



Graphs, GraphRAG and Generative AI: An Introduction

UKOUG Conference Discover 2024

Hans Viehmann

Director Product Management

Oracle Spatial and Graph

December 2nd, 2024



“You are a travel agent planning a trip to Nuremberg Convention Center. Create a list of at least 15 hotels with their name, street address, postal code, city, phone number, and website. Provide the results as a csv file.”



Using Chat interface in OCI GenAI

Model: cohere.command-r-08-2024 v1.7

The screenshot displays the Oracle Cloud Generative AI Chat interface. The left sidebar shows navigation options: Overview, Playground, Chat (selected), Generation, Summarization, Embedding, Dedicated AI clusters, Custom models, and Endpoints. The main chat area shows a conversation with the model 'cohere.command-r-08-2024 v1.7'. The user prompt is: "You are a travel agent planning a trip to Nuremberg Convention Center. Create a list of at least 12 hotels with their name, street address, postal code, city, phone number, and website. Provide the results as a csv file." The model response is a CSV file listing 12 hotels near the Nuremberg Convention Center, including Hotel Drei Raben, Hotel Victoria, Hotel Elch, Hotel Agneshof, Hotel FIVE, Hotel Deutscher Kaiser, Hotel Novotel Nürnberg Centre Ville, Hotel Mövenpick Nürnberg-Airport, Hotel InterCity Nürnberg, Hotel NH Collection Nürnberg City, Hotel Maritim Nürnberg, Hotel Sheraton Nürnberg, and Hotel Leonardo Nürnberg. The right sidebar shows parameters for the chat, including Maximum output tokens (1000), Preamble override, Temperature (0.5), Top p (0.75), Top k (0), Frequency penalty (0), and Presence penalty (0). The bottom of the interface includes a text input field, a Submit button, and a Clear chat button.

Generative AI

Chat

To get started, select a model and a preset prompt example. Then, refine the prompts and parameters to fit your use cases. All model responses have [moderation filtering](#) applied for explicit content. Note that some models have [deprecation/retirement dates](#). View our [model list](#) for more details.

Model: cohere.command-r-08-2024 v1.7 View model details Example: Choose example View code

You are a travel agent planning a trip to Nuremberg Convention Center. Create a list of at least 12 hotels with their name, street address, postal code, city, phone number, and website. Provide the results as a csv file.

Here is a list of 12 hotels near the Nuremberg Convention Center, formatted as a CSV file:

Hotel Name	Street Address	Postal Code	City	Phone Number	Website
Hotel Drei Raben	"Obristplatz 8, 90403"	90403	Nuremberg	" +49 911 274380"	https://www.hoteldreiraben.de/
Hotel Victoria	"Bahnhofstraße 15-17, 90402"	90402	Nuremberg	" +49 911 23430"	https://www.victoria-nuernberg.de/
Hotel Elch	"Königstr. 63, 90402"	90402	Nuremberg	" +49 911 223525"	https://www.hotel-elch.de/
Hotel Agneshof	"Agnesgasse 10, 90403"	90403	Nuremberg	" +49 911 22550"	https://www.hotel-agneshof.de/
Hotel FIVE	"Frauentorgraben 5, 90443"	90443	Nuremberg	" +49 911 234540"	https://www.hotel-five.de/
Hotel Deutscher Kaiser	"Königstr. 93, 90402"	90402	Nuremberg	" +49 911 22290"	https://www.deutscher-kaiser-nuernberg.de/
Hotel Novotel Nürnberg Centre Ville	"Bärenschanzstr. 8, 90439"	90439	Nuremberg	" +49 911 39900"	https://www.accor.com/c
Hotel Mövenpick Nürnberg-Airport	"Ringstr. 120, 90408"	90408	Nuremberg	" +49 911 94540"	https://www.moevenpick-hotels.c
Hotel InterCity Nürnberg	"Bahnhofstr. 1-3, 90402"	90402	Nuremberg	" +49 911 23450"	https://www.intercityhotel.com/en/hc
Hotel NH Collection Nürnberg City	"Königstr. 24, 90402"	90402	Nuremberg	" +49 911 23690"	https://www.nh-collection.com/
Hotel Maritim Nürnberg	"Frauentormauer 8, 90402"	90402	Nuremberg	" +49 911 23650"	https://www.maritim.de/en/hotels/nuer
Hotel Sheraton Nürnberg	"Königsstr. 101, 90402"	90402	Nuremberg	" +49 911 2320"	https://www.marriott.com/hotels/travel/
Hotel Leonardo Nürnberg	"Bahnhofstr. 17-19, 90402"	90402	Nuremberg	" +49 911 23990"	https://www.leonardo-hotels.com/en/

These hotels are conveniently located near the Nuremberg Convention Center and offer a range of amenities and services for travelers.

Type a message...

Submit Clear chat

Parameters

Maximum output tokens 1000

Preamble override

Temperature 0.5

Top p 0.75

Top k 0

Frequency penalty 0

Presence penalty 0

Terms of Use and Privacy Cookie Preferences

Copyright © 2024, Oracle and/or its affiliates. All rights reserved.



Using Chat interface in OCI GenAI

Model: meta.llama-3.1-405b-instruct

The screenshot displays the Oracle Cloud Generative AI Chat interface. The top navigation bar includes the Oracle Cloud logo, a search bar, and the region "US Midwest (Chicago)". The left sidebar shows the "Generative AI" section with options like Overview, Playground, Chat (selected), Generation, Summarization, Embedding, Dedicated AI clusters, Custom models, and Endpoints. The main chat area shows a conversation with the model "meta.llama-3.1-405b-instruct". The user prompt is: "You are a travel agent planning a trip to Nuremberg Convention Center. Create a list of at least 12 hotels with their name, street address, postal code, city, phone number, and website. Provide the results as a csv file." The model response is a CSV list of 12 hotels near Nuremberg Convention Center. The right sidebar shows the "Parameters" section with settings for Maximum output tokens (600), Temperature (0.5), Top p (0.75), Top k (-1), Frequency penalty (0), and Presence penalty (0). The bottom of the interface includes a text input field, "Submit" and "Clear chat" buttons, and footer text for Terms of Use, Privacy, and Copyright.

Generative AI

Chat

To get started, select a model and a preset prompt example. Then, refine the prompts and parameters to fit your use cases. All model responses have [moderation filtering](#) applied for explicit content. Note that some models have [deprecation/retirement dates](#). View our [model list](#) for more details.

Model: meta.llama-3.1-405b-instruct View model details Choose example View code

You are a travel agent planning a trip to Nuremberg Convention Center. Create a list of at least 12 hotels with their name, street address, postal code, city, phone number, and website. Provide the results as a csv file.

Here is the list of 12 hotels near Nuremberg Convention Center in CSV format:

Hotel Name	Street Address	Postal Code	City	Phone Number	Website
"Hilton Nuremberg"	"Valznerweiherstraße 200"	"90480"	"Nuremberg"	" +49 911 60030"	" https://www.hilton.com/en/hotels/germany/hilton-nuremberg-NUEHNH/index.html "
"Maritim Hotel Nürnberg"	"Frauentorgraben 11"	"90443"	"Nuremberg"	" +49 911 23630"	" https://www.maritim.com/en/hotels/germany/nuremberg/hotel-nurnberg "
"Le Méridien Grand Hotel Nuremberg"	"Bahnhofstraße 1"	"90402"	"Nuremberg"	" +49 911 23220"	" https://www.marriott.com/hotels/travel/NUEMD-Le-Meridien-Grand-Hotel-Nuremberg/ "
"Park Plaza Nuremberg"	"Bahnhofstraße 5"	"90402"	"Nuremberg"	" +49 911 94480"	" https://www.parkplaza.com/nuremberg-hotel-de-90402/germany "
"NH Collection Nürnberg City"	"Bahnhofstraße 17"	"90402"	"Nuremberg"	" +49 911 92940"	" https://www.nh-hotels.com/hotel/nh-collection-nurnberg-city "
"Ringhotel Loew's Merkur"	"Pillenreuther Straße 1"	"90459"	"Nuremberg"	" +49 911 52160"	" https://www.ringhotels.com/en/hotels/ringhotel-loews-merkur-nuremberg/ "
"Hotel Europa"	"Hauptbahnhof 1"	"90443"	"Nuremberg"	" +49 911 23250"	" https://www.hotel-europa-nuernberg.de/en/ "
"Hotel Fürth-Nürnberg"	"Würzburger Straße 45"	"90762"	"Fürth"	" +49 911 79430"	" https://www.hotelfuerth-nuernberg.de/en/ "
"Hotel Am Heidehoffplatz"	"Heidehoffplatz 7"	"90478"	"Nuremberg"	" +49 911 52270"	" https://www.hotel-am-heidehoffplatz.de/en/ "
"Congress Hotel Mercure Nürnberg an der Messe"	"Münchener Straße 283"	"90471"	"Nuremberg"	" +49 911 52150"	" https://www.accorhotels.com/gb/hotel-5325-mercure-nuernberg-an-der-messe/index.shtml "
"Novotel Nürnberg Messezentrum"	"Münchener Straße 340"	"90471"	"Nuremberg"	" +49 911 52140"	

Type a message...

