mindsdb

NVIDIA

Making Enterprise Data Intelligent and Responsive for AI

The Potential of AI

AI Search

Example
Get the right answer in seconds from millions of sources

AI Analytics

Example
Real-time biz intelligence queries w/o data science resources

AI Agents

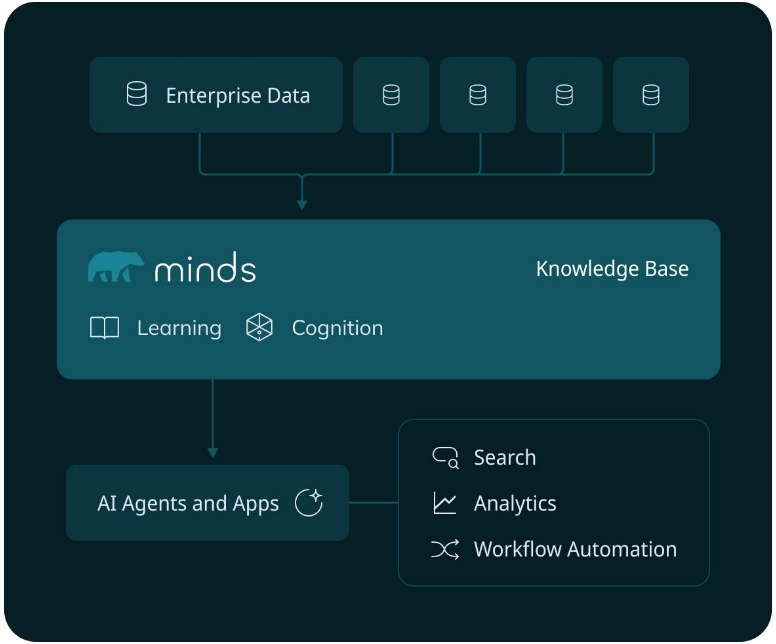
Example
Smarter chatbots leveraging enterprise data

Your AI Initiatives, Backed by Your Data

Powered by NVIDIA and MindsDB

Minds is the intelligent bridge

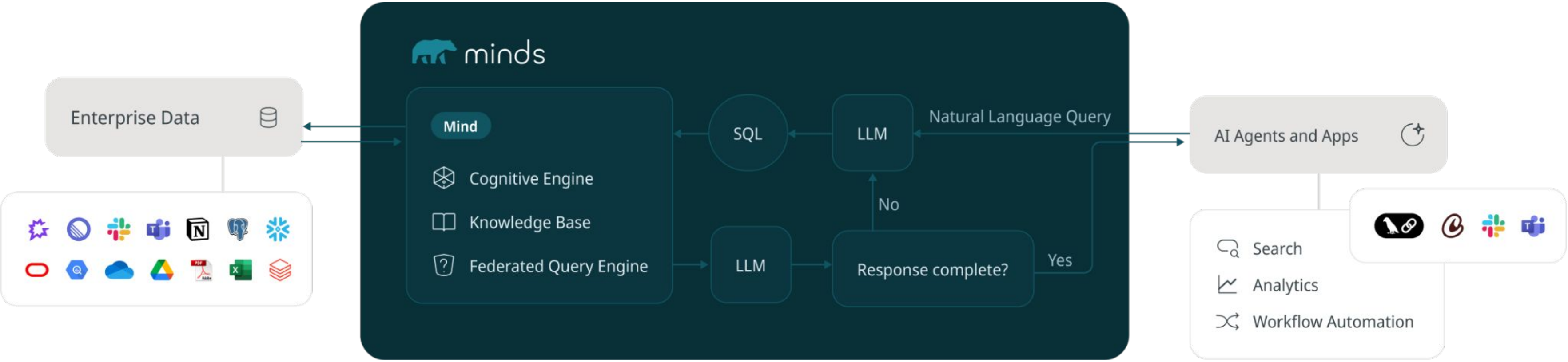
- Query any data, anywhere
- Secure and scalable to any infra
- No data engineering required



