

# Build your own RAG solution

Heli Helskyaho  
Pekka Kanerva  
Miracle Finland Oy

# Heli

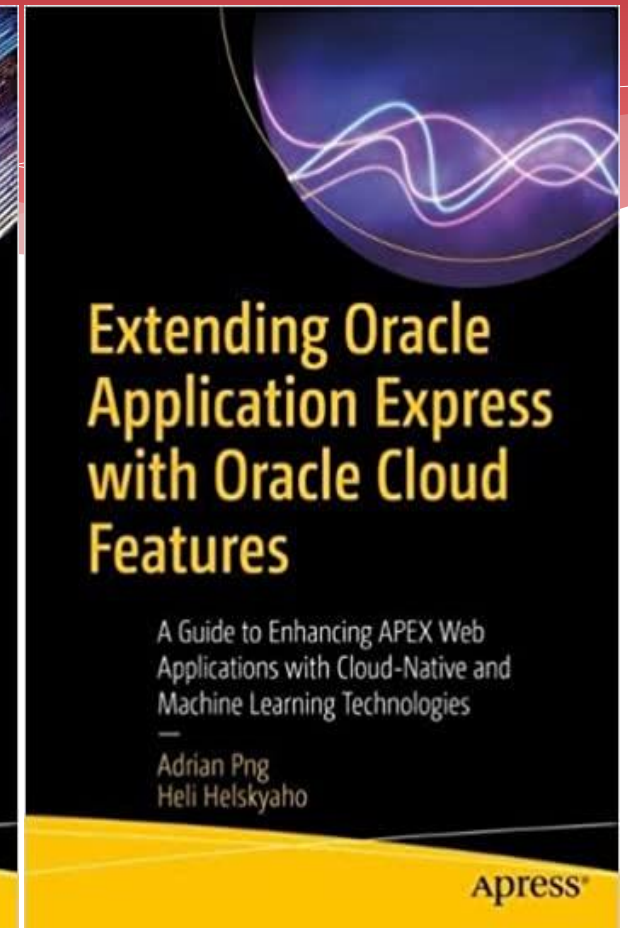
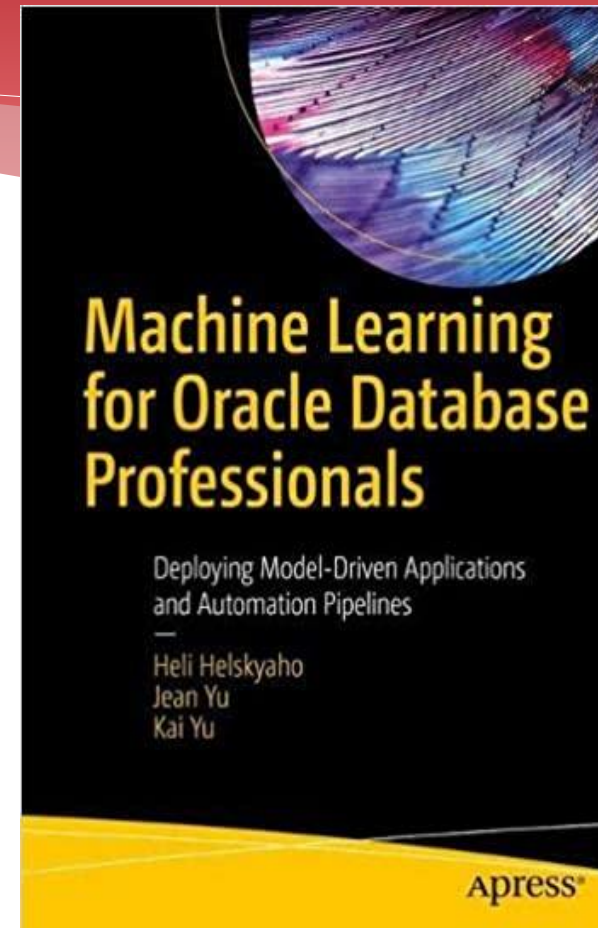
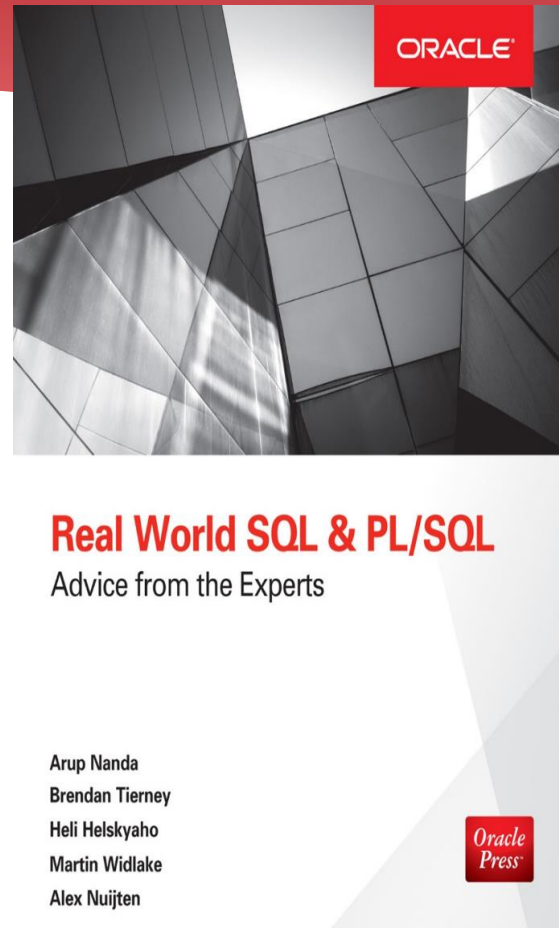
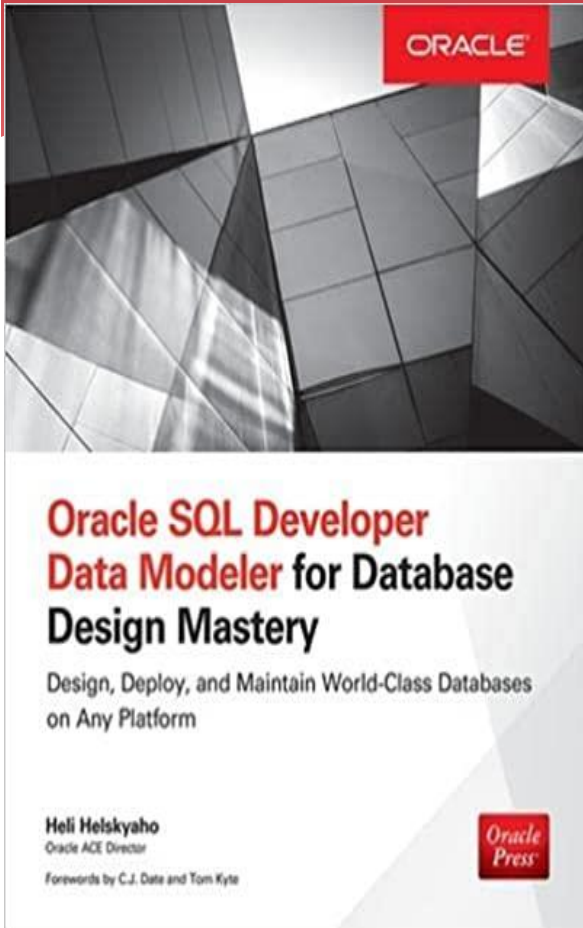


