



Powering Generative AI Apps With Redis Cloud and Amazon Bedrock

April 11, 2024
Boston

Introductions



- Solutions Architect
Manish Arora
Redis



- Senior Partner
Solutions Architect
Antony Prasad Thevaraj
AWS

Agenda

01 Intro to Redis 1:00 - 1:45 p.m.

02 Break 1:45 - 2:00 p.m.

03 Intro to Amazon Bedrock 2:00 - 2:45 p.m.

04 Break 2:45 - 3:00 p.m.

05 Hands-on lab 3:00 - 4:30 p.m.

06 Open discussion 4:30 - 5:00 p.m.

How familiar are you with Redis?

1. Redis? Sounds like a mystery!! 🕵️
2. Redis in the Grapevine, but not in my toolbox! 🍇
3. A dash of sporadic usage! ↲
4. Redis: My trusted sidekick in the daily grind! 🚀



How familiar are you with Vector Stores?

1. Vector Stores: alien territory! 🚂
2. Heard, but haven't explored! 🔎
3. Dipping my toes in! 🌊
4. My everyday arsenal! 🔥



Devs love us. Apps need us.

● MOST ADMIRE

NoSQL Database,
2023



● MOST POPULAR

Container Images -
2020, 2021, 2023

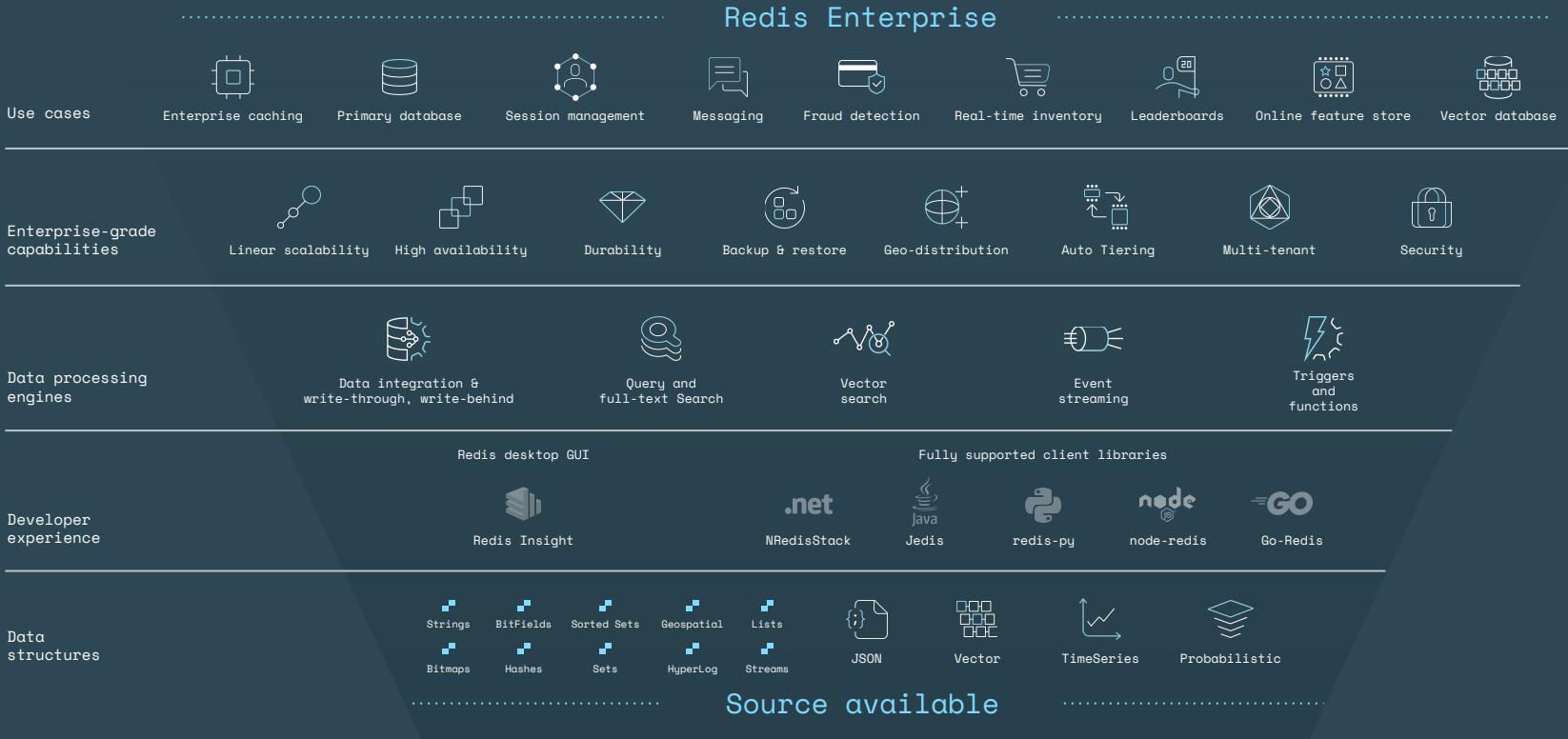


● MOST LAUNCHED

Database
technology



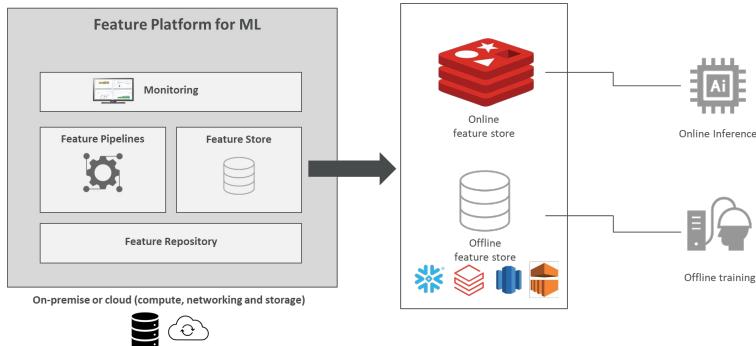