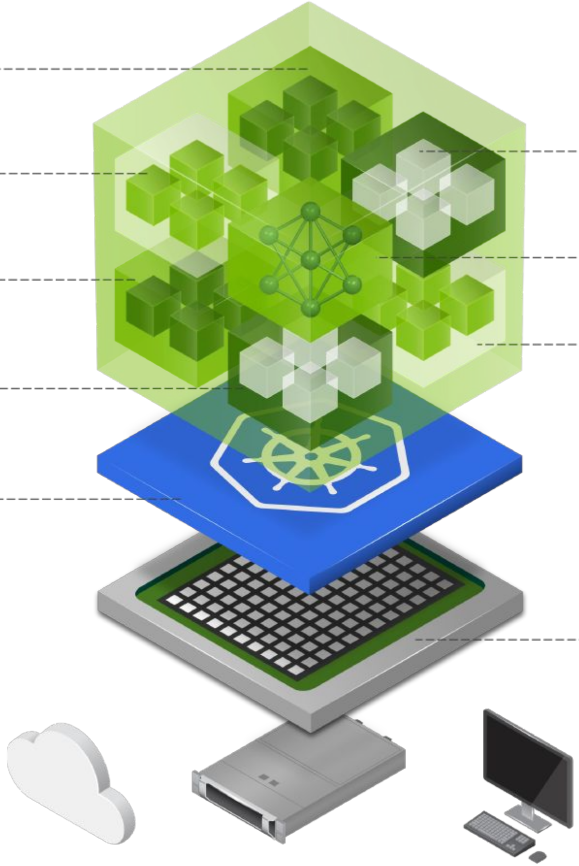
What is a Mind?

A "Mind" is the intelligent bridge between your enterprise data and AI solutions. Powered by the Cognitive Engine, Knowledge Base, and Federated Query Engine, it transforms raw data into actionable insights.

Build AI Search, Analytics, and Agents seamlessly—all from one solution.



NVIDIA NIM



Standard APIs
Text, Speech, Image,
Video, 3D, Biology

NVIDIA Triton Inference Server
cuDF, CV-CUDA, DALI, NCCL,
Postprocessing Decoder

Cloud-Native Stack
GPU Operator, Network Operator

Enterprise Management
Health Check, Identity, Metrics,
Monitoring, Secrets Management

Kubernetes

NVIDIA TensorRT and TensorRT-LLM
cuBLAS, cuDNN, In-Flight Batching,
Memory Optimization, FP8 Quantization

Optimized Model
Single GPU, Multi-GPU, Multi-Node

Customization Cache
P-Tuning, LoRA, Model Weights

CUDA

NVIDIA NIM Microservices

Fast, Enterprise AI Inference Anywhere

- **What It Is:** Prebuilt, optimized microservices to deploy AI models in minutes. Packages models, engines, standard APIs & dependencies into secure containers.

- **How to Start:**
    - **Prototype:** Free API access (NVIDIA Developer Program).
    - **Deploy:** Self-host (Free dev/test; NVIDIA AI Enterprise for production).

