Safe harbor statement

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, timing, and pricing of any features or functionality described for Oracle's products may change and remains at the sole discretion of Oracle Corporation.

Copyright © 2025, Oracle and/or its affiliates

ORACLE
Data Deep Dive
at AI World

RAG in a Box

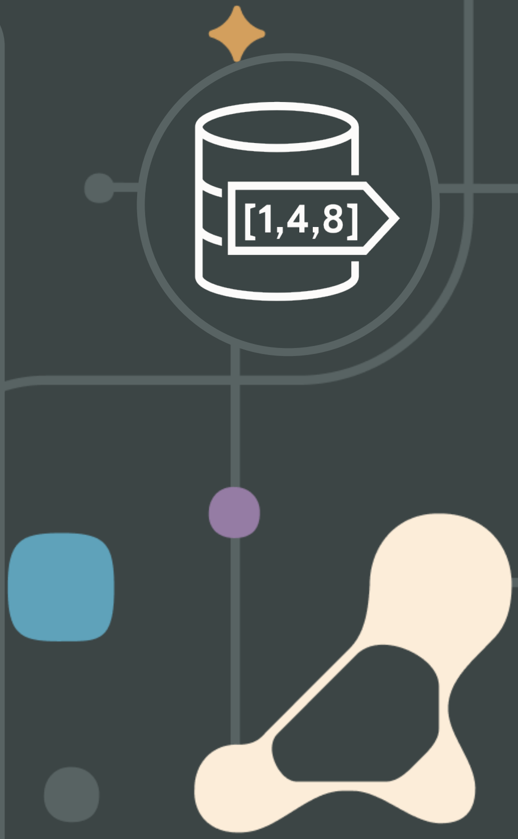
Simplify and Accelerate Data Searches

Massimo Castelli
Vice President AI & Data Platform Strategy

Doug Hood
Product Manager for AI Vector Search and Private AI Services Container

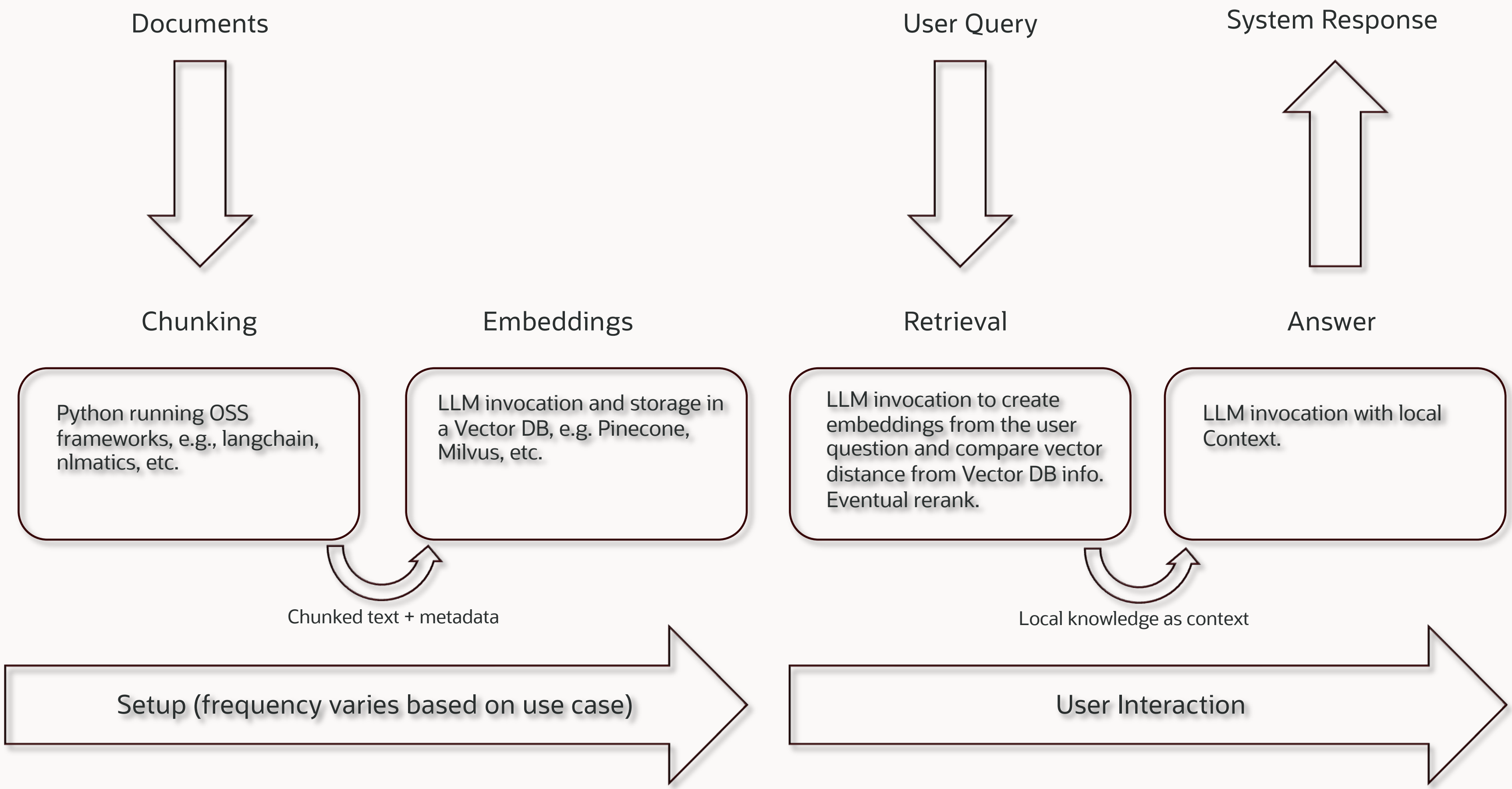
Retrieval Augmented Generation (RAG)