We made the most-used database even better.



Redis in AI/ML Space

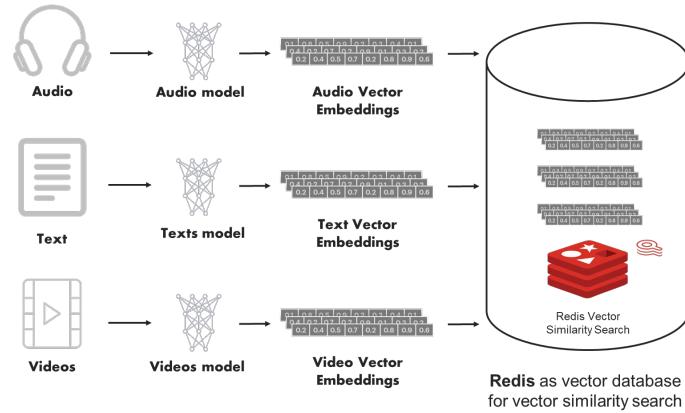
Single-digit millisecond Feature Retrieval

Redis as an online Feature Store



A composable platform for Intelligent Applications

Redis as a Vector Database



What is Generative AI?

It is type of artificial intelligence that focuses on creating new content, like text, images, music, audio, and videos.

It is powered by Large AI models(a.k.a. foundation models).

Generative AI apps

APPLICATION LAYER	Marketing (content)							
	Sales (email)	Code generation	Image generation					Gaming
	Support (chat / email)	Code documentation	Consumer / Social					RPA
	General writing	Text to SQL	Media / Advertising					Music
	Note taking	Web app builders	Design	Voice Synthesis	Video editing / generation			Audio
	Other					3D models / scenes		Biology & chemistry
	TEXT	CODE	IMAGE	SPEECH	VIDEO	3D	OTHER	

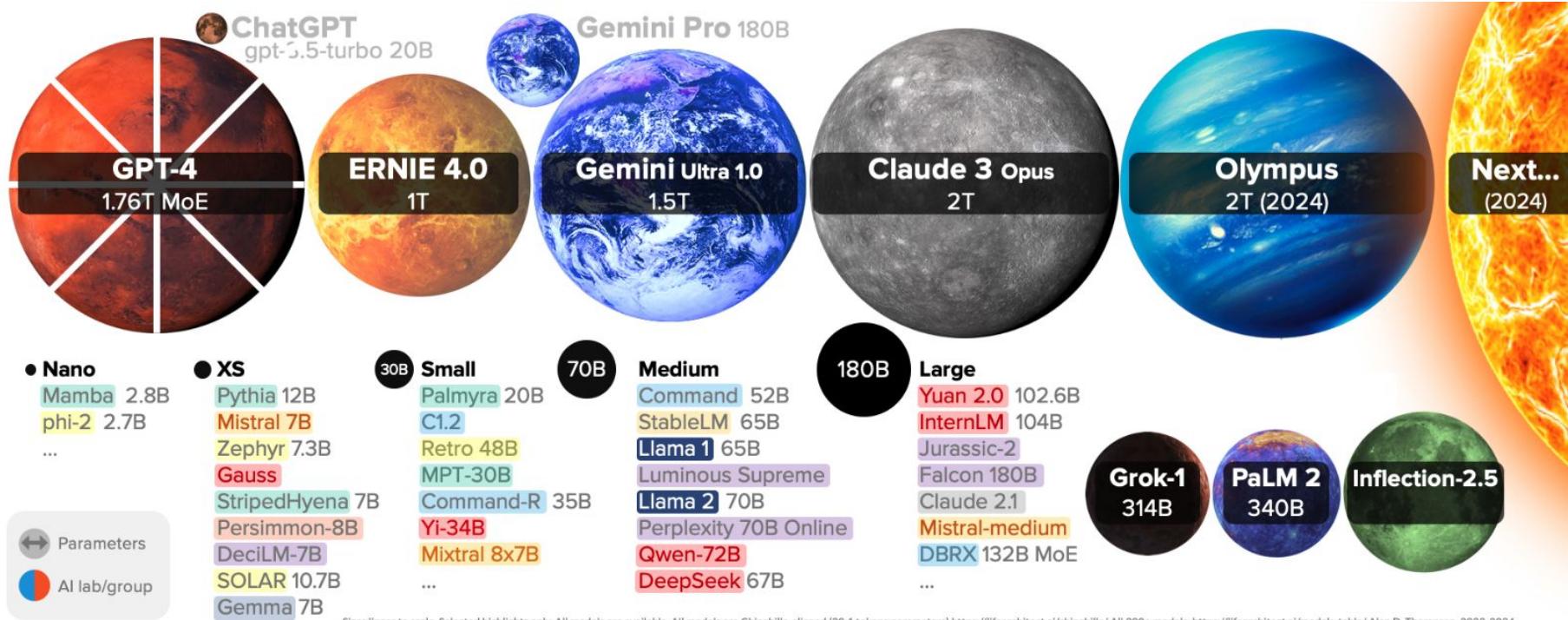
Source: <https://anjanasusarla.substack.com/p/the-second-coming-or-slouching-towards>