- \* Graduated from University of Helsinki (Master of Science, computer science), currently a doctoral student at University of Helsinki
- \* Worked with Oracle products since 1993, worked for IT since 1990
- \* Data and Database!
- \* CEO for Miracle Finland Oy
- \* Oracle ACE Director
- \* Public speaker and an author
- \* Author of the book Oracle SQL Developer Data Modeler for Database Design Mastery (Oracle Press, 2015), co-author for Real World SQL and PL/SQL: Advice from the Experts (Oracle Press, 2016), Machine Learning for Oracle Database Professionals: Deploying Model-Driven Applications and Automation Pipelines (Apress, 2021), and Extending Oracle Application Express with Oracle Cloud Features: A Guide to Enhancing APEX Web Applications with Cloud-Native and Machine Learning Technologies (Apress, 2022)



**Oracle ACE  
Director**

# Books



# Pekka

- \* Pekka Kanerva is Oracle technology consultant working for Miracle Finland Oy
- \* Pekka holds a Master's degree at the Helsinki University of Technology and he has been working with Oracle from the early 90'.
- \* He is specialized in Oracle technologies, application development with PL/SQL and APEX, and AI/ML Services in Oracle Cloud. Pekka has acquired several certificates on Oracle, Snowflake and Data Vault 2.0 technologies.

# Generative AI

- \* Can help us
  - \* Automate tasks, create efficiency, better outcomes and money savings
  - \* Create possibilities that did not exist before
  - \* ...



# Use cases (Tasks) are unlimited

- \* Content generation, augmentation
- \* Summarization
- \* Content personalization, customer segmentation
- \* Question answering, Conversations
- \* Virtual assistants
- \* Language style transferring, adjust the tone
- \* Creative writing, technical writing, articles, letters, emails,...
- \* Translations, localizations
- \* Feedback analysis, automated customer responses
- \* Code generation, error detection, debugging, code conversions to another language
- \* Code documentation, automated testing
- \* ...

# Concerns about GenerativeAI

- \* Legal issues
  - \* Privacy
  - \* Security
  - \* Intellectual property rights (IPR), protection
  - \* Acts, laws,...
- \* Ethical issues
  - \* Bias
  - \* Misinformation
- \* ...

# Hallucination and old data

- \* When the LLM does not know (does not have the data for the answer), it invents the answer: hallucinates
- \* When was the foundation model trained? Long time ago? Years ago? Old data!
- \* The data used for model training might be erroneous (internet).
- \* How to check the answer is correct? Where did the model find the answer?
- \* Note: remember the "job" of an LLM is to create content? So hallucination is actually exactly what we were expecting.



