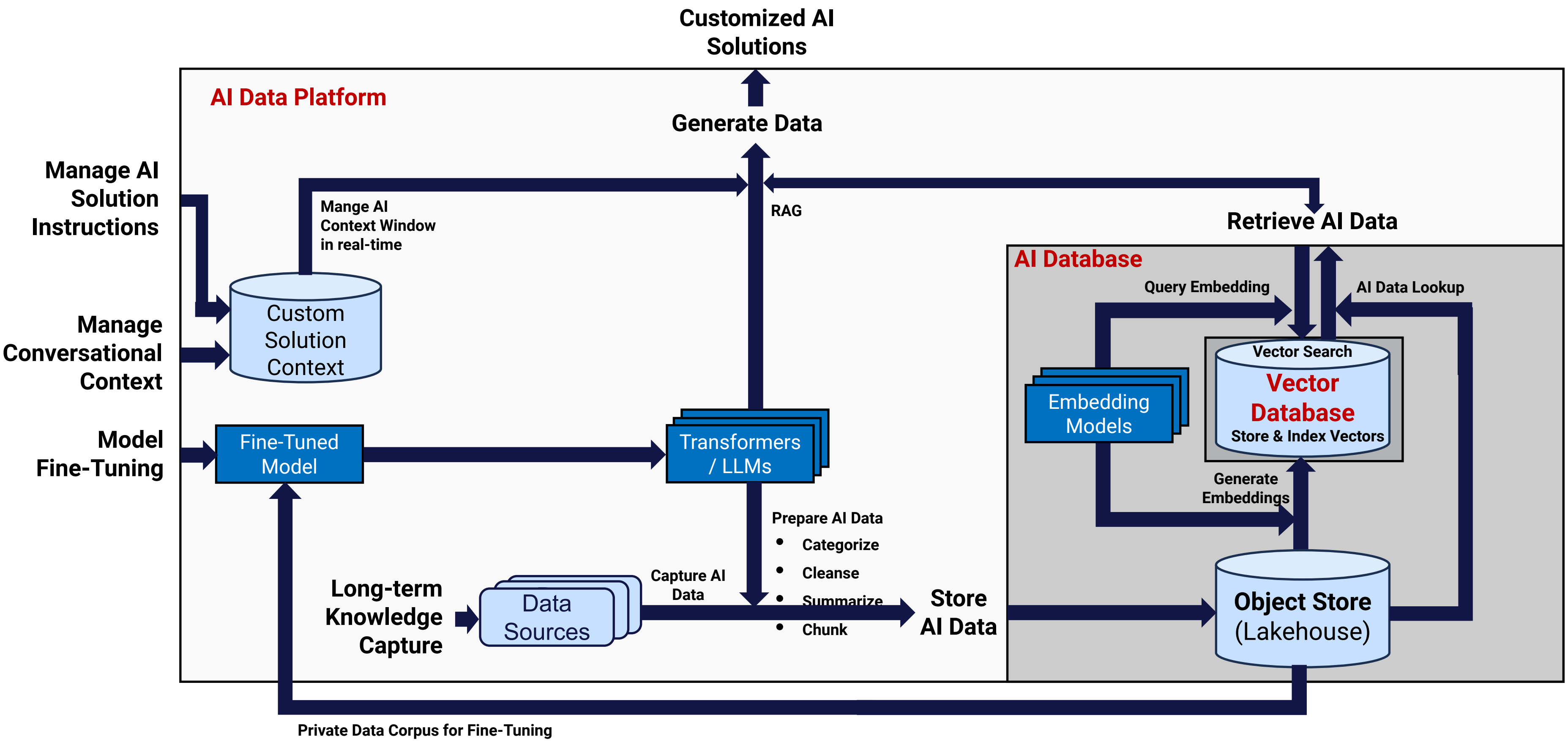


What needs to be in place for RAG?