Generative AI is powered by

	TEXT	CODE	IMAGE	SPEECH	VIDEO	3D	OTHER
MODEL LAYER	OpenAI GPT-3	OpenAI GPT-3	OpenAI Dall-E 2	OpenAI	Microsoft X-CLIP	DreamFusion	TBD
DeepMind Gopher	Tabnine	Stable Diffusion			Meta Make-A-Video	NVIDIA GET3D	
Facebook OPT	Stability.ai	Craiyon				MDM	
Hugging Face Bloom							
Cohere							
Anthropic							
AI2							
Alibaba, Yandex, etc.							

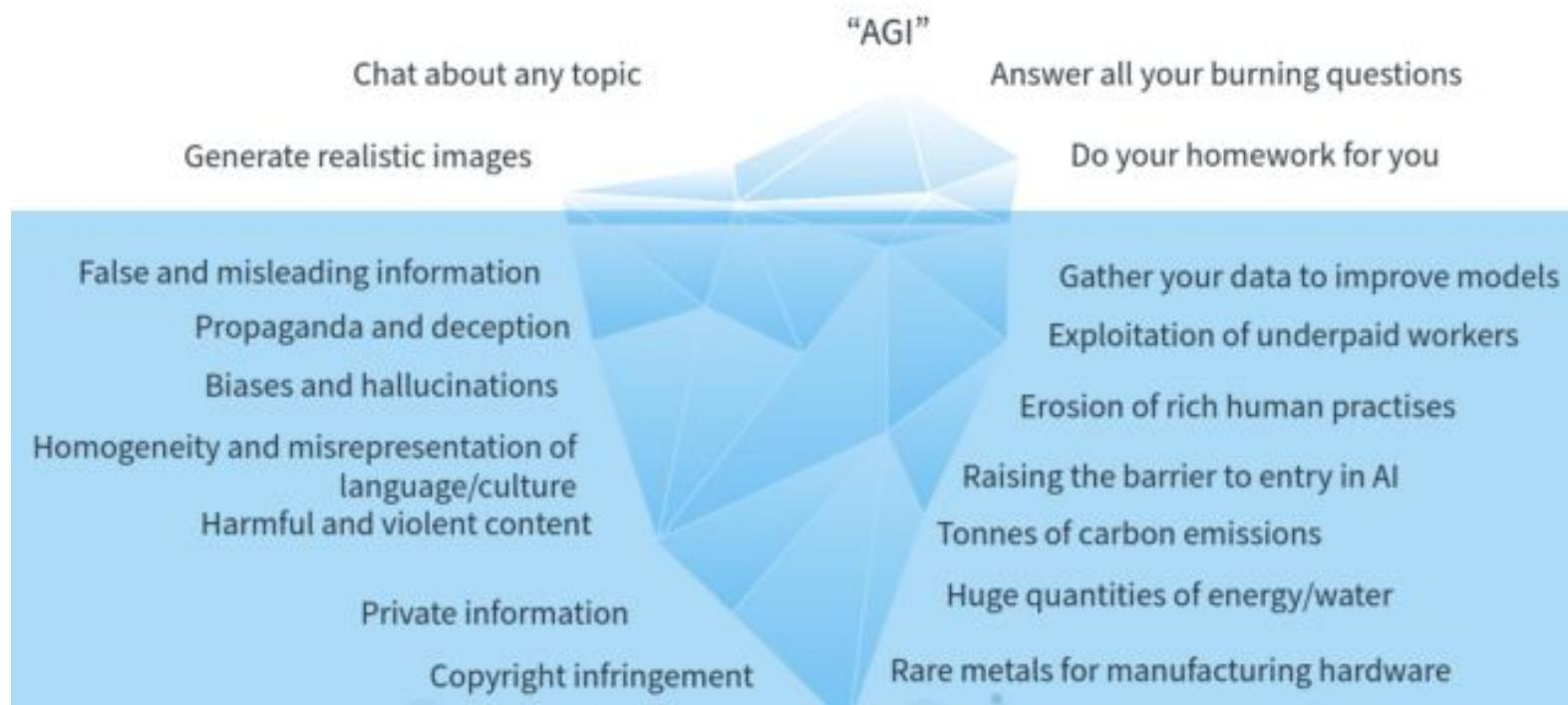
Source: <https://anjanasusarla.substack.com/p/the-second-coming-or-slouching-towards>