# What about my own data?

- \* Data in the database
- \* Data in our documentation
- \* Data in our knowledge base
- \* Data in our support system
- \* Data in our...

# The solution?

- \* Retrieval-Augmented Generation, RAG?

# RAG

- \* Adds the **context**
- \* **Augments** the LLM with your **own data**
- \* Adding data **without** training
- \* **Grounds** the responses to your data
- \* Reduces **hallucination**

# RAG

- \* Retrieval-Augmented Generation
- \* "Retrieval"
  - \* the ability to retrieve data from a data source; database, internet,...
- \* "Augmented Generation"
  - \* augmenting the generation process with retrieved data

# Generative AI without the context (ChatGPT)

what is OML?



OML can refer to a few different things depending on the context:

1. **Operational Machine Learning (OML):** This refers to the practice of deploying, monitoring, and managing machine learning models in a production environment. It involves the entire lifecycle of machine learning, from model training and validation to deployment and ongoing maintenance. OML is important for ensuring that models perform well over time and adapt to new data.
2. **One More Level (OML):** In gaming, "OML" is often used as an abbreviation for "One More Level," which refers to the addictive nature of some games that encourage players to keep playing "just one more level."
3. **Oracle Machine Learning (OML):** This is a platform provided by Oracle that allows data scientists and developers to build, train, and deploy machine learning models within Oracle databases. It integrates machine learning capabilities directly into Oracle's database systems.
4. **Outright Monetary Transactions (OMT):** In finance, "OML" might be confused with OMT, which refers to a program by the European Central Bank aimed at stabilizing bond markets in the Eurozone.
5. **Online Math Learning (OML):** A platform or website that provides resources and tutorials for students to learn mathematics online.
6. **Oh My Lord (OML):** A casual expression used in online conversations or social media to express surprise, excitement, or disbelief, similar to "Oh My God (OMG)."

The meaning of OML depends on the context in which it's used.

Copyright © Miracle Finland Oy

# RAG solution

Question



Question  
What is OML?

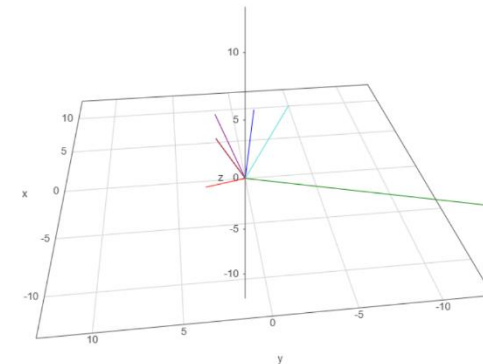
Response  
OML stands for Oracle Machine Learning.

# Augmentation

- \* The response can be augmented with data in the **prompt** (prompt engineering)
- \* Or in a **database, internet**, or any data source we can reach
- \* But if we want to augment it with, for example, by data from pdf documents or images, we need **vectors**

# Vectors

- \* Unstructured data (text, image, voice, video,...) is transformed (encoded, embedded) into numeric representation and stored as vectors
- \* The data is stored as its *semantic content*, not the actual content, obtained by vectorizing





# Distance and Similarity metrics

- \* Examples of these metrics
  - \* Euclidean (*EUCLIDEAN*) and Euclidean Squared Distances (*EUCLIDEAN\_SQUARED* or *L2\_SQUARED*)
  - \* Cosine Similarity (*COSINE*)
  - \* Dot Product Similarity (*DOT*)
  - \* Manhattan Distance (*MANHATTAN*)
  - \* Hamming Similarity (*HAMMING*)
- \* Used as a *vector distance operand* to the **VECTOR\_DISTANCE** Function in the Oracle Database 23ai

# Distance and Similarity metrics

- \* The metric to be chosen depends on the embedding model chosen!
- \* Use the distance/similarity metric that was used to train your embedding model.
- \* Documentation!

embed-multilingual-v2.0	multilingual classification and embedding support. <a href="#">See supported languages here.</a>	768	256	Dot Product Similarity	<a href="#">Classify, Embed</a>
embed-english-v3.0	A model that allows for text to be classified or turned into embeddings. English only.	1024	512	Cosine Similarity	<a href="#">Embed, Embed Jobs</a>
embed-english-light-v3.0	A smaller, faster version of <code>embed-english-v3.0</code> . Almost as capable, but a lot faster. English only.	384	512	Cosine Similarity	<a href="#">Embed, Embed Jobs</a>
embed-multilingual-v3.0	Provides multilingual classification and embedding support. <a href="#">See supported languages here.</a>	1024	512	Cosine Similarity	<a href="#">Embed, Embed Jobs</a>

# Oracle Database Vector datatype

Define dimension and format.

Dimension: how many dimensions in a vector.  
[1.1, 2.2, 3.3] has three dimensions.