Submit Clear chat

Parameters

Maximum output tokens 600

Input + output tokens should be less than 128000

Temperature 0.5

Top p 0.75

Top k -1

Frequency penalty 0

Presence penalty 0

Terms of Use and Privacy Cookie Preferences

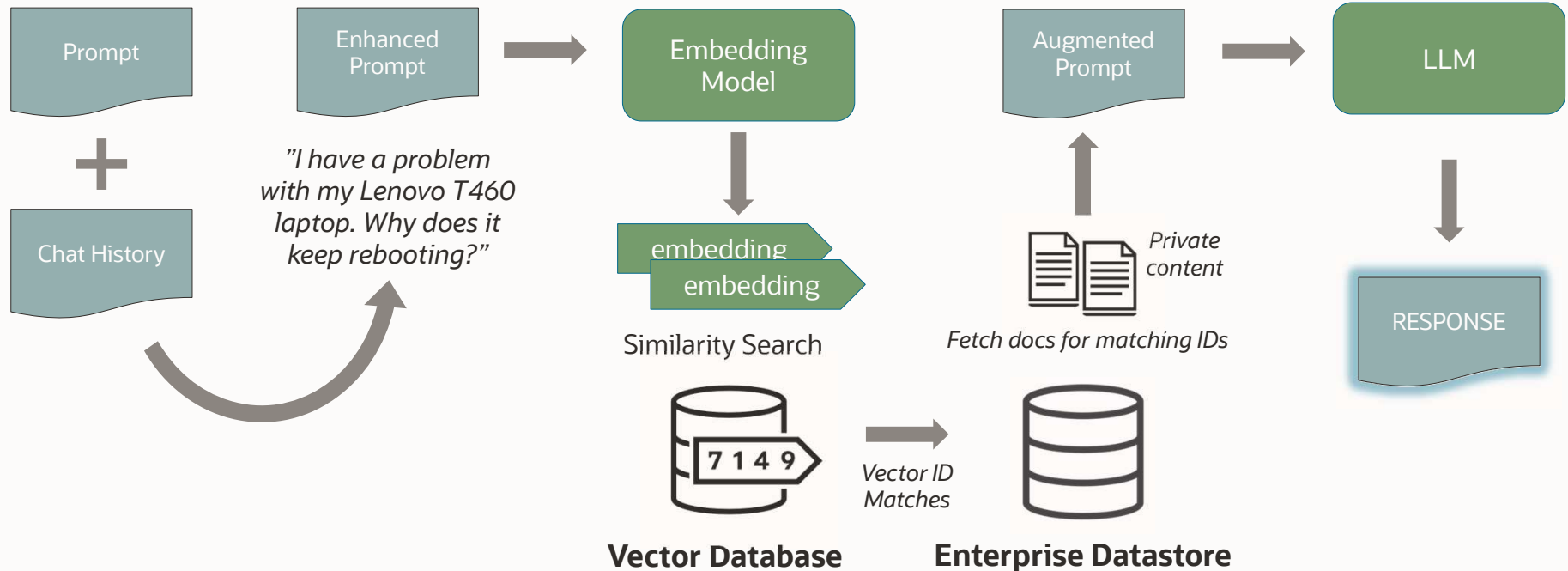
Copyright © 2024, Oracle and/or its affiliates. All rights reserved.



Retrieval Augmented Generation workflow

Supplying LLM-based chatbot with “enterprise knowledge”

"Why does my laptop keep rebooting?"



RAG based on vector search

Development

Frameworks and tools evolving rapidly

- Langchain, LlamaIndex, ...

Comprehensive support in Oracle 23ai

- Document storage
- Vector similarity search
- OML with ONNX support
- APIs for external GenAI services
- ... All available through SQL and PL/SQL

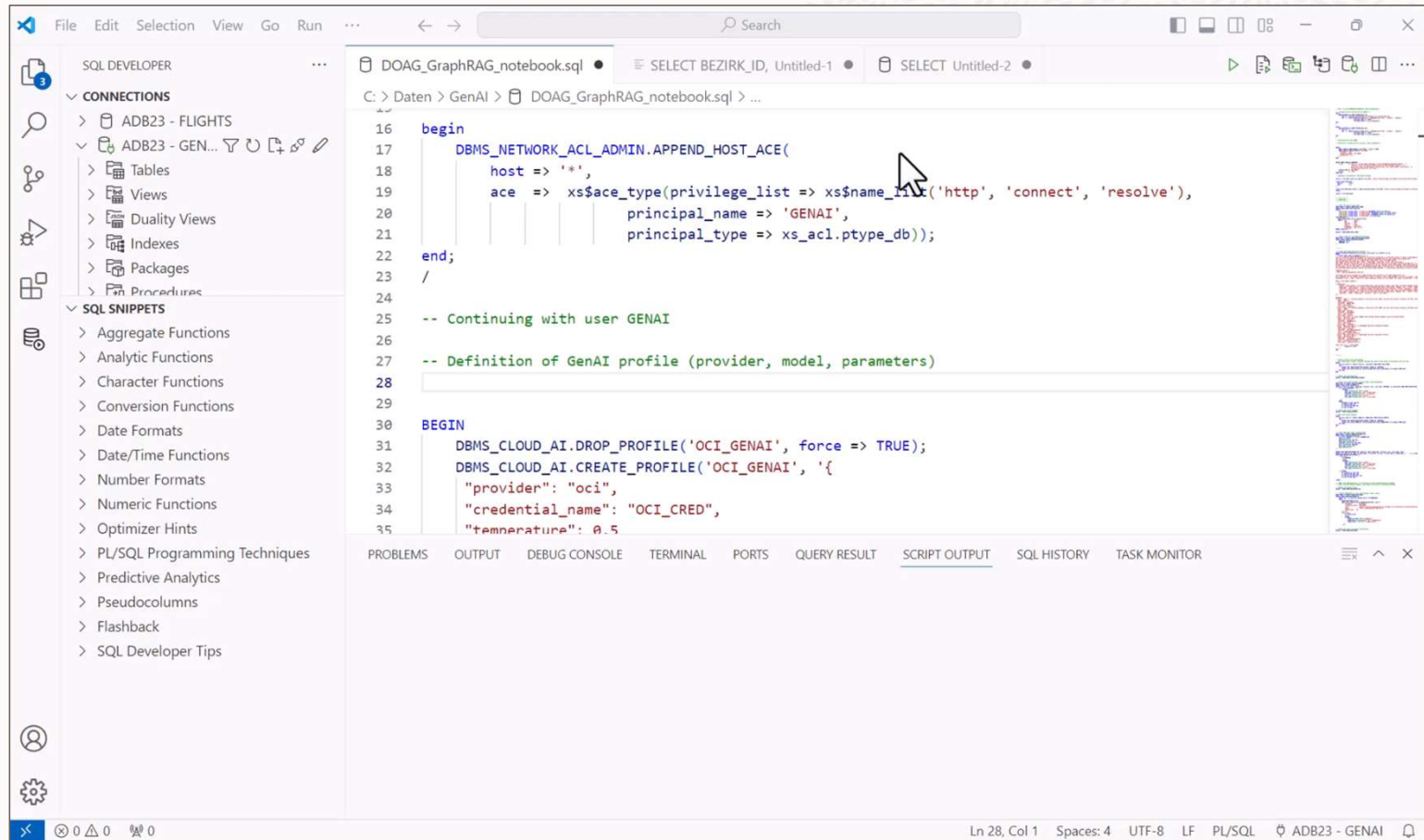
Accessing GenAI APIs from Oracle 23ai

- Allow REST callouts
- Set up credentials
- Define GenAI profile

```
SELECT DBMS_CLOUD_AI.GENERATE(  
    prompt      => 'You are a travel agent  
<...> .',  
    profile_name => 'OCI_GENAI',  
    action      => 'chat')  
FROM dual;
```