Current Large Language Models landscape

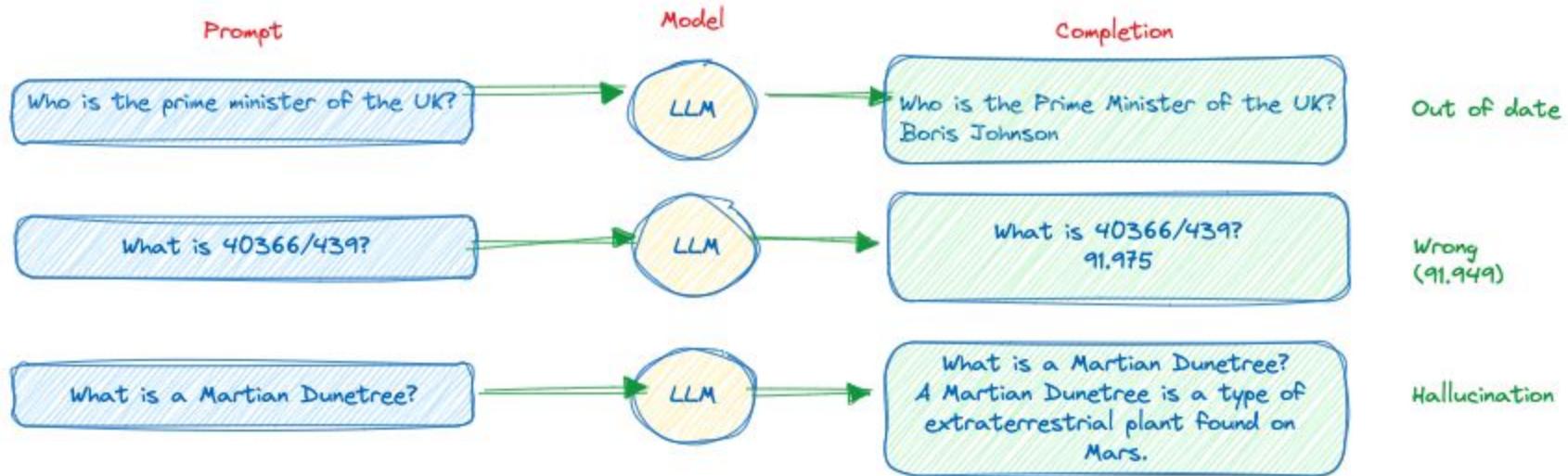


LifeArchitect.ai/models

Generative AI challenges



Challenges of Adopting Generative AI in Enterprises



Challenges of adopting Generative AI in enterprises cont.

We discussed:

- Hallucinations
- Staleness / Revision

Others:

- Attribution
- Customization
- Cost
- Performance
- Security

Solutions:

- Couple with an external memory (i.e. Vector Store)

External Memory: addressing LLM challenges

- Hallucinations are significantly reduced.
- Data in external memory can be swapped out on demand.
- Addresses shortcomings like Staleness, Revision and Customization
- Attribution or Data Grounding
- Security & Privacy
- Cost
- Performance
- Overcomes token limit

Vector Store

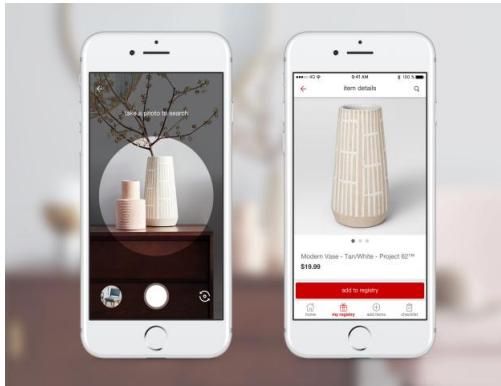
Use cases

Large Language Models



Semantic search, RAG, Q&A,
Document Retrieval, Chatbots,
LLM Caching, Chat memory