Operations for using the new datatype.

```
CREATE TABLE t2 (  
  v1 VECTOR,  
  v2 VECTOR(384, *),  
  v3 VECTOR(768, FLOAT32),  
  v4 VECTOR(1024, FLOAT64),  
  v5 VECTOR(4096, INT8),  
  v6 VECTOR(*, *)  
);
```

```
DESC t2;
```

Name	Null?	Type
V1		VECTOR(* , FLOAT32)
V2		VECTOR(384 , *)
V3		VECTOR(768 , FLOAT32)
V4		VECTOR(1024 , FLOAT64)
V5		VECTOR(4096 , INT8)
V6		VECTOR(* , *)

# A table with data and a vector

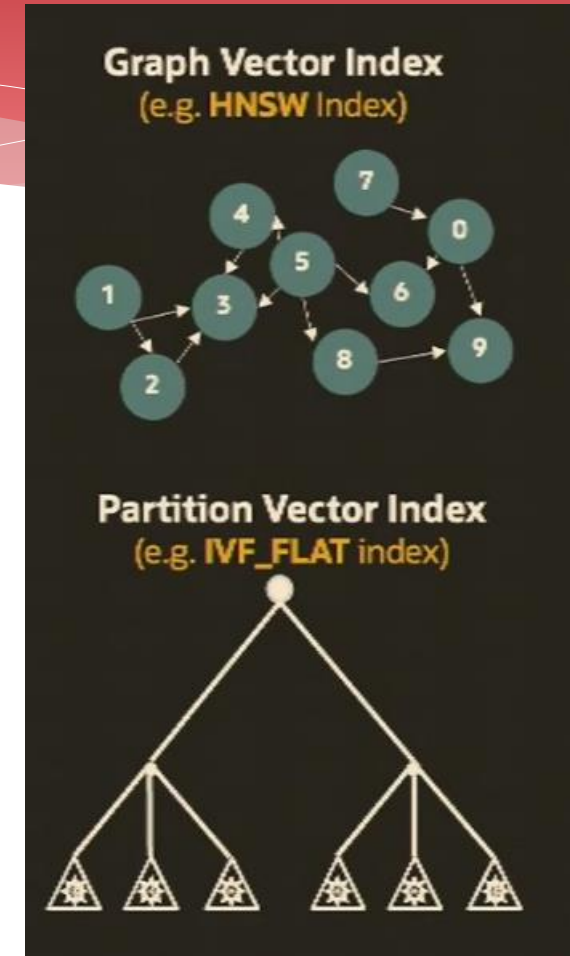
```
Create table MyText (  
  TextID Number(16),  
  TextClause (CLOB),  
  Text_vector VECTOR);
```

# The RAG Process (simplified)

- \* **Generate vectors** for unstructured data using an embedding model
- \* **Save** vectors into the database in a column of VECTOR datatype
- \* Create Approximate Vector **Index** for the VECTOR column
- \* **Query** using AI Vector Search (SQL)

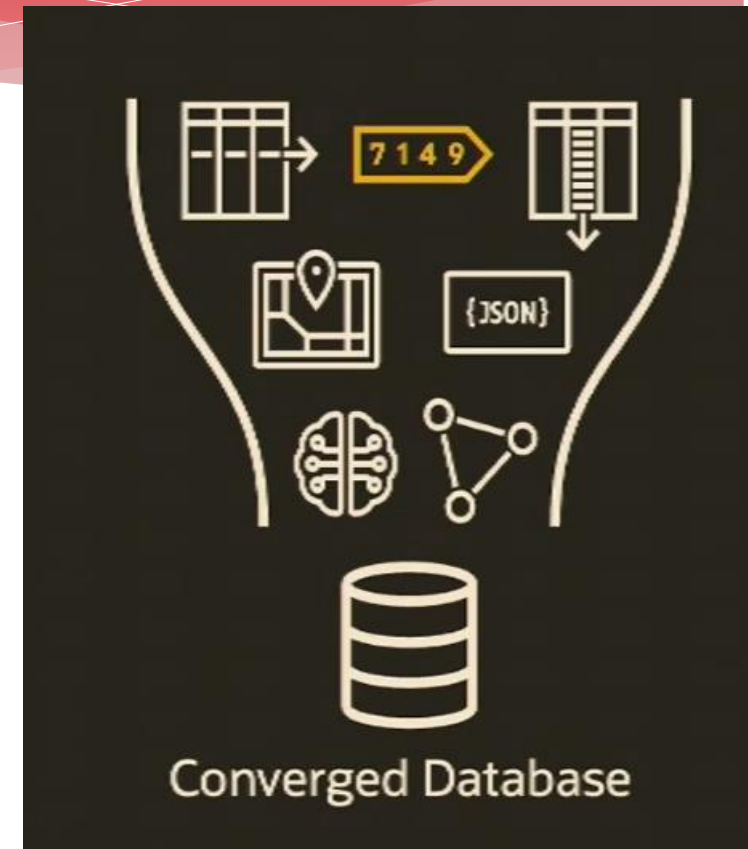
# Approximate Vector Indexes

**CREATE VECTOR INDEX** text\_idx ON Customer(text\_vector)  
**ORGANIZATION** [INMEMORY NEIGHBOR GRAPH | NEIGHBOR PARTITIONS]  
**DISTANCE** EUCLIDEAN | COSINE\_SIMILARITY | HAMMING ...