Accessing GenAI APIs from Oracle 23ai



Demo



Graph RAG

Improving baseline RAG with graphs

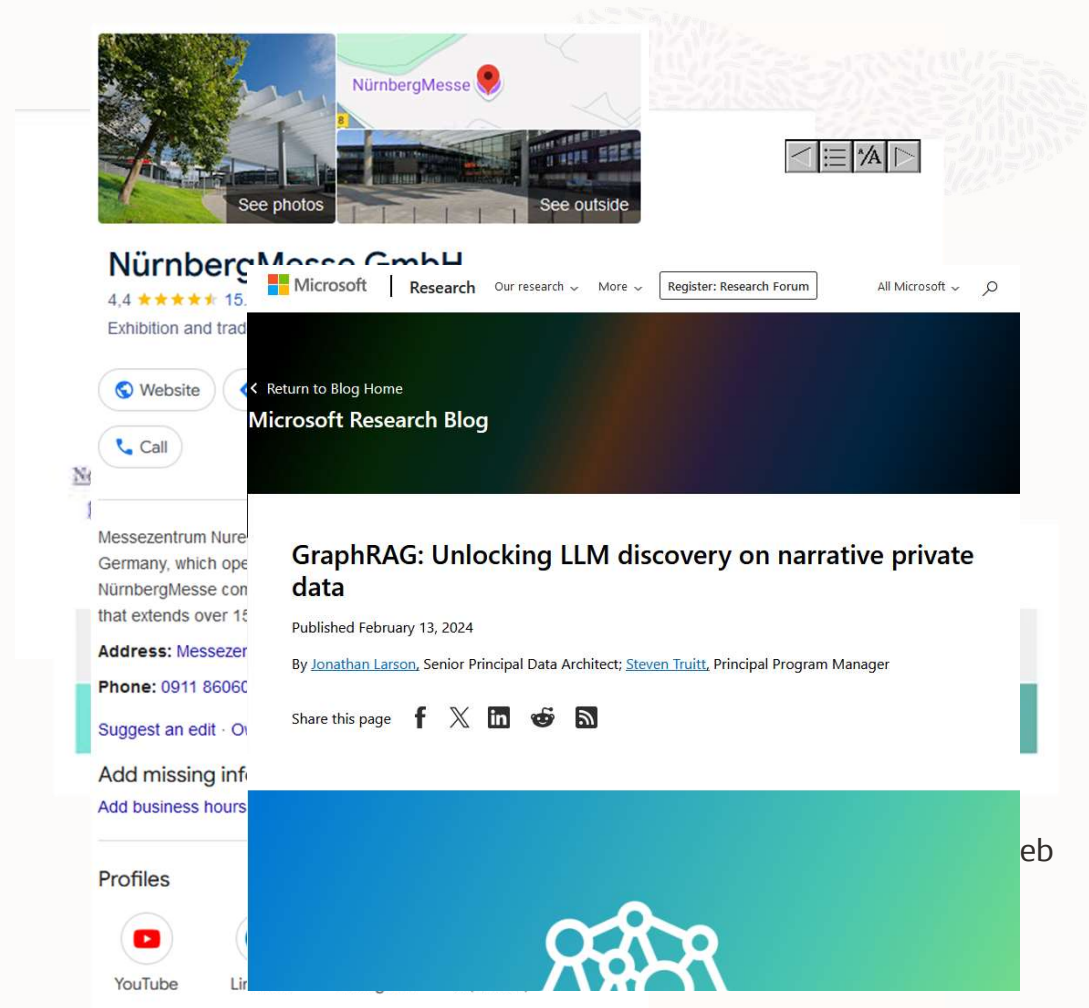


Evolution of search

- Keyword search (~1994)
- Pagerank (~1999)
- Semantic web (~1999)
- Google Knowledge Graph (2012)
- Graphs with LLMs (GraphRAG, 2024)

GraphRAG

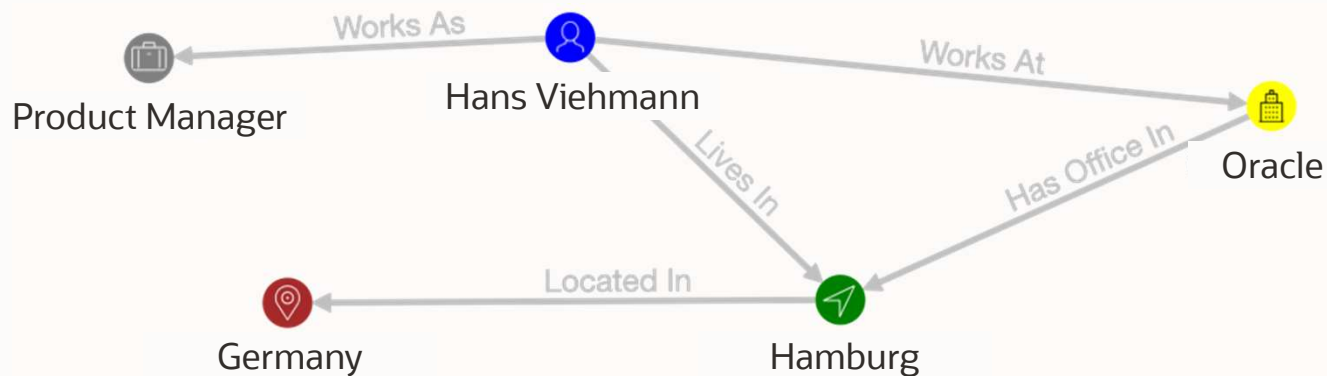
- Includes graph retrieval step as part of RAG workflow
- Not necessarily graph retrieval only



What are Knowledge Graphs?

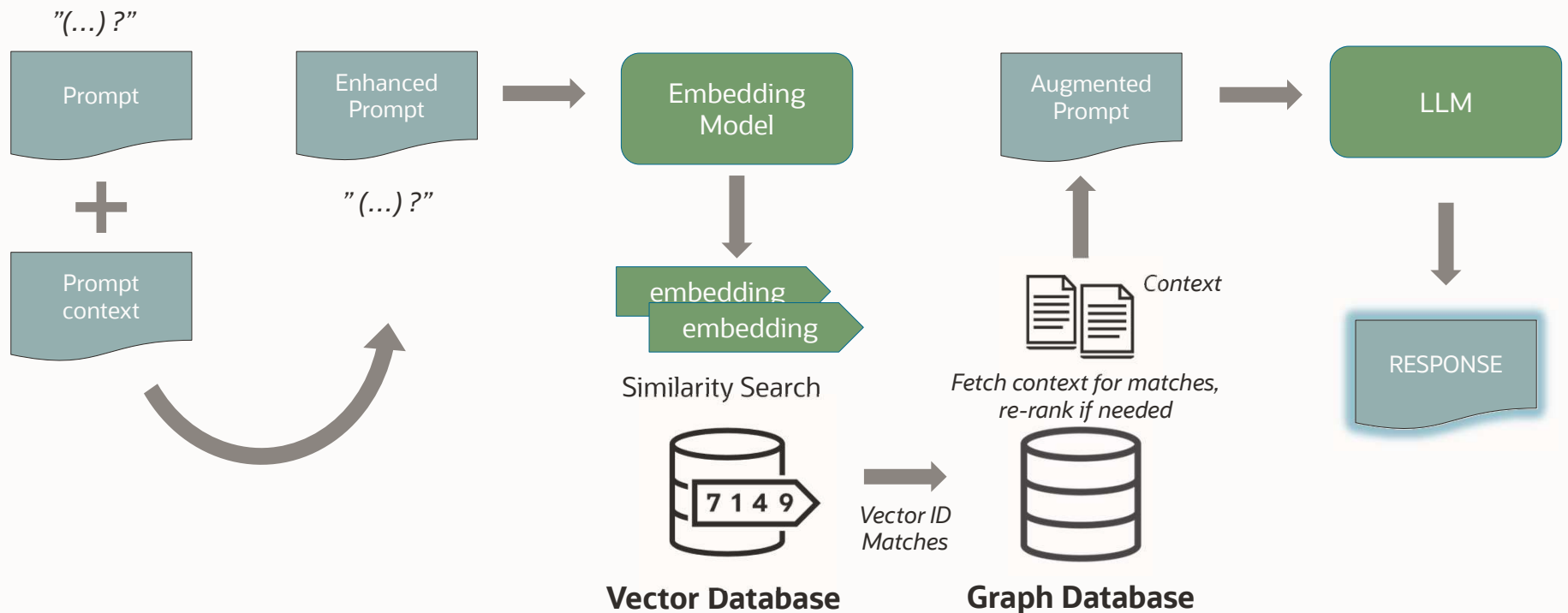
Knowledge graphs use a graph structure to represent information about entities and their relations