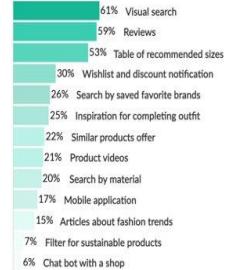
Visual Search



Find similar products through
image data

Recommenders

Features that customers desire

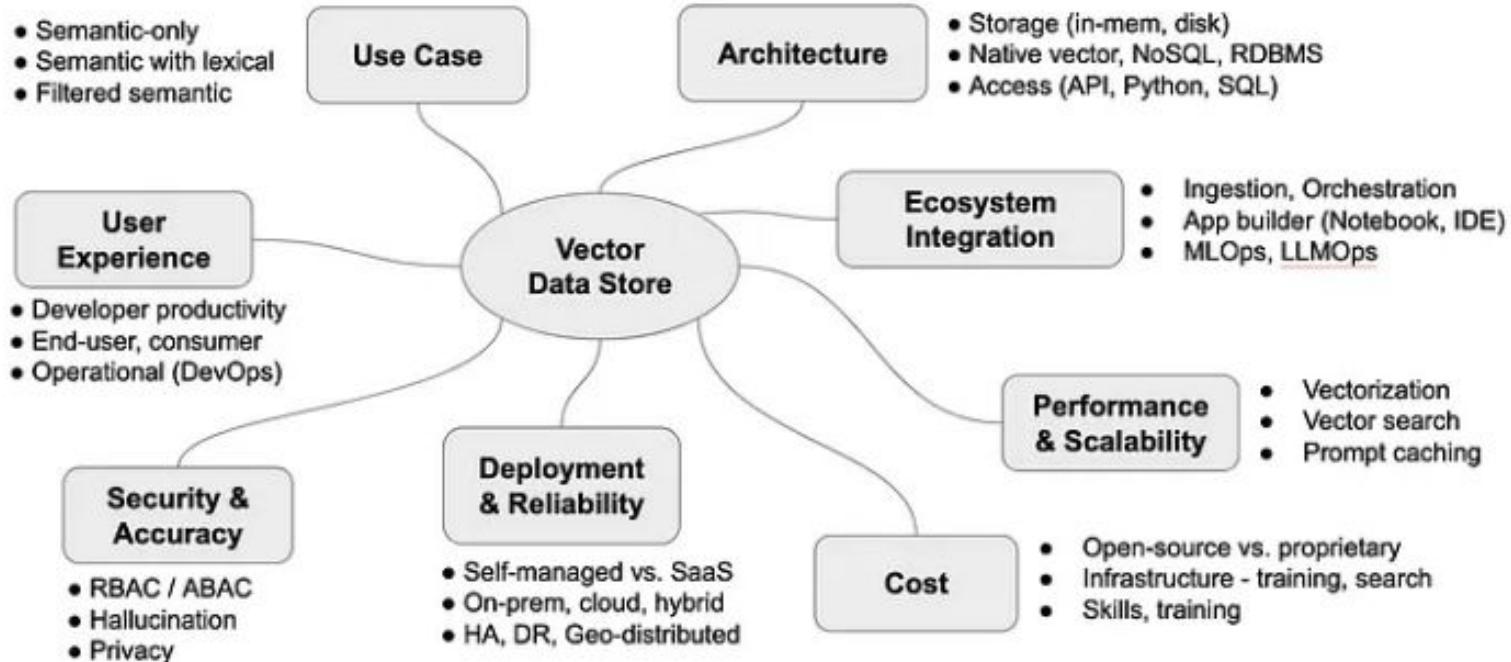


"Choose 3-5 features you consider important while shopping online for fashion!"



Recommend personalized
products, brands, properties,
offers, etc.