Enables LLMs to use business data to produce better and more contextually relevant answers to user questions while keeping business data secure



Copyright © 2025, Oracle and/or its affiliates

Typical RAG Pipeline



Documents

User Query

System Response

Chunking

Python running OSS frameworks, e.g., langchain, nlmatics, etc.

Embeddings

LLM invocation and storage in a Vector DB, e.g. Pinecone, Milvus, etc.

Chunked text + metadata

Retrieval

LLM invocation to create embeddings from the user question and compare vector distance from Vector DB info. Eventual rerank.

Answer

LLM invocation with local Context.

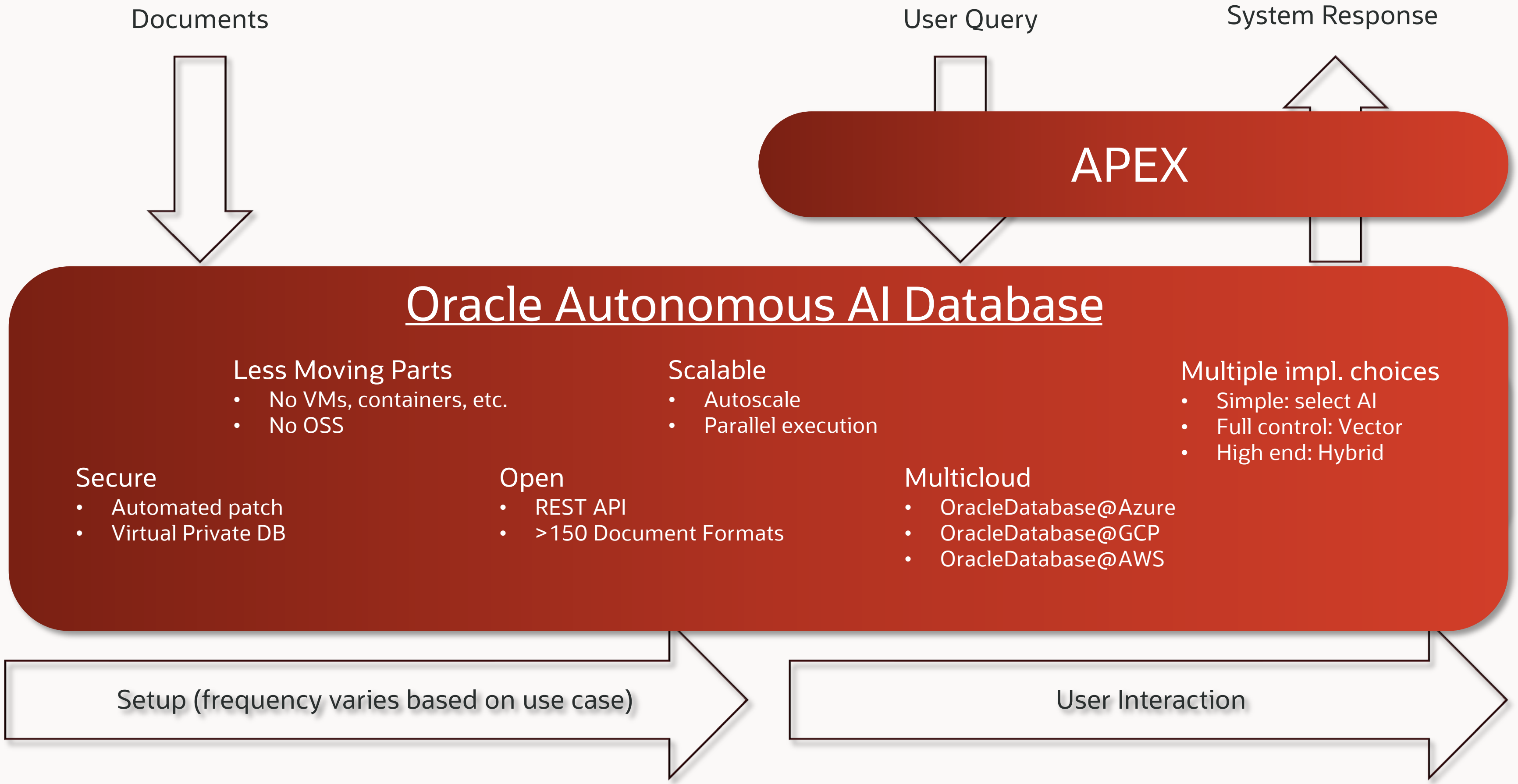
Local knowledge as context

Setup (frequency varies based on use case)

User Interaction

Oracle Copyright © 2025, Oracle and/or its affiliates

Typical RAG Pipeline



Documents

User Query

System Response

APEX

Oracle Autonomous AI Database

Less Moving Parts
- No VMs, containers, etc.
- No OSS

Scalable
- Autoscale
- Parallel execution

Multiple impl. choices
- Simple: select AI
- Full control: Vector
- High end: Hybrid