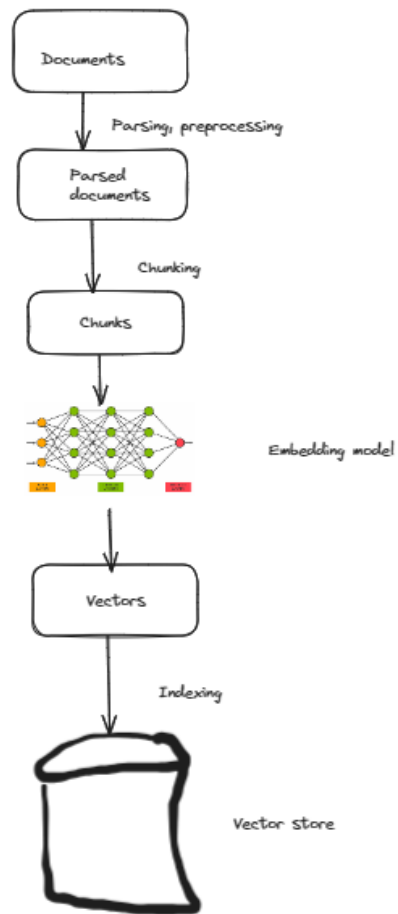


# SQL joins (Authors, Books, Pages)

```
Select pageID  
from Authors, Books, Pages  
where Authors.authorID = Books.authorID and  
Books.bookID = Pages.bookID and  
Books.bookGenre = 'Fiction' and  
Author.authorCountry = 'Finland'  
order by vector_distance(pageVec, :queryVec)  
fetch approx first 5 rows only;
```