Performance & Scale
Up to 3x faster inference

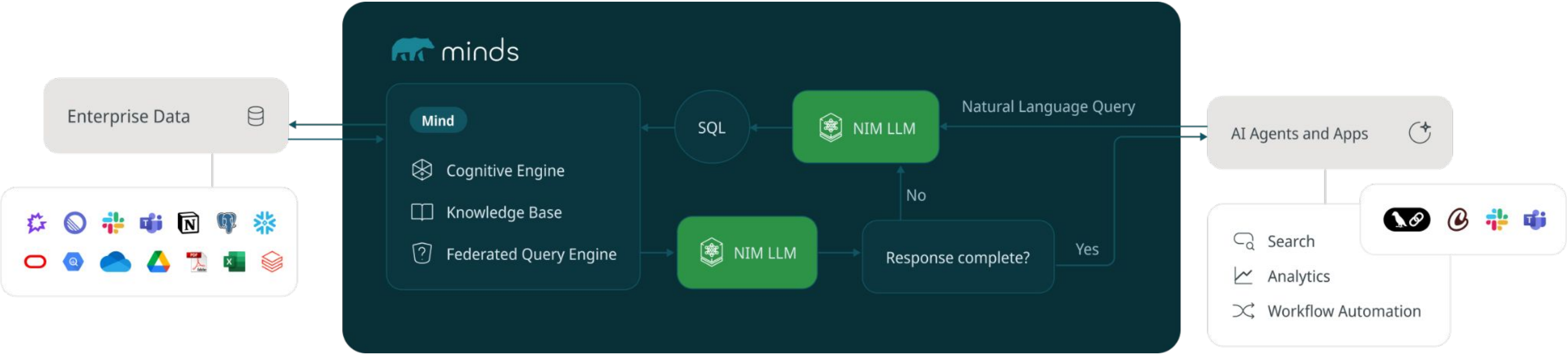
Ease of use
Deploy quickly with standard APIs

Flexibility
Run anywhere - Cloud, Data Center, Edge etc

Enterprise-Ready
Validated, managed, secure, and supported.