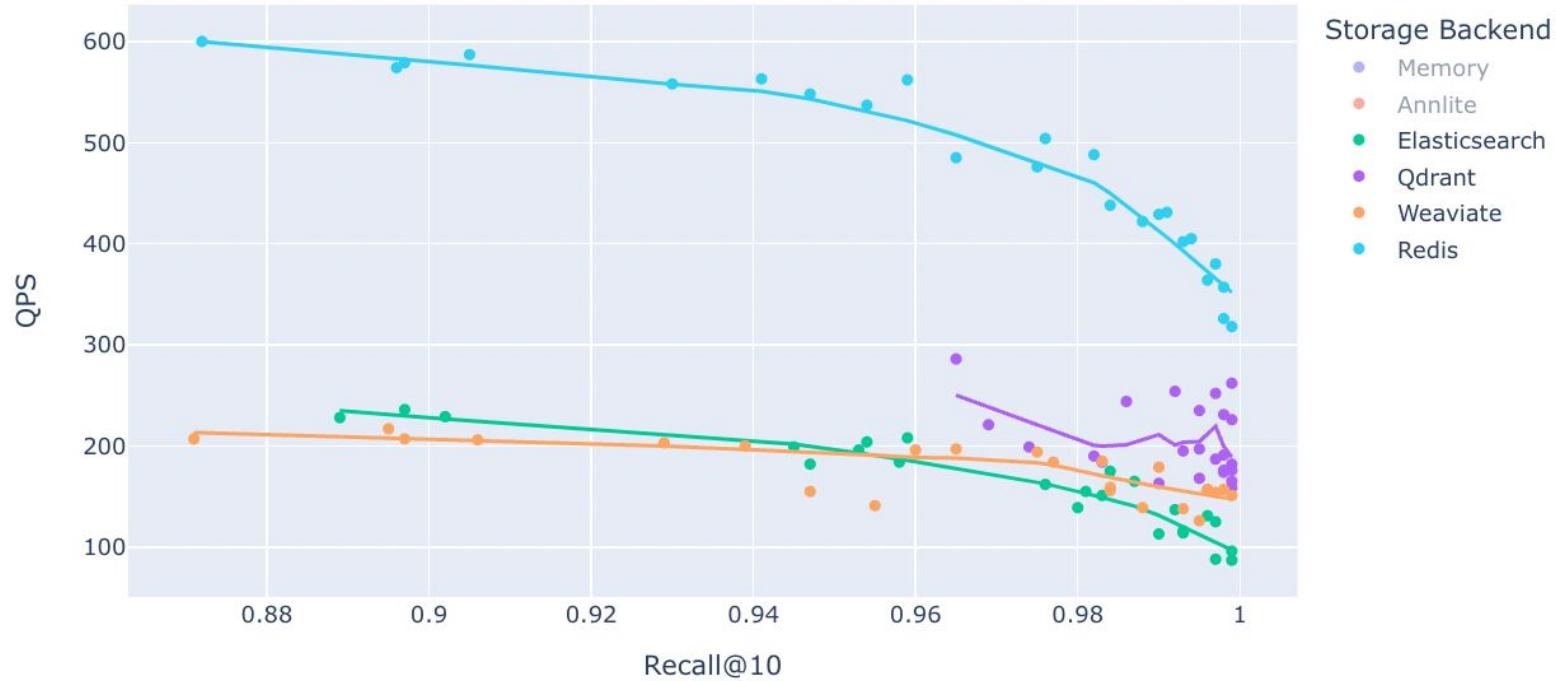
Vector Store: evaluation criteria



Source: Sanjeev Mohan (<https://sanjmo.medium.com/vector-data-store-evaluation-criteria-6d7677ef3b60>)

3rd Party Assessed Performance

QPS to Recall



<https://jina.ai/news/benchmark-vector-search-databases-with-one-million-data/>

-

Customer References



About Docugami

Document engineering empowers business users with impact on Day One, without any massive investment in machine learning, staff training, or IT development. Docugami developed a proprietary Business Document Foundation Model, an LLM for Generative AI applied to your documents.

Scan for
full story



Challenge

Docugami is a document engineering company that built its own Business Document Foundation Model.

The company analyzes and searches tens of thousands of documents (and millions of pages). That meant Docugami needed to scale easily.

Speed and reliability of its data layer was a crucial factor. But the company encountered bottlenecks with its Apache Spark document processing pipeline.



Solution

Redis Enterprise was employed as a vector database to handle embeddings more efficiently, which improved consistency and accuracy.

Docugami also uses Redis Enterprise to store Apache Spark checkpoint data, which means the company can use a single data platform.

It provides its enterprise customers with valuable data from their extensive set of business documents-and generates new content.



Results

Redis Enterprise makes it easy to store, search, and update vector embeddings at scale, improving the user experience by ensuring that Docugami's foundation model receives the most timely, relevant and up-to-date context.