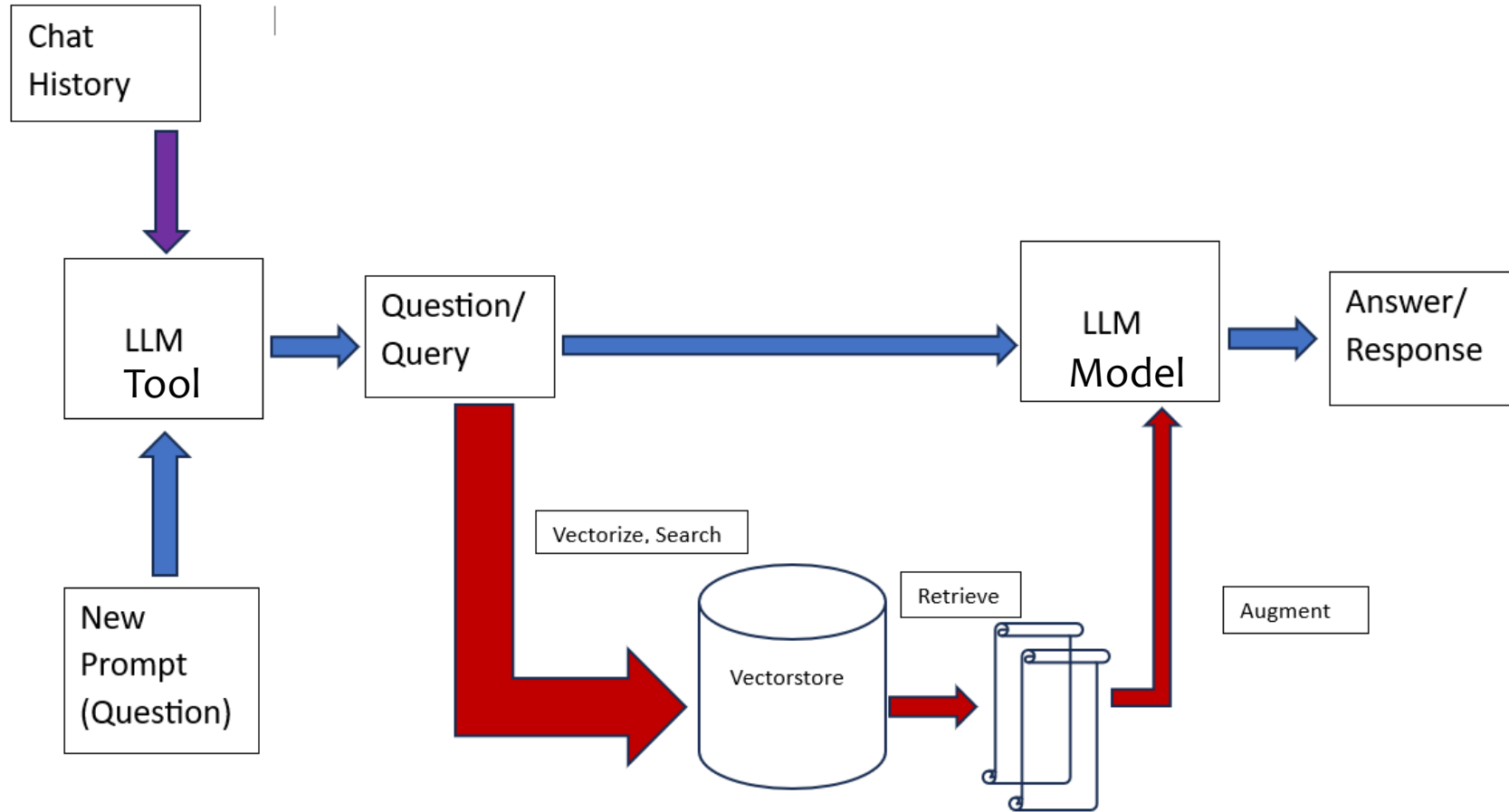


# Storing Documents

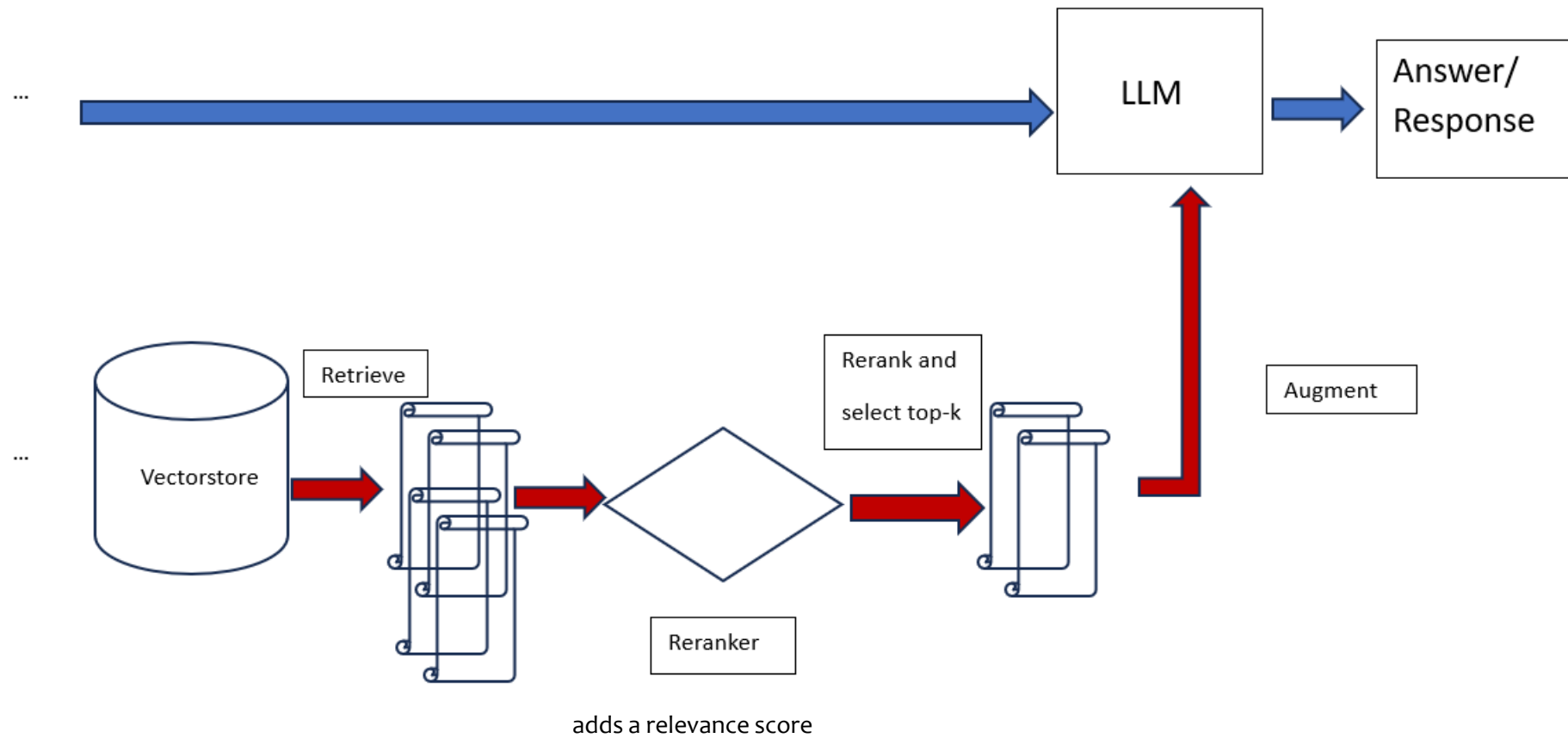




# RAG



# Reranking



# Oracle?

# Data Science Service

The image displays four screenshots of the Oracle Cloud Data Science Service interface, arranged in a 2x2 grid. The top-left screenshot shows the 'Launcher' page with a welcome message and instructions. The top-right screenshot shows a terminal window with the command `odsc conda install -i pygpx2410_p38_cpu_v1` and its output, including warnings and download progress. The bottom-left screenshot shows the 'Environment Explorer' page with a table of available conda environments. The bottom-right screenshot shows a notebook editor with a code cell containing Python code for data analysis.