Minds – Powered by NVIDIA NIMs

Run a 'Mind' on bleeding edge infrastructure

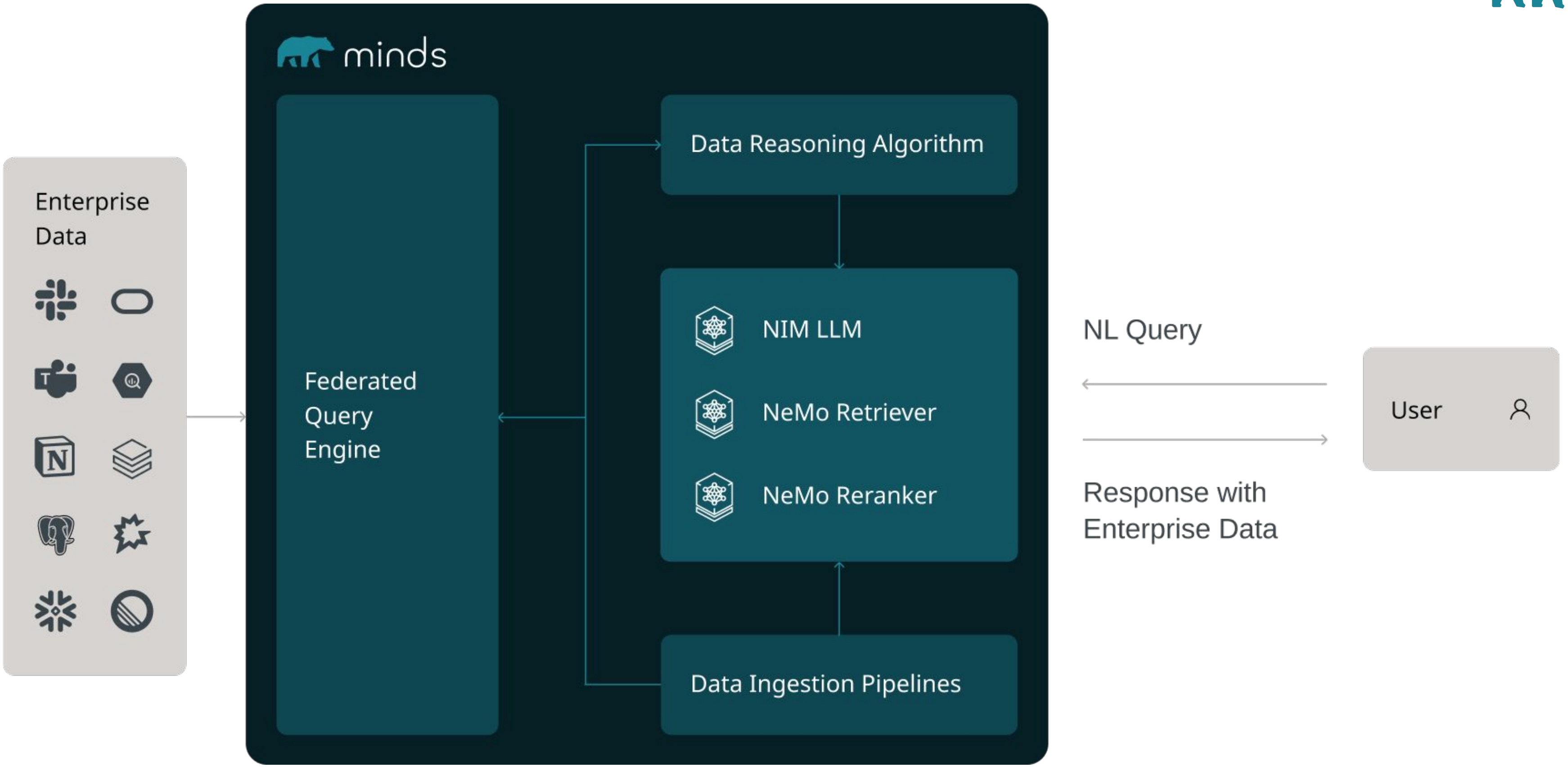


Minds, Powered by NVIDIA AI Enterprise

AI Data Automation backed by NVIDIA

Minds are an AI system to analyze and interact with custom data sets from structured and unstructured data. Minds are integrated with NVIDIA AI Enterprise using open and de facto standards.

- LLM inference via OpenAI compatible APIs (NVIDIA NIM)
- Retrieval using embeddings (Retriever NIM)
- Reranking (Reranker NIM)
- Vector databases (cuVS)
- Query plan and execution