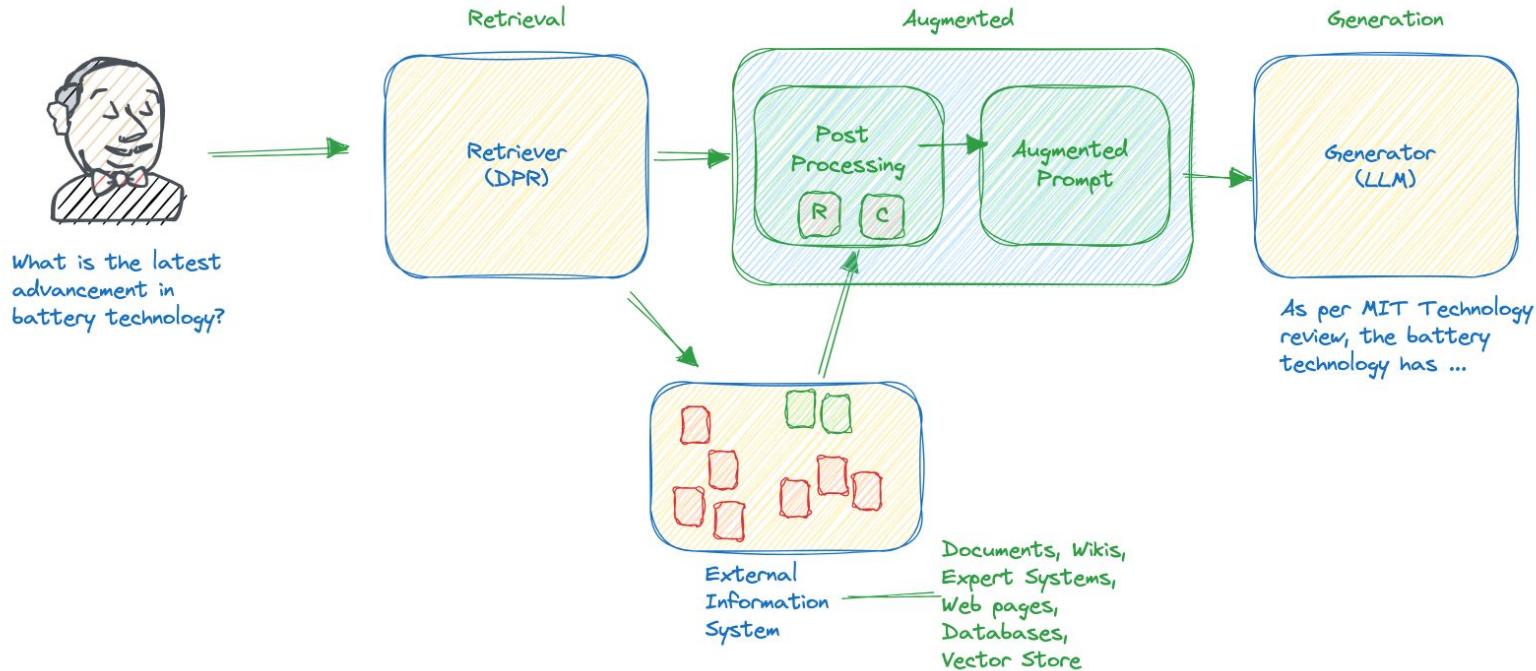
Redis also mitigated the bottlenecks in Docugami's Apache Spark processing pipeline, which was stymied due to high I/O.

“With Redis Enterprise as our vector database and our backing data store for Apache Spark, we've achieved superior latency, throughput, and embedding search result accuracy. Redis is helping us turn dense, lengthy documents into valuable data as a Document XML Knowledge Graph for our customers at massive scale.”