**Launcher**

Welcome to the Data Science service

The Launcher provides easy access to your notebooks, console, text editor, terminal, Environment Explorer, Notebook Explorer.

To get started, use the Environment Explorer to install a conda environment.

To be able to publish your own conda environments, specify the location to store published conda environments and how to authenticate with object storage.

**Environment Explorer**

Conda Environments: Data Science (26 of 26) Published (0 of 0) Installed (0 of 0) Clear

Architecture: ALL CPU GPU Show Deprecated (0 of 33)

Name	Environment Version	Type	Language	Architecture	Created	Size
PyTorch 2.1 for GPU on Python 3.9	1.0	Data Science	Python 3.9	GPU	2 weeks ago	7.38 GB
PySpark 3.2 and Feature Store	3.0	Data Science	Python 3.8	CPU	2 weeks ago	2.77 GB
Oracle Property Graph 24.1 for CPU on Python 3.8	1.0	Data Science	Python 3.8	CPU	2 weeks ago	2.91 GB
ARM Pack for Machine Learning	1.0	Data Science	Python 3.8	CPU	2 months ago	674.56 MB

**Terminal**

```
(base) bash-4.2$ odsc conda install -i pygpx2410_p38_cpu_v1
WARNING:ODSC:The current platform matches the default platform (x86_64). However, if the pack being installed was not built on x86_64 it will not work.
Environment slug: pygpx2410_p38_cpu_v1
INFO:ODSC:Downloading conda pack pygpx2410_p38_cpu_v1...
INFO:ODSC:Writing to /home/datascience/pygpx2410_p38_cpu_v1.tar.gz
Downloading pack pygpx2410_p38_cpu_v1: 100%
INFO:ODSC:download complete
INFO:ODSC:Extracting conda environment "/home/datascience/pygpx2410_p38_cpu_v1.tar.gz"
INFO:ODSC:Running conda-unpack script.
INFO:ODSC:Downloading Notebooks for the pack: Oracle Property Graph 24.1 for CPU on Python 3.8
INFO:ODSC:Checking for notebooks with prefix notebooks/pygpx2410_p38_cpu_v1/
Saving Notebooks: 0it [00:00, 717/s]
INFO:ODSC:Conda environment has been successfully installed.
Removing: /home/datascience/pygpx2410_p38_cpu_v1.tar.gz
The environment setup is complete. To activate, run "conda activate /home/datascience/conda/pygpx2410_p38_cpu_v1" in your terminal. It may take a few seconds for the kernel to appear in the JupyterLab Launcher. To change the description of the environment, update /home/datascience/conda/pygpx2410_p38_cpu_v1/*_manifest.yaml.
(base) bash-4.2$
```

**Notebook Explorer**

You can access the diabetes dataset license here.

The notebook is compatible with the following Data Science conda environment:

- Parallel Graph AnalytIX 23.1 and Oracle Property Graph 23.1 for CPU on Python 3.8 (version 1.0)

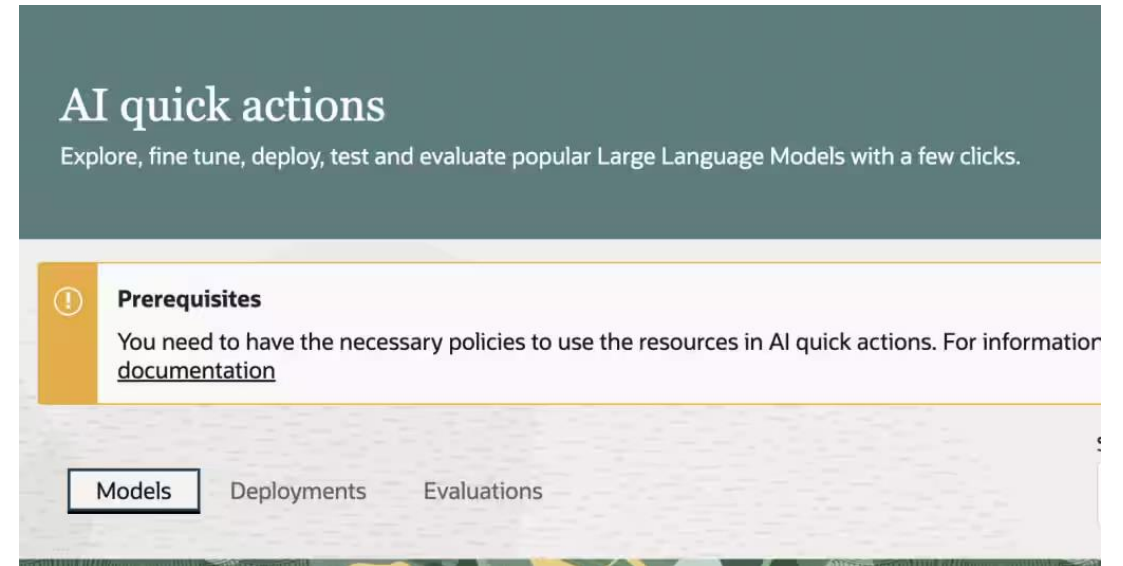
```
[1]: import logging
      logging.basicConfig(level=logging.ERROR)
      import warnings
      warnings.filterwarnings('ignore')

      import json
      import numpy as np
      import os
      import pandas as pd
      import pygpx
      import seaborn as sns
      import sklearn
      import shutil
      import tempfile

      from ads.common.model import ADSModel
      from ads.dataset.dataset_browser import DatasetBrowser
      from ads.dataset.factory import DatasetFactory
      from ads.evaluations.evaluator import ADSEvaluator
      from sklearn import datasets
      from sklearn.ensemble import RandomForestClassifier
      from sklearn.preprocessing import StandardScaler
```