Enterprise Data

minds

Federated Query Engine

Data Reasoning Algorithm

NIM LLM

NeMo Retriever

NeMo Reranker

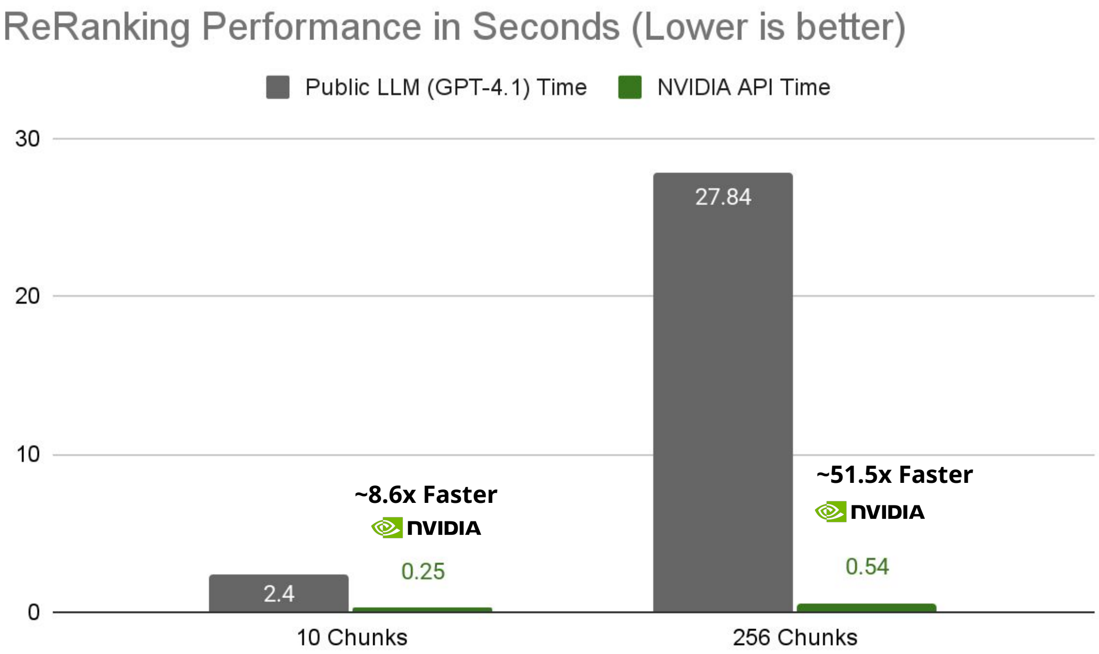
Data Ingestion Pipelines

NL Query

Response with Enterprise Data

User

The road to production, from public LLM to NVIDIA NIMs



Methods Compared:

- **Public LLM:** prompted for relevance classification.
- **NVIDIA:** Dedicated Reranking API (`llama-3.2-nv-rerankqa-1b-v2`)
.

Key Findings:

- NVIDIA's specialized API provides significantly faster performance & scalability, directly benefiting from its optimized design for handling lists efficiently in one call.
- The manual batching needed for the general Public LLM Chat API adds overhead and contributes to longer processing times for bulk tasks.

DEMO

NVIDIA and MindsDB

NVIDIA AI Enterprise GPU power for embedding and search at scale.

1. Setup **MindsDb**
2. **Define your use cas**e — What will you build?
3. Create an **Nvidia Developer Account**
4. Select **Nvidia NIM Models**
    - Instantiate models via a third-party platform or directly through the API
5. **Connect** Models to MindsDB
6. **Build!**

llama-3.3-nemotron-super-49b-v1

From Documents to Insights

MindsDB + NVIDIA

NVIDIA AI Enterprise GPU power for embedding and search at scale.

- **Leveraging NVIDIA's Enterprise RAG Pipeline:** Utilize GPU-accelerated NIM microservices for scalable embedding, search, and accurate information retrieval from enterprise data.
- **High-Performance Embeddings:** Use pre-trained models like **llama-3.2-nv-embedqa** or fine-tune for specific domains, achieving higher accuracy (~50% fewer incorrect answers).
- **Accelerated Multimodal Data Extraction**: Go beyond text – rapidly process complex PDFs (including tables, charts, graphics) with specialized NIMs (15x faster extraction).
- **Optimized Vector Search & Retrieval:** MindsDB RAG techniques combined with NVIDIA's accelerated hybrid search and reranking deliver relevant context.
- **Enriched Metadata:** Custom MindsDB parameters enhance retrieval for faster, smarter results.

5+ Million Documents, One Private AI

Case Study: MindsDB + NVIDIA

AI Search at Scale

- Customer: High-profile org in a regulated industry
- Goal: Search, reason, and answer from 5M+ private docs in a VPC, no external APIs
- Solution: Custom RAG pipeline on NVIDIA AI Enterprise, powered by NIM-served LLMs

$120K

Fines per day avoided

Solution: AI Search

Industry: Energy Sector
Unified 5.2M docs compiled across 70 years

Bring your use cases to life with MindsDB

AI Search

Connect your teams to the information they need, quickly.

Make enterprise search easier, and AI driven. Use natural language to query vast amounts of structured and unstructured data with precision.

AI Analytics

Get insightful answers you can trust, faster.

Turn data into insights instantly. Generate complex analysis of your data, using natural language and SQL across multiple data sources.

AI Agents

Enable AI agents to interact with your data like never before.

Your intelligent AI-powered agents now can easily retrieve, analyze, and act on data. Enhance automation and decision-making across your organization.

mindsdb | NVIDIA

Thank You

Alejandro Cantu, AI Product Manager, MindsDB

Vijaykumar Ranganathapura, Strategic Alliances, NVIDIA