Implementing Private Cloud RAG with Milvus & llmware

Agenda

Live Demo: end-to-end private cloud RAG system in action

Chalk Talk:

- Quick Intro to AI Bloks & our open source arm (**llmware**)
- Private Cloud RAG
- Conceptual Architecture
- Demo Scenarios –
  - Fast document parsing, text chunking and embedding
  - RAG with semantic embedding retrieval
  - Web App powered by Milvus + llmware

Why RAG in Private Cloud

- Data privacy of sensitive business documents
- Open source innovation curve + ability to customize
- Lower cost of smaller, specialized models
- Task type – extractive fact-based
- Integrate AI to the process, not the other way around

What is DRAGON ?

dRAGon – “Delivering RAG On …” model series – launched in mid-November 2023

- 7 open source RAG-finetuned models available in llmware repo on HuggingFace and the transformers library
- Built on leading open source 6-7B model bases – Llama, Mistral, Yi, Red-Pajama, StableLM, Falcon, Deci
- Easy to get started generation scripts
- First-tier integration in llmware (pip install llmware)
- Benchmark tested - “common-sense” RAG Instruct Benchmark

Small, Specialized LLMs for RAG

BLING

1-3B parameter CPU-inference "laptop" LLMs tuned for RAG instructions - designed for testing, prototyping and simple RAG extraction tasks in production

DRAGON

6-7B parameter production-grade, single GPU models optimized for RAG, and built on top of leading foundation models

INDUSTRY BERT

Industry-fine-tuned embedding models built on BERT Sentence Transformer base

High quality open source models, tuned for RAG tasks, built on industry-leading foundation models

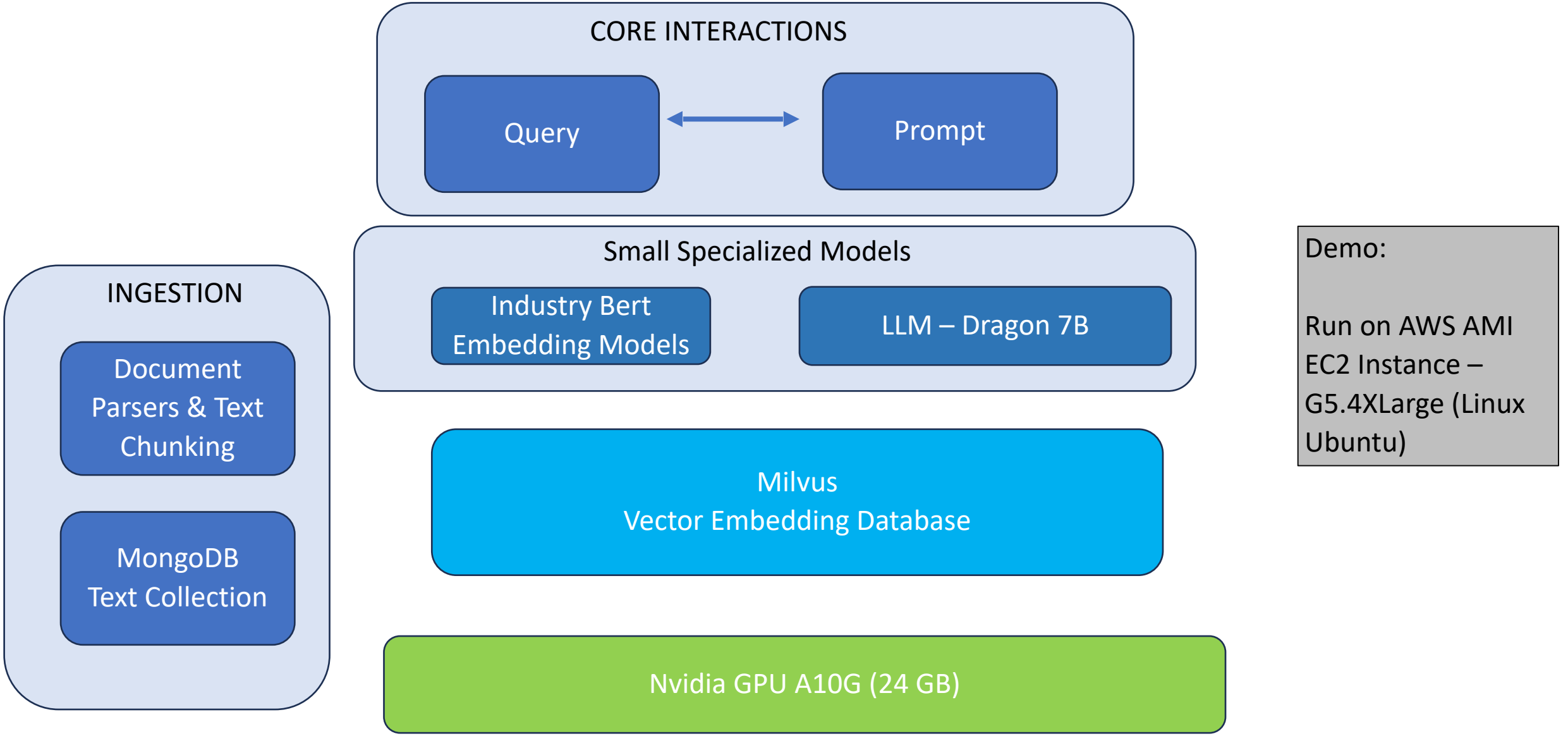
LLMWare

- New Python open source development framework for enterprise-grade LLM-based applications with focus on Private Cloud & Open Source Smaller Specialized Models
  - Designed for enterprise workflows
  - Scalable document ingestion
  - Integrated end-to-end data model with persistent data stores
  - Priority is Open Source Models & HuggingFace integration

- Getting started is as fast and easy as Langchain …

pip install llmware or repo at: Github.com/llmware-ai/llmware.git

Private Cloud RAG in a Box Architecture



Demo:

Run on AWS AMI EC2 Instance – G5.4XLarge (Linux Ubuntu)

DEMO