- Schema-flexible model

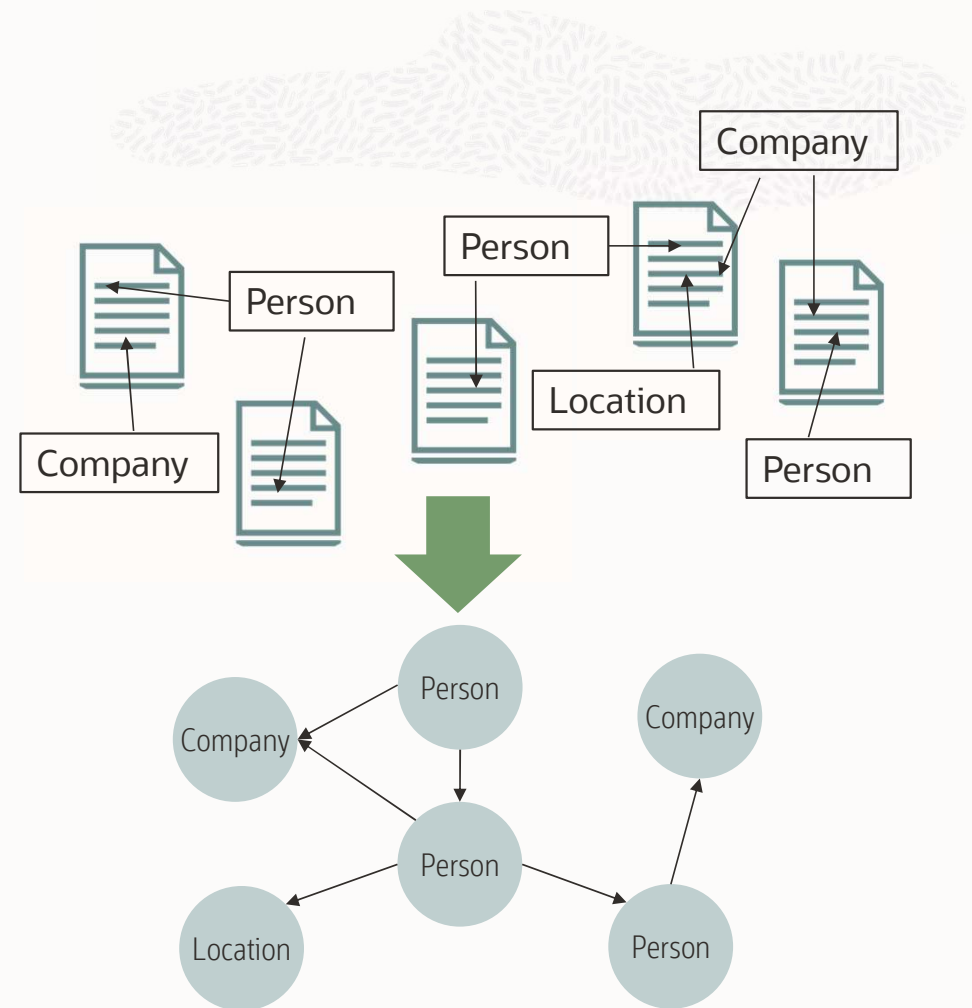
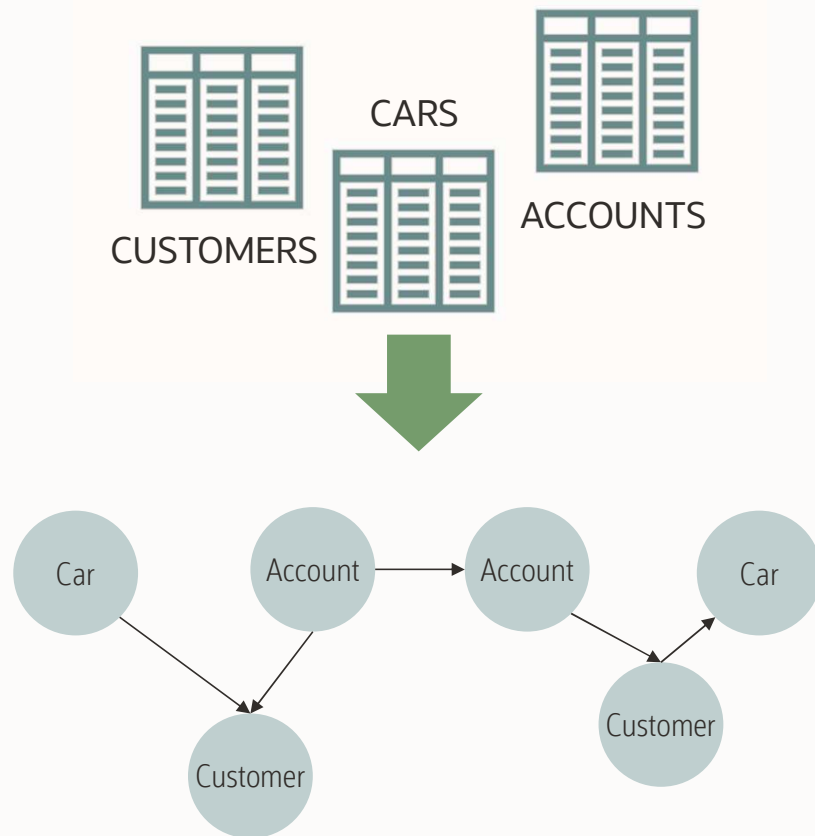


- Typical use cases include [improving search engines](#), [smart assistants](#), [detecting fraud and analyzing financial crime](#), [predicting diagnoses in healthcare](#).
- Expressive data model, powerful tool for data integration, unification, and analysis.

GraphRAG Workflow: LLM-based chatbot using Knowledge Graph

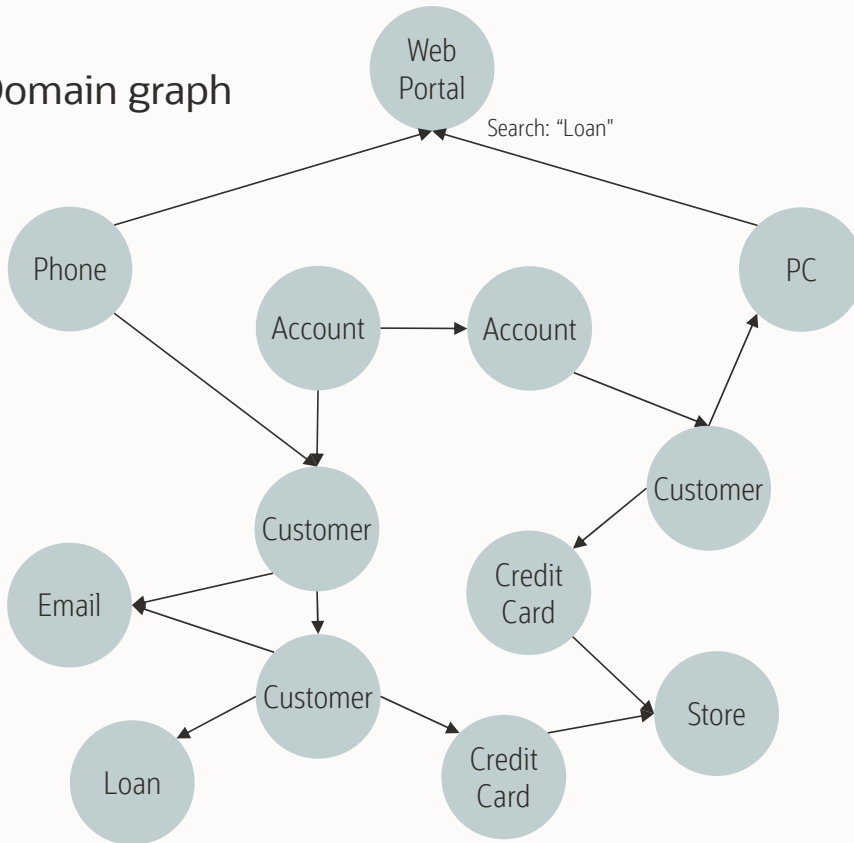


Where do graphs come from?