Secure
- Automated patch
- Virtual Private DB

Open
- REST API
- >150 Document Formats

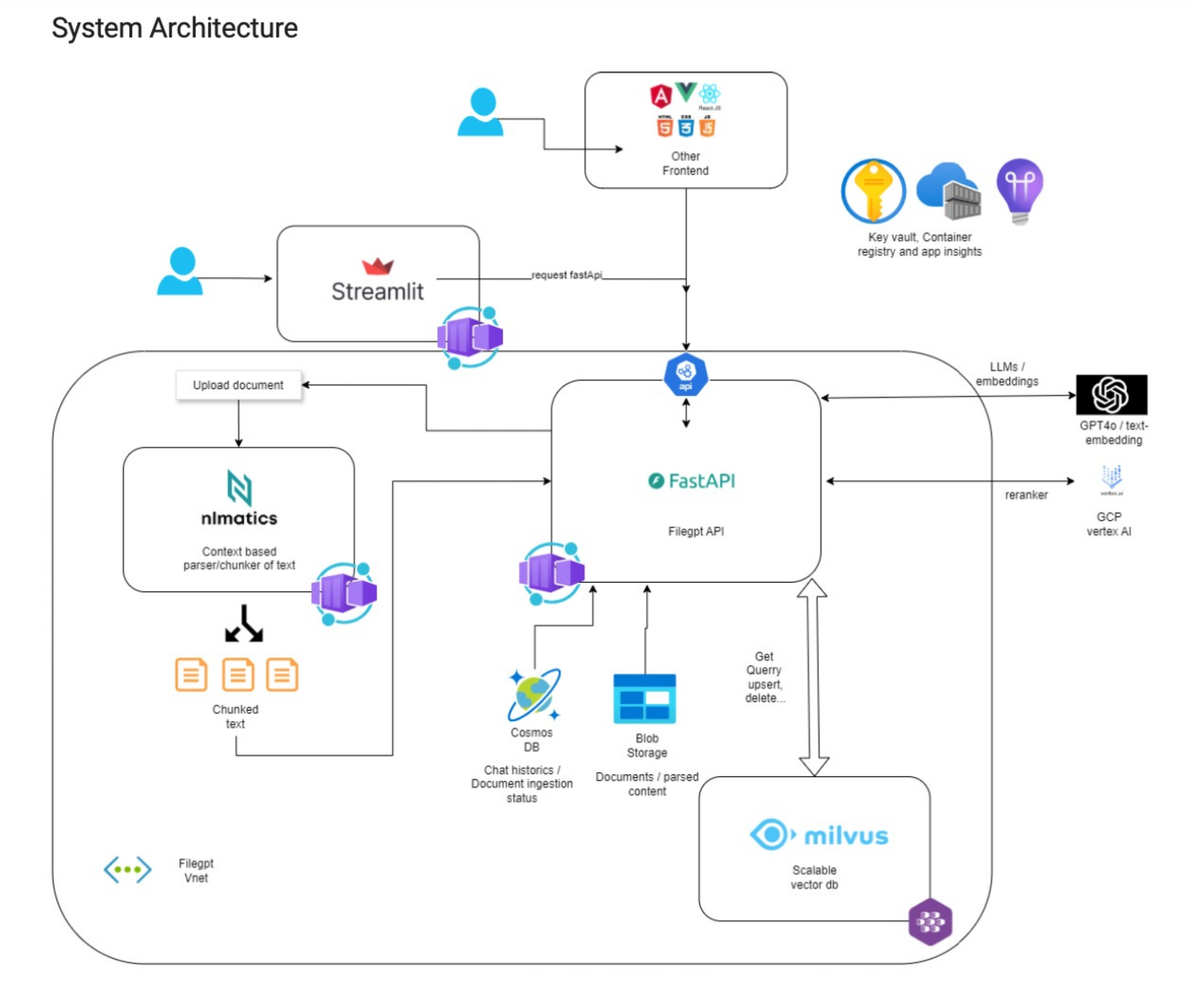
Multicloud
- OracleDatabase@Azure
- OracleDatabase@GCP
- OracleDatabase@AWS

Setup (frequency varies based on use case)

