ODSC Europe 2024

How to run scalable, fault-tolerant RAG with a vector database

Prerequisites:

- A MacOS / Linux / Windows device with 8GB RAM or higher (preferably 16GB+)
- Python 3.8+
  - We will use `weaviate-client` (4.7.1+)
  - `pip install weaviate-client`
- Docker desktop (https://www.docker.com/products/docker-desktop/)
- Ollama (https://ollama.com/)
  - Alternative: An API key for an API-based embedding & LLM model provider
  - E.g. Cohere, OpenAI
- Minikube (https://minikube.sigs.k8s.io/docs/start/)
  - Optional, but preferable. The workshop could be done with Docker alone.
- Helm (https://helm.sh/docs/intro/install/)
  - Optional, if using Minikube.