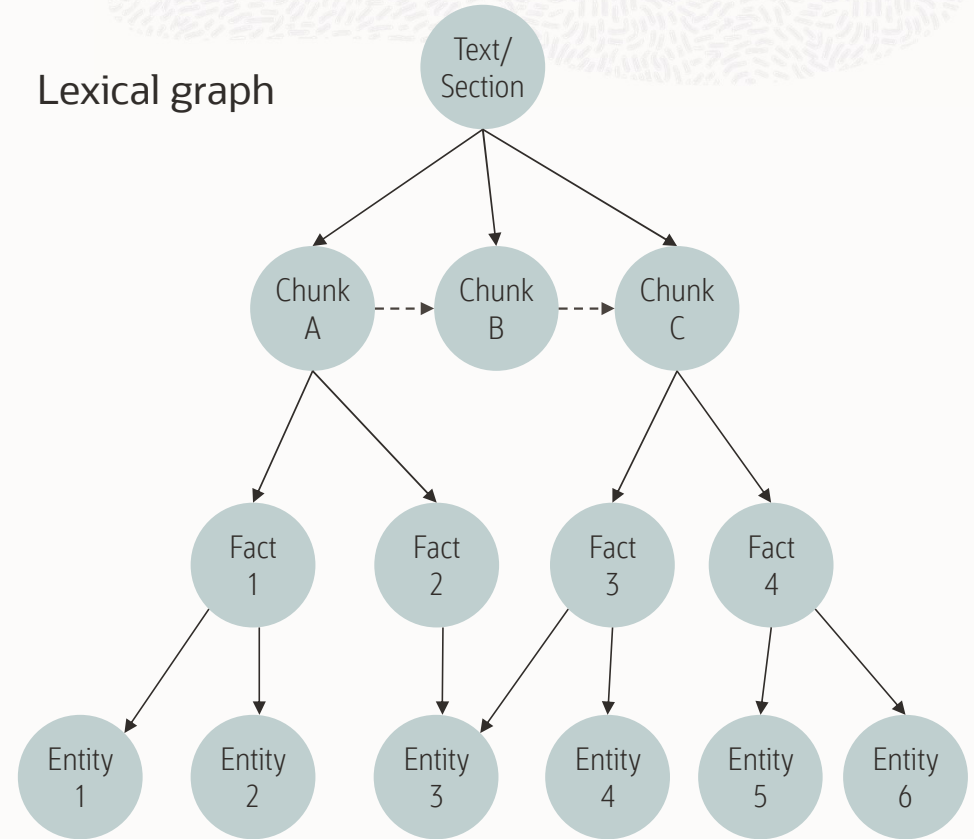


Exatracting graphs from documents

Domain graph



Lexical graph



Benefits of using graphs for RAG

Accuracy

Retrieval-Augmented Generation with Knowledge Graphs for Customer Service Question Answering

Microsoft Research Blog

GraphRAG: Unlocking data

Published February 13, 2024

By [Jonathan Larson](#), Senior Principal

Share this page [f](#) [X](#) [in](#)

data.world Product Solutions Customers Developers Resources Company Sign In Get a demo

The Knowledge Graph Difference

MIT News
ON CAMPUS AND AROUND THE WORLD

Graph-based AI model maps the future of innovation

An AI method developed by Professor Markus Buehler finds hidden links between science and art to suggest novel materials.

Stephanie Martinovich | Department of Civil and Environmental Engineering
November 12, 2024

Knowledge Graph improved accuracy of LLM responses average of 3x.

constraints of embedding models; an embedding vector for retrieval. During phase, the system retrieves Is them as contexts for LLMs to eries. Despite its straightforward several limitations:

- **Limitation 1 - Compromised Retrieval Accuracy from Ignoring Structures:** Issue tracking documents such as Jira [2] possess inherent structure and are interconnected, with references such as "issue A is related to/copied from/caused

and/or a fee. Request permissions from permissions@acm.org.
SIGIR '24, July 14–18, 2024, Washington, DC, USA
© 2024 Copyright held by the owner/authors. Publication rights licensed to ACM.
ACM ISBN 978-0-4007-0431-4/24/007
<https://doi.org/10.1145/3626772.3661370>

▼ PRESS INQUIRIES

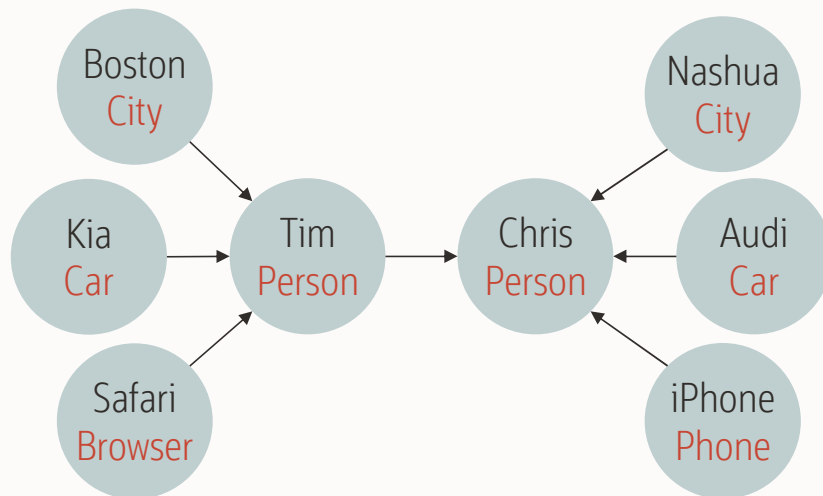


Benefits of using graphs for RAG

Explainability

Graph query

- Explicit representation of relationships
- Takes semantics into account
- Search based on graph traversal



Vector similarity search

- Opaque data structure
- Different distance metrics
- Search based on approximate nearest neighbor analysis

Tim: [0.247, 0.889, 0.144, 0.183,...]

Chris: [0.551, 0.795, 0.201, 0.229,...]

Why use Oracle for GraphRAG?

Converged database

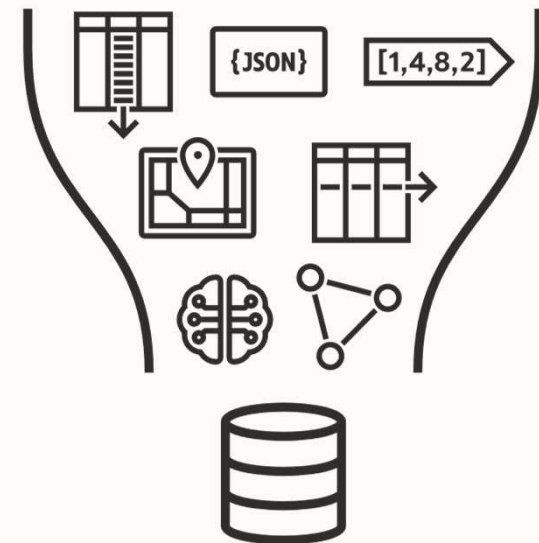
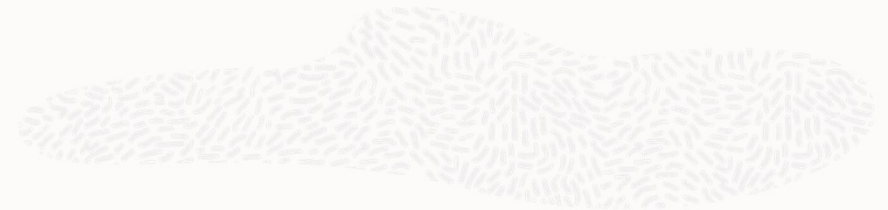
- Graph support natively in the database kernel
 - Graph visualization for validation, debugging
- Vector similarity search built in
- Document store (Text, JSON, XML, ...)

Integration of external services

- Secure REST calls from database
- PL/SQL APIs

OCI GenAI Services

- Various LLM services and options
- Comfortable tooling

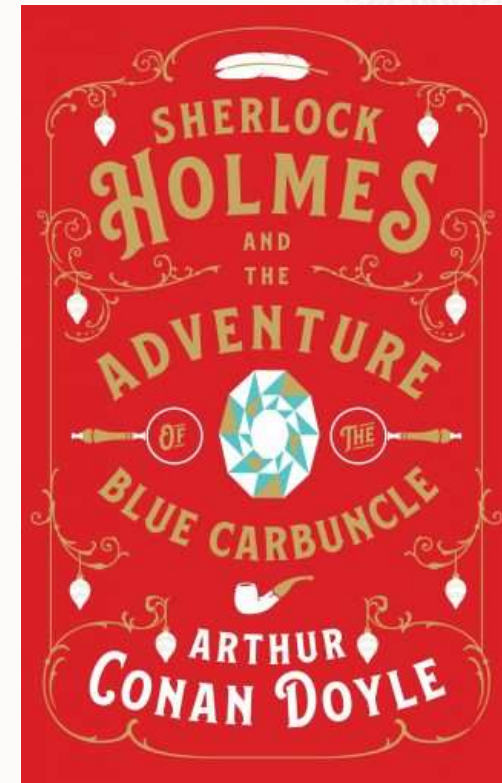


Demo scenario

Analysis of Sherlock Holmes story

Demo flow:

- Load text from object storage
- Split into chunks
- Extract relationships from each chunk in JSON format using LLM with self-defined prompt
- Extract entities, relationships and respective types from each JSON string
- Compute embedding for JSON string using embedding model, create vector index
- Create property graph from entities and relationships for graph queries
- Compute query embedding and select most relevant chunks
- Use LLM to generate answer



Demo



Using the new Generative AI Agents in OCI GenAI

The screenshot displays the Oracle Cloud console dashboard for a user in the US Midwest (Chicago) region. The interface includes a top navigation bar with the Oracle Cloud logo, a search bar, and a dropdown menu for the region. The main content area is divided into several sections:

- Service links:** A section with three columns of links. The first column, labeled 'PINNED', includes links for Autonomous Database, Instances Compute, and Virtual cloud networks Networking. The second column, labeled 'RECENTLY VISITED', includes links for Buckets Object Storage & Archive Storage, Generative AI Agents AI Services, Generative AI AI Services, Policies Identity, Limits, Quotas and Usage, Tenancy Management, and Domains Identity. The third column, labeled 'RECOMMENDED - (Update)', includes links for Logging Logging and Tenancies Organization Management. A mouse cursor is hovering over the 'Tenancies Organization Management' link.
- Quickstarts:** A section with six cards, each representing a quickstart guide. The cards are: 'Deploy a WordPress website' (6-8 mins), 'Deploy the CIS Landing Zone' (15-20 mins), 'Deploy a low-code app on Autonomous Database using APEX' (3-5 mins), 'Deploy RStudio in a container', 'Deploy a baseline landing', and 'Visualize and analyze Strava'. Each card includes an icon representing the application or service.
- Tenant status:** A section on the right side of the dashboard showing the tenant name 'oradbcloudcm' and a status indicator 'All services operational'. It also displays 'Cost savings opportunities' with an estimated savings of 0 and a link to 'View recommendations (823)'. Below this, there are sections for 'Experience previews' (New) and 'Marketplace' (New), each with a brief description and a link to learn more.

The bottom of the dashboard includes a footer with 'Terms of Use and Privacy' and 'Cookie Preferences' links, and a copyright notice: 'Copyright © 2024, Oracle and/or its affiliates. All rights reserved.'



Using the new Generative AI Agents in OCI GenAI

The screenshot displays the OCI GenAI Chat interface. The top navigation bar includes a search bar and a location dropdown set to 'US Midwest (Chicago)'. The left sidebar shows the 'Generative AI Agents' section with options for Overview, Agents, Knowledge Bases, and Chat. The main chat area is titled 'Chat' and shows a conversation with the 'DOAG Agent' at endpoint 'genaiaagentendpoint2024111...'. The chat history includes a user query 'Who stole the jewel?' and a detailed response from the agent. The response text is: 'James Ryder, the upper-attendant at the Hotel Cosmopolitan, and Catherine Cusack, the Countess of Morcar's waiting maid, conspired to steal the jewel. They framed John Horner, a plumber, by planting the jewel in his possession and alerting the authorities.' Below the response, there are three citations from 'The Adventure of the Blue Carbuncle.txt', each with an object storage path and a source text snippet. The right sidebar shows a 'Logs' section with a 'Close logs' button, displaying a retrieval request log and a generation request log. The bottom of the interface includes a 'Type a message...' input field, a 'Submit' button, and a 'Reset chat session' button. The footer contains 'Terms of Use and Privacy', 'Cookie Preferences', and a copyright notice for 2024.

Cloud

Search resources, services, documentation, and Marketplace

US Midwest (Chicago)

Generative AI Agents

Overview

Agents

Knowledge Bases

Chat

Agent compartment

SpatialAndGraph

oradbcloudum (root)/SpatialAndGraph

Agent endpoint compartment

SpatialAndGraph

oradbcloudum (root)/SpatialAndGraph

Chat

Agent

DOAG Agent

Agent endpoint

genaiaagentendpoint2024111...

Chatting with DOAG Agent

Who stole the jewel?

James Ryder, the upper-attendant at the Hotel Cosmopolitan, and Catherine Cusack, the Countess of Morcar's waiting maid, conspired to steal the jewel. They framed John Horner, a plumber, by planting the jewel in his possession and alerting the authorities.

View citations

1. The Adventure of the Blue Carbuncle.txt

Object storage path: ...buncle.txt Show Copy

Source text:

'That is the reward, and I have reason to know that there are sentimental considerations in the background which would induce the Countess to part with half her fortune, if she could but recover the g...

Show more

2. The Adventure of the Blue Carbuncle.txt

Object storage path: ...buncle.txt Show Copy

Source text:

'I had a friend once called Maudsley, who went to the bad, and has just been serving his time in Pentonville. Once day he had met me, and fell into talk about the ways of thieves and how they could ge...

Show more

3. The Adventure of the Blue Carbuncle.txt

Object storage path: ...buncle.txt Show Copy

Source text:

Logs

Close logs

2024-11-13T16:19:24.407Z: Retrieval request

Search string: "Who stole the jewel?"

> Sources retrieved

2024-11-13T16:19:27.253Z: Generation request

Generated text:

James Ryder, the upper-attendant at the Hotel Cosmopolitan, and Catherine Cusack, the Countess of Morcar's waiting maid, conspired to steal the jewel. They framed John Horner, a plumber, by planting t...

Show more

Type a message...

Submit

Reset chat session

Terms of Use and Privacy

Cookie Preferences

Copyright © 2024, Oracle and/or its affiliates. All rights reserved.



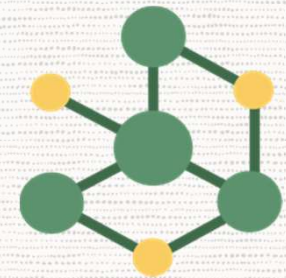
Wrap-up

Summary

GraphRAG produces more accurate,
explainable results than baseline RAG

Using Oracle 23ai and Oracle Graph simplifies
development of GraphRAG workflows

GraphRAG as a technique offers huge
potential and is evolving rapidly



Further information on Oracle Graph

Oracle Graph technologies landing page on oracle.com

- www.oracle.com/database/graph

Videos on Youtube

- www.youtube.com/c/OracleSpatialandGraph

Oracle LiveLabs: Search for „graph“ on

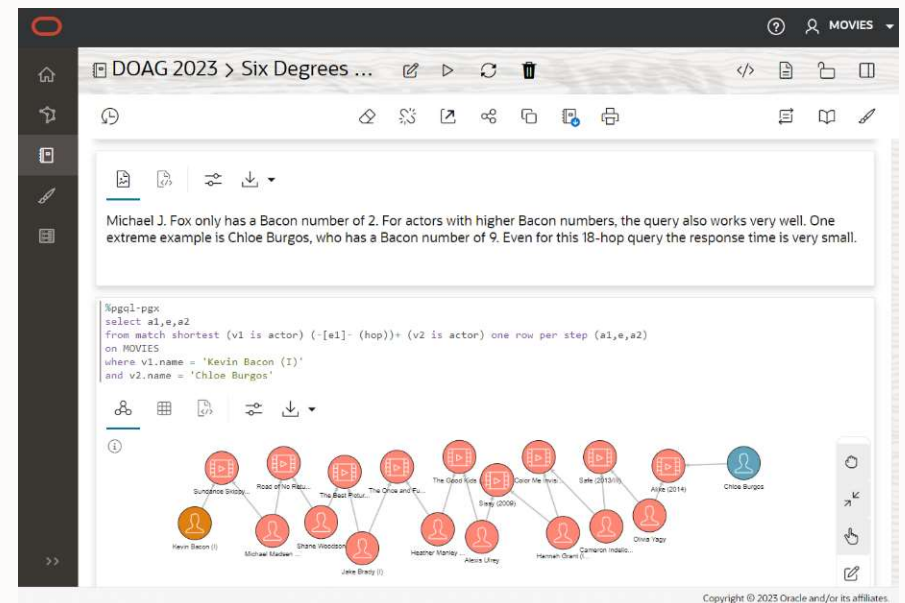
- livelabs.oracle.com/

Blogs

- [Database Insider \(category: Graph\)](#)
- [medium.com \(tag: Oracle Graph\)](#)

Documentation

- [Oracle Property Graph documentation](#)



Appendix

GraphRAG use case in Financial Services

Use Case

Case Graph Construction from Suspicious Activity Report (SAR)

Introduction:

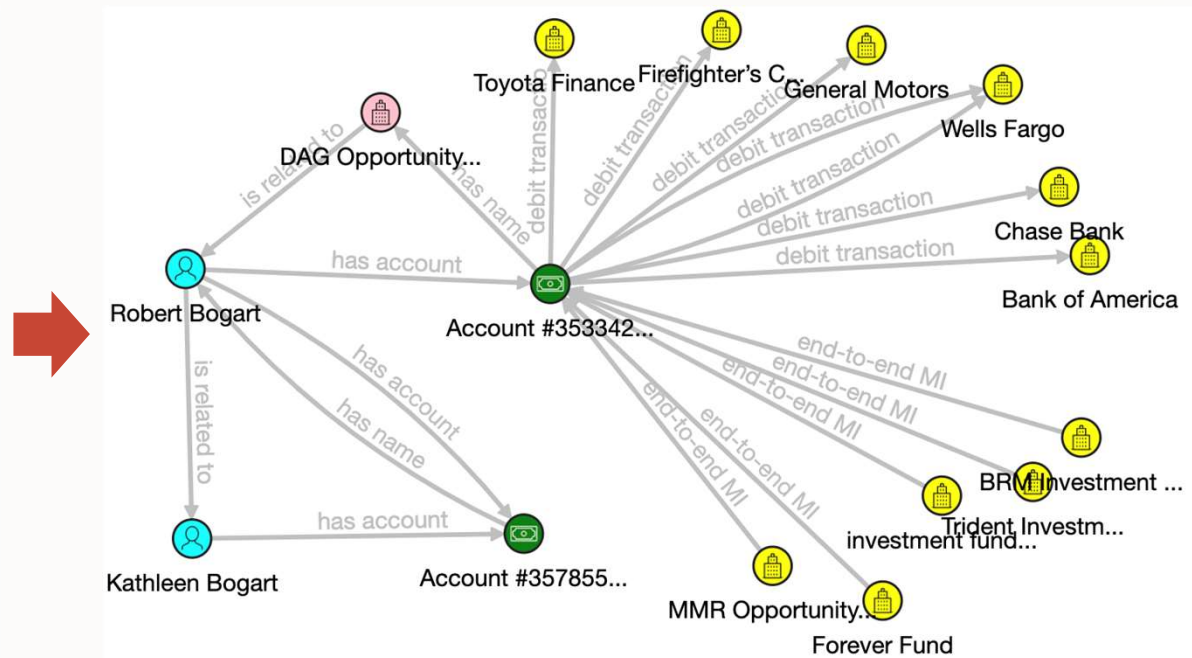
This case was referred for investigation by the MyBank AML Detection unit. The referral identifies potential cash deposit structuring activity in account number 353342287 in the name of DAG Opportunity Fund, LLC.

This investigation, which covers the time period of 1/1/20 through 7/6/20 (account closure), revealed suspicious activity totaling \$1,399,185.00, occurring between 1/8/20 and 6/18/20.

Details of Investigation:

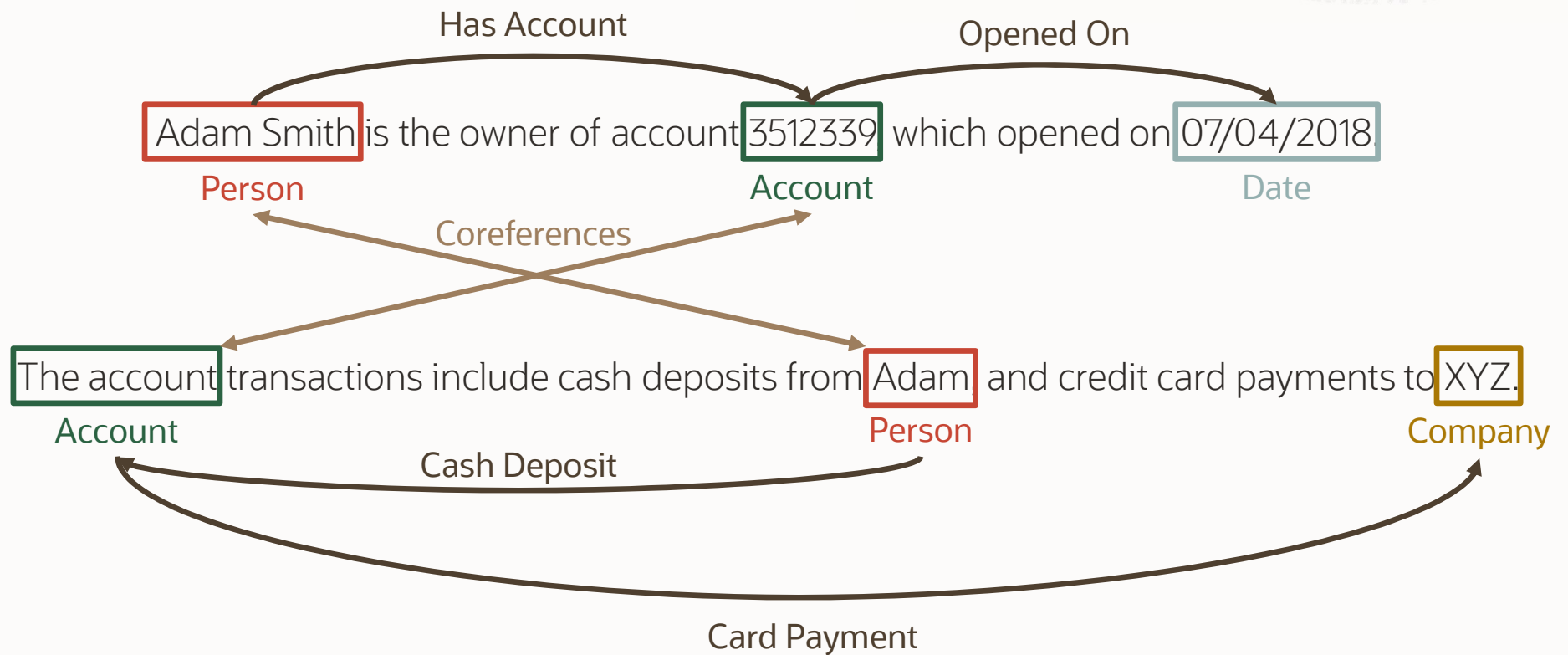
According to internal bank records, DAG Opportunity Fund, LLC is a pooled | investment account. The account signer, Robert Bogart, age 55, along with his spouse, Kathleen, age 48, are listed as an investment broker and schoolteacher respectively, and maintain the following account relationship with MyBank:

Account #353342287 – DAG Opportunity Fund, LLC - opened 4/16/15 – closed 7/6/20 - signer is Robert Bogart. Credits to this MyBank Business Premier Checking account are cash deposits, and checks drawn against the accounts of various individuals and other investment fund accounts. A considerable number of cash deposits, totaling \$96,400.00, appear to be structured to avoid the filing of a Currency Transaction Report (CTR) and are therefore deemed suspicious. A sample of the structuring activity is as follows:



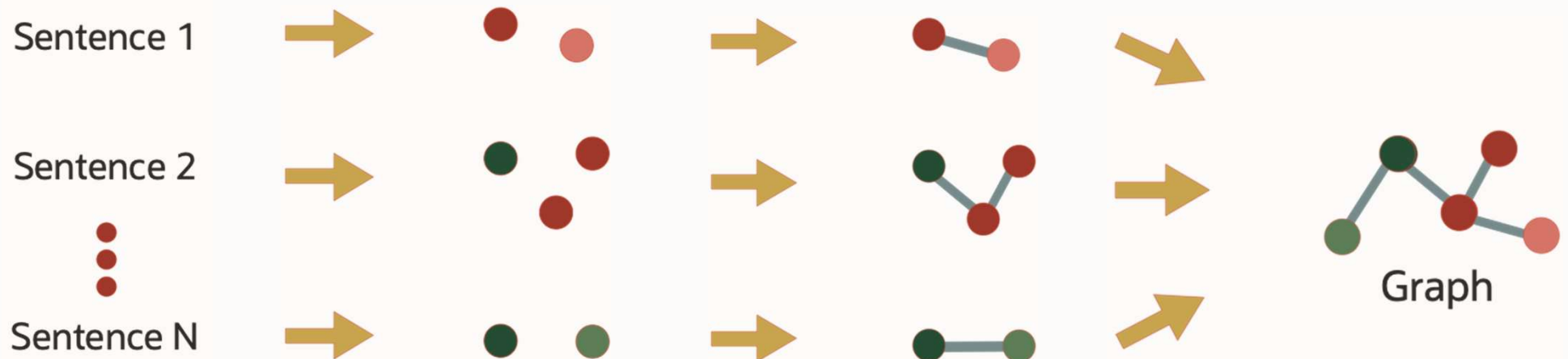
Tasks to Solve

Example



Pipeline Approach

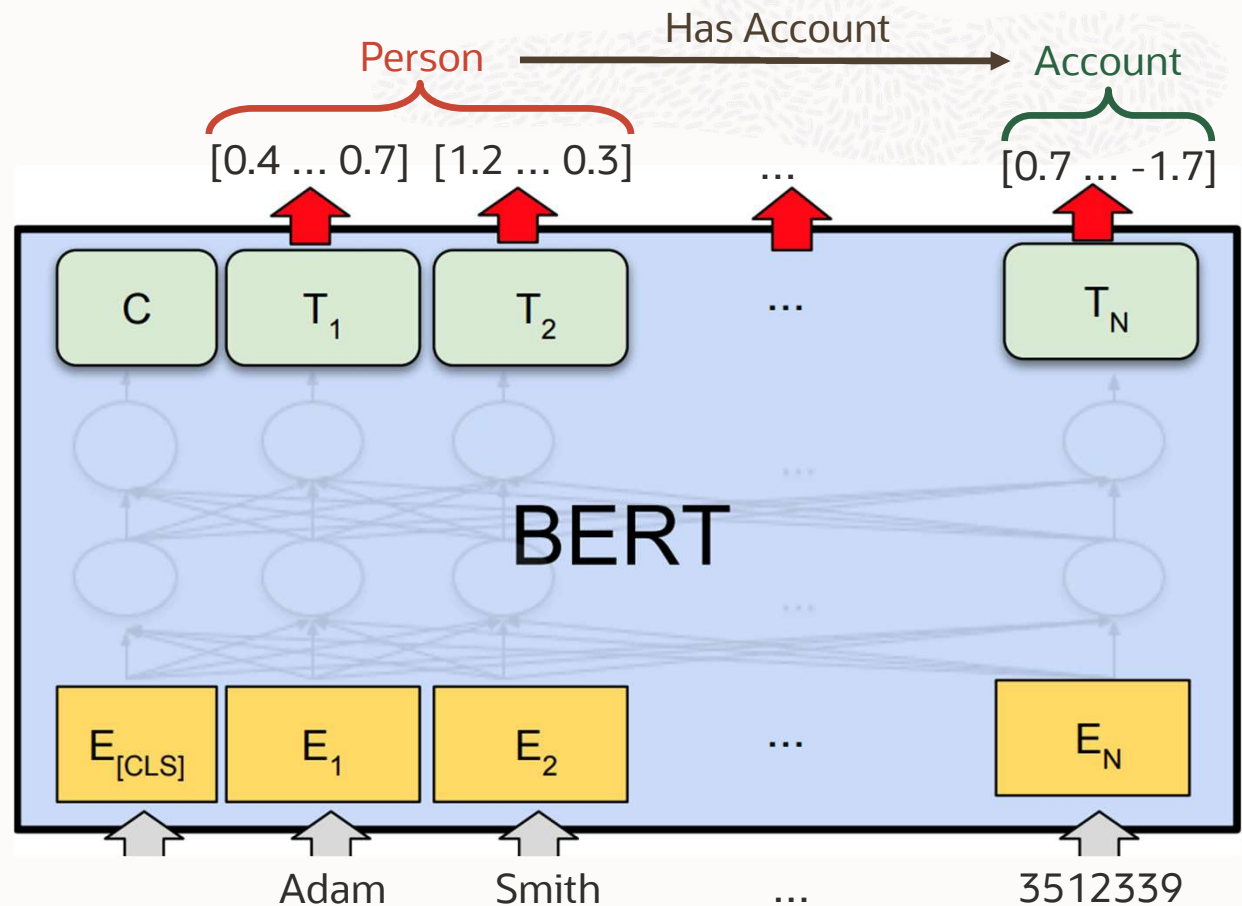
Overview



Pipeline Approach

Models

- We built three models, one for each of the tasks in the pipeline.
- All models are based on pre-trained transformer models.
- Transformer models have 100s of millions of parameters, pre-trained on a language modeling task using large datasets.
- They capture information about syntax and semantics of natural language.
- Simple classifiers on generated embeddings are enough to solve all three tasks **after fine-tuning**.



Pipeline Approach

Data Annotation

Introduction:

This case was referred for investigation by the MyBank AML Detection unit. The referral identifies potential cash deposit structuring activity in account number 353342287 in the name of DAG Opportunity Fund, LLC.

This investigation, which covers the time period of 1/1/20 through 7/6/20 (account closure), revealed suspicious activity totaling \$1,399,185.00, occurring between 1/8/20 and 6/18/20.

Details of Investigation:

According to internal bank records, DAG Opportunity Fund, LLC is a pooled investment account. The account signer, Robert Bogart, age 55, along with his spouse, Kathleen, age 48, are listed as an investment broker and schoolteacher respectively, and maintain the following account relationship with MyBank:

Account #353342287 – DAG Opportunity Fund, LLC - opened 4/16/15 – closed 7/6/20 - signer is Robert Bogart. Credits to this MyBank Business Premier Checking account are cash deposits, and checks drawn against the accounts of various individuals and other investment fund accounts. A considerable number of cash deposits, totaling \$96,400.00, appear to be structured to avoid the filing of a Currency Transaction Report (CTR) and are therefore deemed suspicious. A sample of the structuring activity is as follows:

Introduction:

This case was referred for investigation by the MyBank AML Detection unit.

The referral identifies potential **cash** deposit structuring activity in account number **353342287** in the name of DAG Opportunity Fund, LLC.

This investigation, which covers the time period of 1/1/20 through 7/6/20 (account closure), revealed suspicious activity totaling

\$1,399,185.00, occurring between 1/8/20 and 6/18/20.

Details of Investigation:

According to internal bank records, DAG Opportunity Fund, LLC is a pooled investment account.

The account signer, Robert Bogart, age 55, along with his spouse, Kathleen, age 48, are listed as an investment broker and schoolteacher respectively, and maintain the following account relationship with MyBank:

Account #353342287 – DAG Opportunity Fund, LLC - opened 4/16/15 – closed 7/6/20 - signer is Robert Bogart. Credits to this MyBank Business Premier Checking account are cash deposits, and checks drawn against the accounts of various individuals and other investment fund accounts.

A considerable number of cash deposits, totaling \$96,400.00, appear to be structured to avoid the filing of a Currency Transaction Report (CTR) and are therefore deemed suspicious.

A sample of the structuring activity is as follows:

To train the pipeline, entity + relation + coreference annotations are needed to be done manually on a number of documents using an annotation tool, such as:

brat



Pipeline Approach

Results

On benchmark dataset SciERC (**Scientific Entity Relation Coreference**):

NAMED ENTITY RECOGNITION	RELATION EXTRACTION	COREFERENCE RESOLUTION
70.1	50.5	60.2

F1 Scores

There are now better scores published for individual tasks on this dataset, e.g.:

- **NER:** 71.1 ([Jeong and Kim, 2022](#))
- **RE:** 51.3 ([Santosh et al., 2021](#))

But the state-of-the-art has not progressed too far.

On our financial dataset SAR (more structured documents):

NAMED ENTITY RECOGNITION	RELATION EXTRACTION	COREFERENCE RESOLUTION
88.82	81.26	87.06

F1 Scores

SAR Graph Quality Metrics:

NODES	EDGES
83.12	82.13

F1 Scores



Generalizability & Annotation Requirements

While the pipeline approach works well, it requires **fine-tuning three models on many manually-annotated documents**. SAR dataset has 105 annotated documents and SciERC has 500.

Fine-tuned models end up being specific to the domain and schema they were trained with. This means for each new use case we need to retrain on a set of annotated documents for that use case.

Most stakeholders don't have such a dataset and would prefer not to go through the annotation process. So we have been focusing our research on reducing these annotation requirements.

Appendix

Other Ways of Using Graphs with GenAI

Generate Graph Queries



- Use LLMs to translate human language queries into graph query language
- If the LLM is not trained to formulate such queries, they can have the capacity to learn with a few examples ('few-shot' learning)
- Eliminates the need for developers to learn new syntax

Steps

- Embed as many queries as possible in a vector database
- From a user query identify the top k matches and retrieve the corresponding graph queries
- Use these related pairs as prompt examples to instruct our LLM to generate SPARQL queries

Semantic Knowledge Graphs based on RDF in RAG

Generating SPARQL queries using OCI GenAI Service

- Graphs can be complex to navigate, often requiring the use of query languages like SPARQL
- Typically, LLMs are not directly trained to formulate these queries. However, their capacity for learning from a few examples, or 'few-shot learning,' especially through prompt engineering, is a game-changer.
- By adopting prompt-based few-shot learning, we can teach LLMs to write SPARQL queries with a high level of proficiency.

Strategy:

1. **Indexing:** Cataloging all the queries in our dataset into a vector store to streamline the retrieval process.
2. **Match User Queries:** When a user submits a query, we identify the most closely related queries in our index.
3. **Retrieval:** We then extract the corresponding SPARQL queries for these matches.
4. **Educate the LLM:** Finally, we use these related pairs as prompt examples to instruct our LLM in generating accurate SPARQL queries.
5. **Final Answer Generation:** After obtaining the SPARQL query, we run it against the graph to gather the necessary results. These results, combined with the initial user query are then fed to the LLM to craft a well-rounded final response.



Thank you

oracle.com/database/graph/



Our mission is to help people see
data in new ways, discover insights,
unlock endless possibilities.



ORACLE