User Interaction

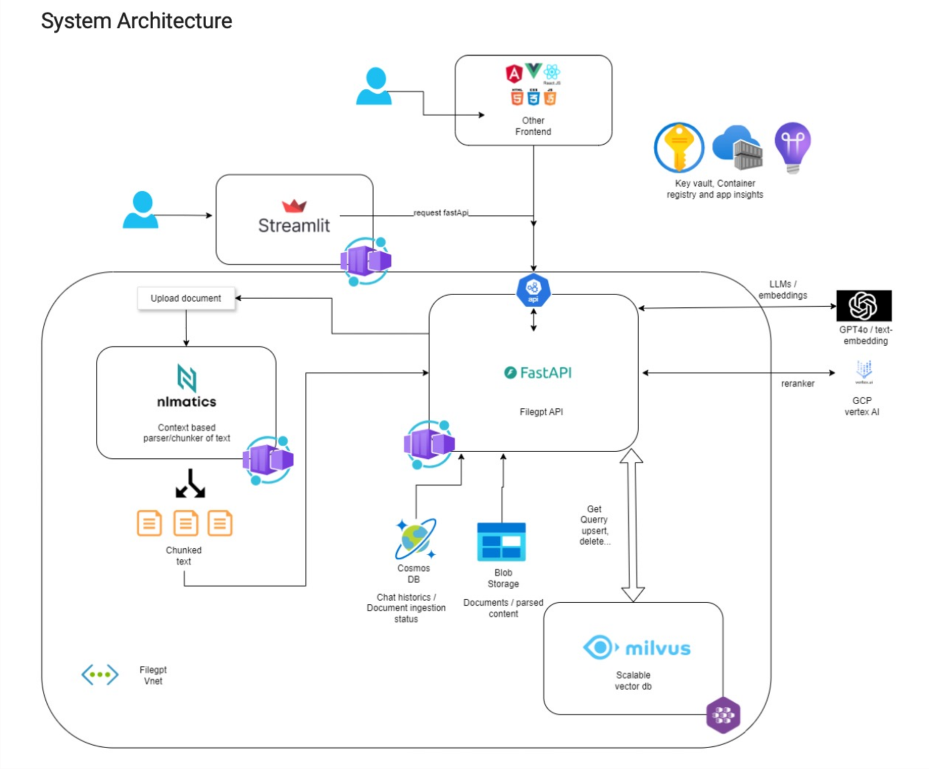
Oracle Copyright © 2025, Oracle and/or its affiliates

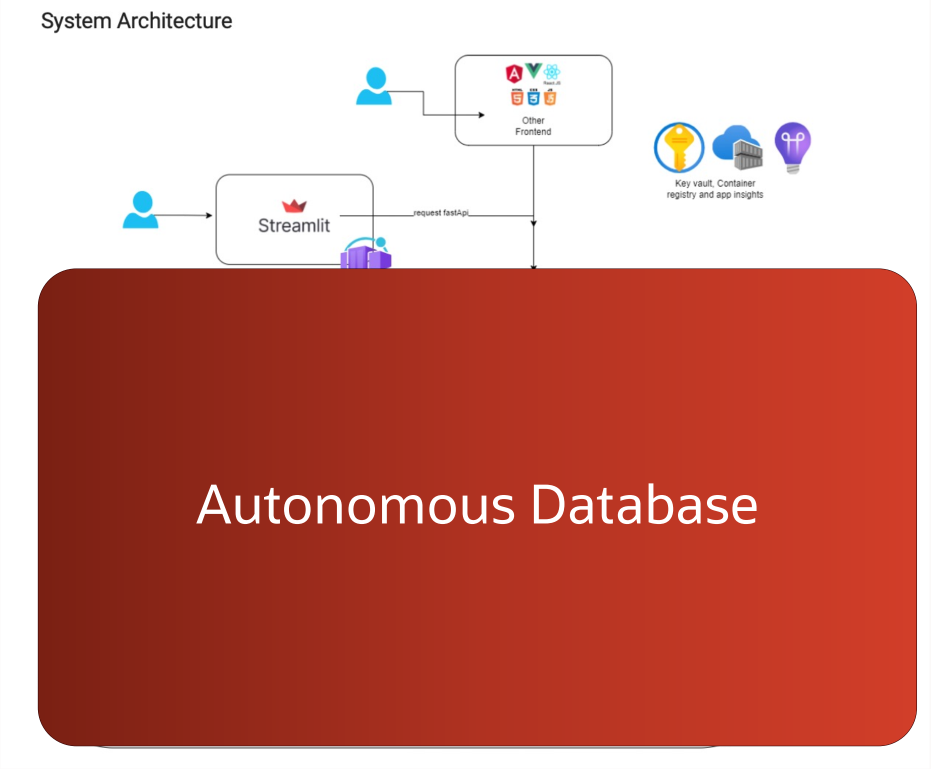
A real architecture as example…



Copyright © 2025, Oracle and/or its affiliates

A real architecture as example…





Copyright © 2025, Oracle and/or its affiliates

The “Three Ways of RAG” using Autonomous AI Database

Write your own server code
this is the recommended way when you want total control