A **RAG flow** and a **vector store** are **just** some of the **elements** of a a full AI data architecture and lifecycle:

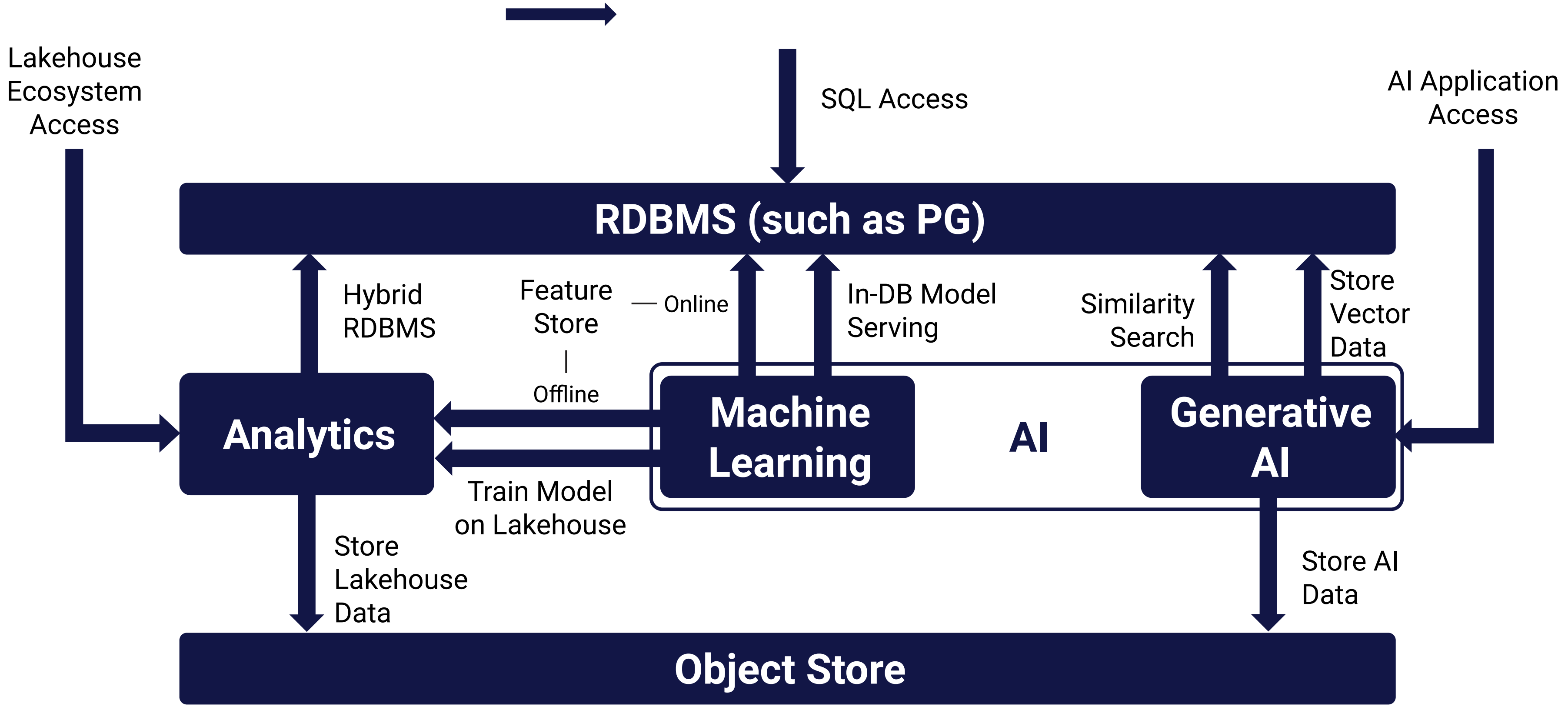


From Vector Database to AI Solution Platform



Analytics & AI Synergies

How does everything they rely on each other


Lakehouse Ecosystem Access
SQL Access
AI Application Access
RDBMS (such as PG)
Hybrid RDBMS
Feature Store
— Online
|
Offline
In-DB Model Serving
Similarity Search
Store Vector Data
Analytics
Machine Learning
AI
Generative AI
Train Model on Lakehouse
Store Lakehouse Data
Store AI Data
Object Store