# Data Science Service

- \* AI Quick Actions
  - \* *Deploying* a foundation model
  - \* *Fine-tuning* a foundation model
  - \* *Test* a foundation model
  - \* *Evaluating* a model
  - \* ...



# AI Service: Generative AI

The screenshot shows the Oracle Cloud Generative AI Playground interface. At the top, the Oracle Cloud logo and a search bar are visible. The location is set to "US Midwest (Chicago)". The left sidebar contains navigation links: Overview, Playground, Generation (selected), Summarization, Embedding, Dedicated AI clusters, Custom models, and Endpoints. Below these is a "Scope" section with a "Compartment" dropdown set to "AIML" and a sub-compartment "heliace (root)/AIML".

The main area is titled "Generation" and includes instructions: "To get started, choose a model and a preset prompt example. Then, refine the prompts and parameters to fit your use cases. See [model types](#) for more information." It features a "Model" dropdown menu with options: cohere.command v15.6, cohere.command v15.6 (highlighted), cohere.command v14.2, cohere.command-light v15.6, cohere.command-light v14.2, and meta.llama-2-70b-chat. There are also "View model details" and "Choose example" buttons. A "View code" button is present. Below the model selection is a large text area for prompts, with a "Generate" button and "Copy input" and "Clear" buttons. A status bar at the bottom of the text area shows "Character count - 0 | Token limit - 4000".

On the right, the "Parameters" section includes sliders and input fields for: Maximum output tokens (set to 600), Temperature (set to 1), Top p (set to 0,75), Top k (set to 0), and Stop sequences (with a text input field "Enter sequence and press enter").

At the bottom of the interface, there are links for "Terms of Use and Privacy" and "Cookie Preferences", and a copyright notice: "Copyright © 2024, Oracle and/or its affiliates. All rights reserved."

# Oracle Database 23ai

- \* ONNX models and Oracle Machine Learning
- \* Vector datatype
- \* Approximate Vector Indexes
- \* PL/SQL Packages
- \* AI Vector Search
- \* Data dictionary views for DBMS\_VECTOR
- \* Several data models (relational, Graph, Spatial, JSON, ...)
- \* SQL query language to query all this data
- \* ...


# A RAG with APEX



# RAG Demo: GenAI + AI Vector Search

## Search from your own PDFs

1. Chunk and embed PDFs (done inside the database) to vectors and store to DB
2. Ask a question
  - a) Get closest matches using AI Vector Search
  - b) Send question and matches to OCI GenAI and get answer

 **GenAI + AI Vector Search**

**Ask GenAI**

Answer from GenAI

Oracle Cloud Infrastructure Sovereign Cloud is a cloud deployment model that is specifically designed to help customers implement cloud data sovereignty strategies and comply with unique regulatory requirements. This offering from Oracle provides extra assurance and control for customers who require that their data remains in a designated region and is isolated from the commercial public cloud.

The Oracle Sovereign Cloud Principles document outlines several considerations for implementing cloud sovereignty strategies. It is intended to assist customers in understanding the potential business benefits of planning for the implementation of sovereign cloud deployment strategies and associated product features.

Specifically, Oracle's Sovereign Cloud offers a realm that is isolated from their commercial public cloud. This isolation enables Oracle to limit support and operations personnel to residents of a specific region, offering further control and adherence to regional requirements.

Would you like to know more about Oracle's Sovereign Cloud offerings?

# Thank you!

QUESTIONS?

[Heli.helskyaho@miracleoy.fi](mailto:Heli.helskyaho@miracleoy.fi)

[Pekka.kanerva@miracleoy.fi](mailto:Pekka.kanerva@miracleoy.fi)

@HeliFromFinland

Blog: [Helifromfinland.com](http://Helifromfinland.com)