No code / use embedded capabilities
this is the recommended way when you are focused on getting a result fast

Agentic approach
this is the recommended way when RAG is to be one part in a larger set of operations

Copyright © 2025, Oracle and/or its affiliates

The “Three Ways of RAG” using Autonomous AI Database

Write your own server code – project assumptions

- Domain Knowledge stored as documents in an object store bucket
- Don’t want/need to load the documents into the DB first (duplication/cost)
- No front-end dependencies (once I get the answer how it gets visualized is independent from it)

Oracle Copyright © 2025, Oracle and/or its affiliates

The “Three Ways of RAG” using Autonomous AI Database

Write your own server code – technology choices

- Select AI
    - Essentially just two calls
    - Limited control on process settings
    - SQL interface

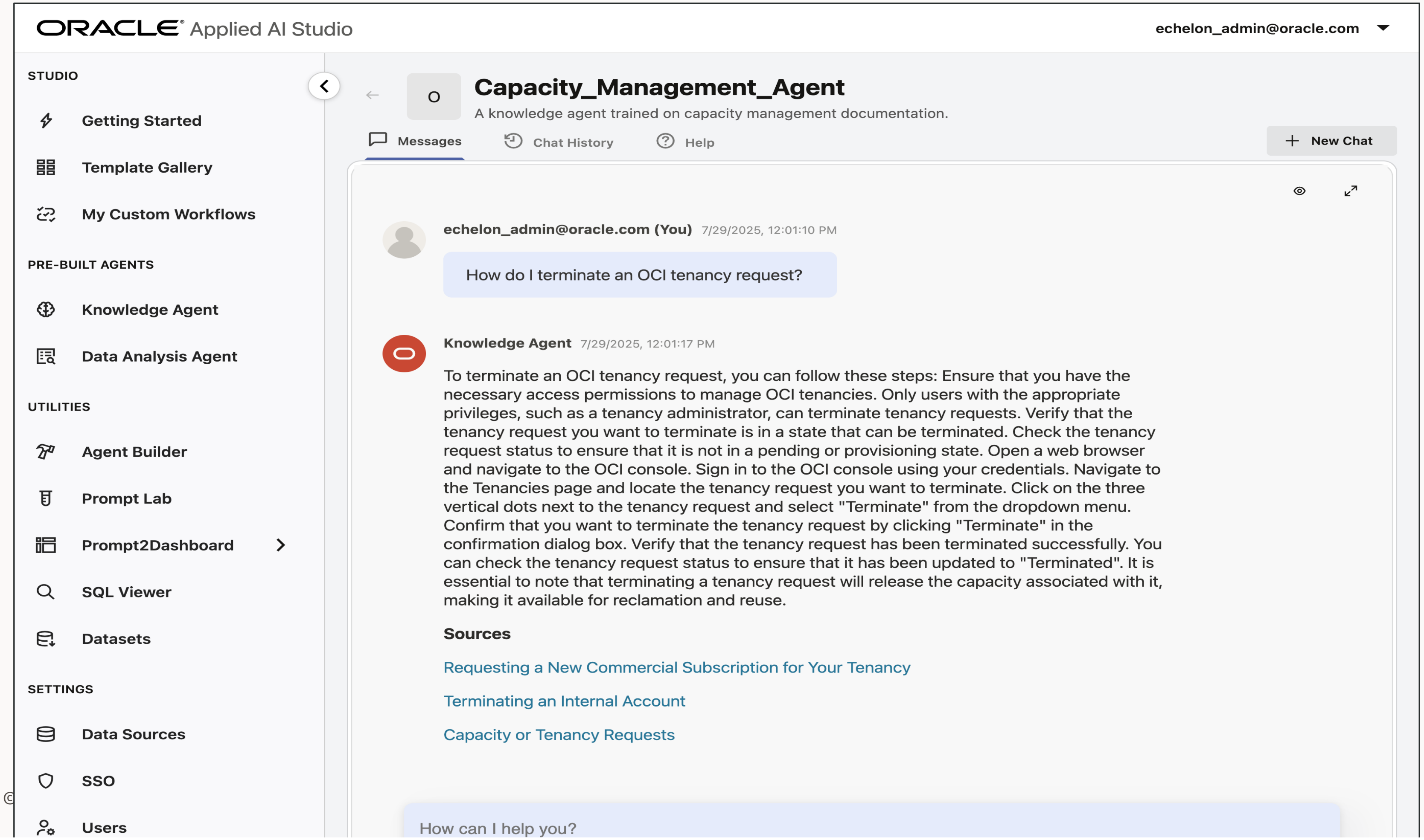
- Vector Index
    - Definitely more than two calls
    - Full control on chunking, embedding, results filtering, and answer creation

- Hybrid Vector Index
    - Complex but very powerful query syntax
    - Less control on the settings
    - Huge optimization (avoid expensive reranking!) through union of Text and Vector searches

Oracle Copyright © 2025, Oracle and/or its affiliates

The “Three Ways of RAG” using Autonomous AI Database

No code approach – just load your documents



echelon_admin@oracle.com (You) 7/29/2025, 12:01:10 PM

How do I terminate an OCI tenancy request?

Knowledge Agent 7/29/2025, 12:01:17 PM

To terminate an OCI tenancy request, you can follow these steps: Ensure that you have the necessary access permissions to manage OCI tenancies. Only users with the appropriate privileges, such as a tenancy administrator, can terminate tenancy requests. Verify that the tenancy request you want to terminate is in a state that can be terminated. Check the tenancy request status to ensure that it is not in a pending or provisioning state. Open a web browser and navigate to the OCI console. Sign in to the OCI console using your credentials. Navigate to the Tenancies page and locate the tenancy request you want to terminate. Click on the three vertical dots next to the tenancy request and select "Terminate" from the dropdown menu. Confirm that you want to terminate the tenancy request by clicking "Terminate" in the confirmation dialog box. Verify that the tenancy request has been terminated successfully. You can check the tenancy request status to ensure that it has been updated to "Terminated". It is essential to note that terminating a tenancy request will release the capacity associated with it, making it available for reclamation and reuse.

Sources

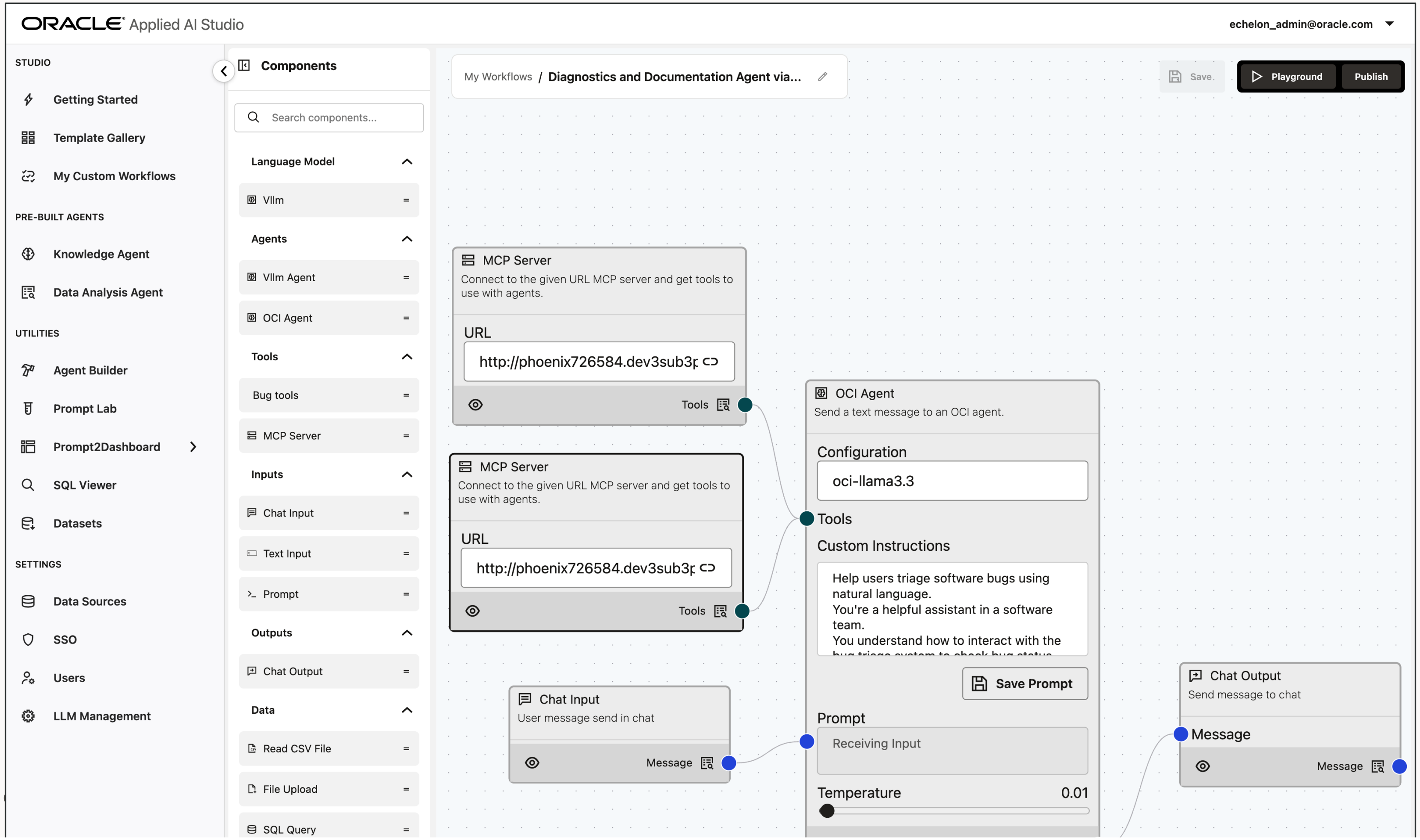
Requesting a New Commercial Subscription for Your Tenancy

Terminating an Internal Account

Capacity or Tenancy Requests

The “Three Ways of RAG” using Autonomous AI Database

Agentic RAG – build your own RAG agent



My Workflows / Diagnostics and Documentation Agent via...

MCP Server

Connect to the given URL MCP server and get tools to use with agents.

URL: http://phoenix726584.dev3sub3p

MCP Server

Connect to the given URL MCP server and get tools to use with agents.

URL: http://phoenix726584.dev3sub3p

OCI Agent

Send a text message to an OCI agent.

Configuration: oci-llama3.3

Custom Instructions:

Help users triage software bugs using natural language.
You're a helpful assistant in a software team.
You understand how to interact with the

Prompt: Receiving Input

Temperature: 0.01

Chat Input

User message send in chat

Chat Output

Send message to chat

Recap & Take Aways

1 **Oracle Autonomous AI Database helps reduce complexity**
It is multicloud, secure and scalable

2 **Multiple technology paths**
Freedom of choice between code, no-code, and agentic approach

3 **Focus on the business case of the RAG solution**
Technology makes it simple, but be very clear on what business problem you are addressing

Copyright © 2025, Oracle and/or its affiliates

RAG in a Box
Blog Article

Select AI

Vector Search

Hybrid Vector Search

Copyright © 2025, Oracle and/or its affiliates

Your feedback is important.

Scan this QR Code or use the Mobile App to share your thoughts on this session.

Copyright © 2025, Oracle and/or its affiliates

To contact us

Massimo Castelli

Doug Hood

Thank you!

Copyright © 2025, Oracle and/or its affiliates

ORACLE