Jean Paoli
CEO and co-founder of
Docugami

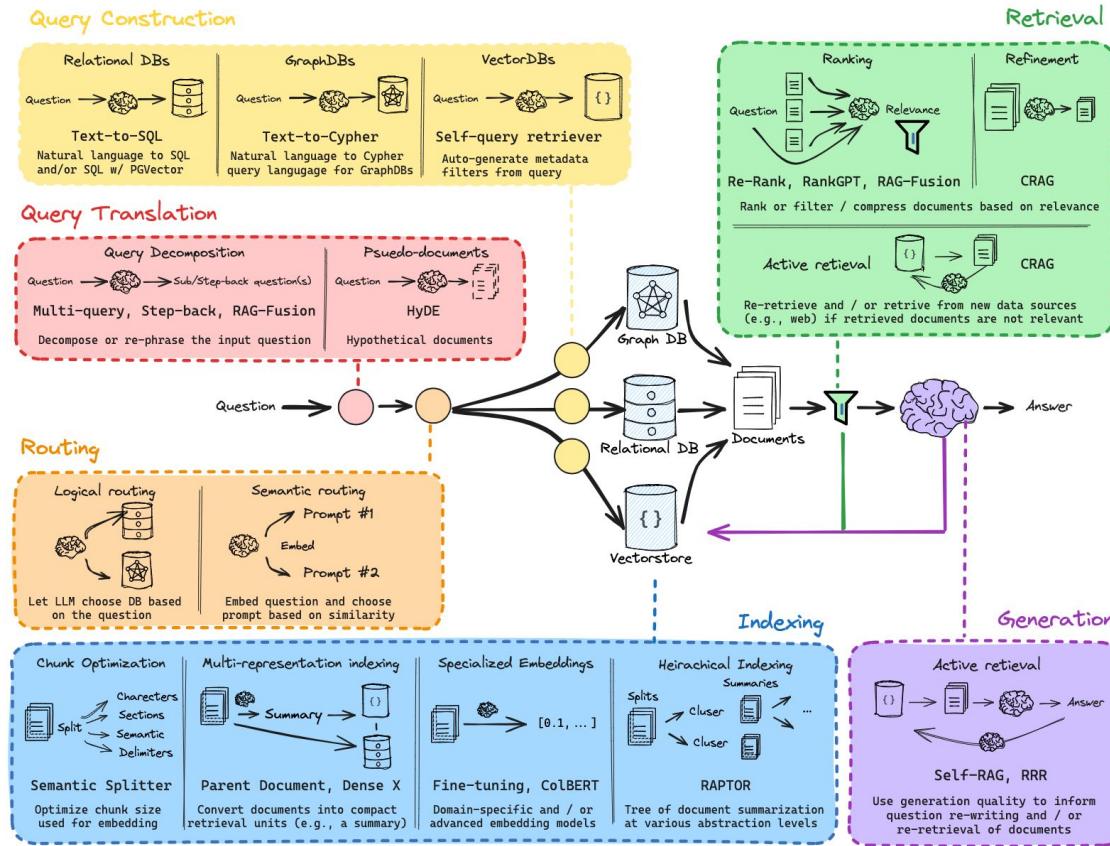
Architecture

Retrieval Augmented Generation framework

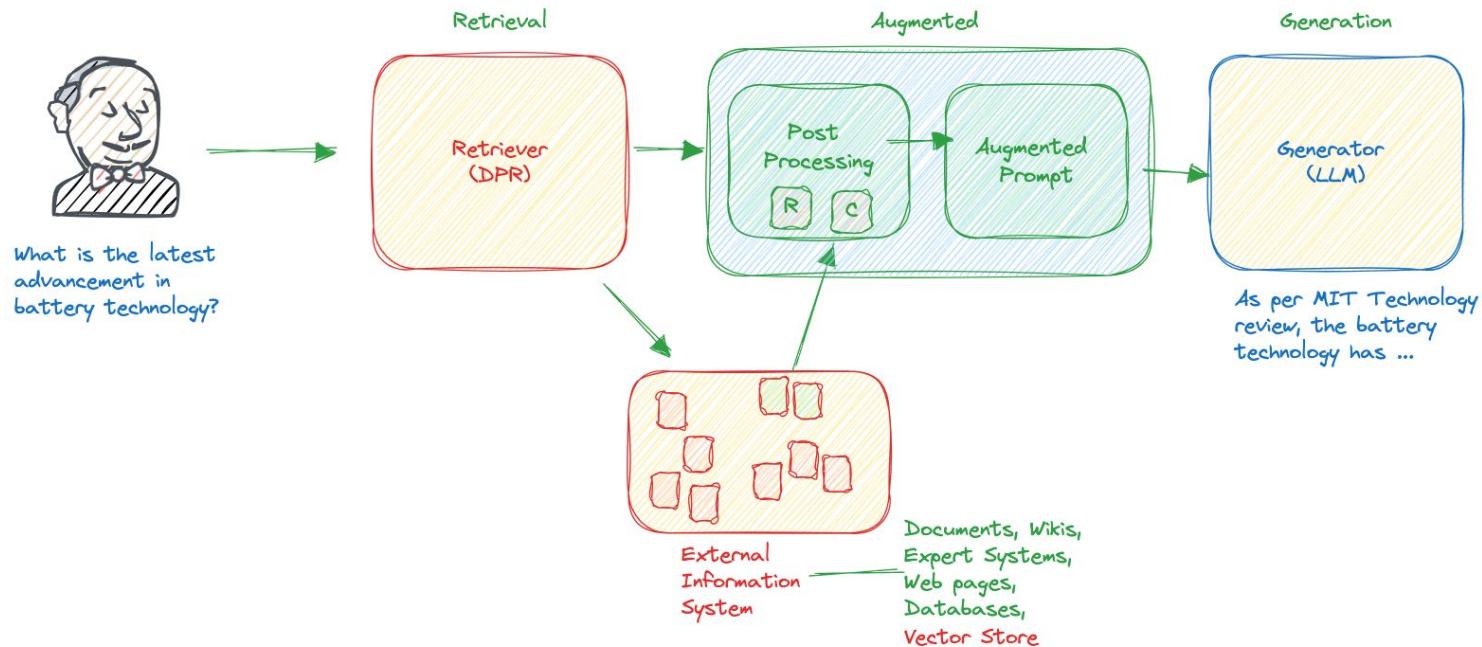


Source: <https://newsletter.artofscience.com/p/beyond-quesswork-the-rise-of-retrieval>

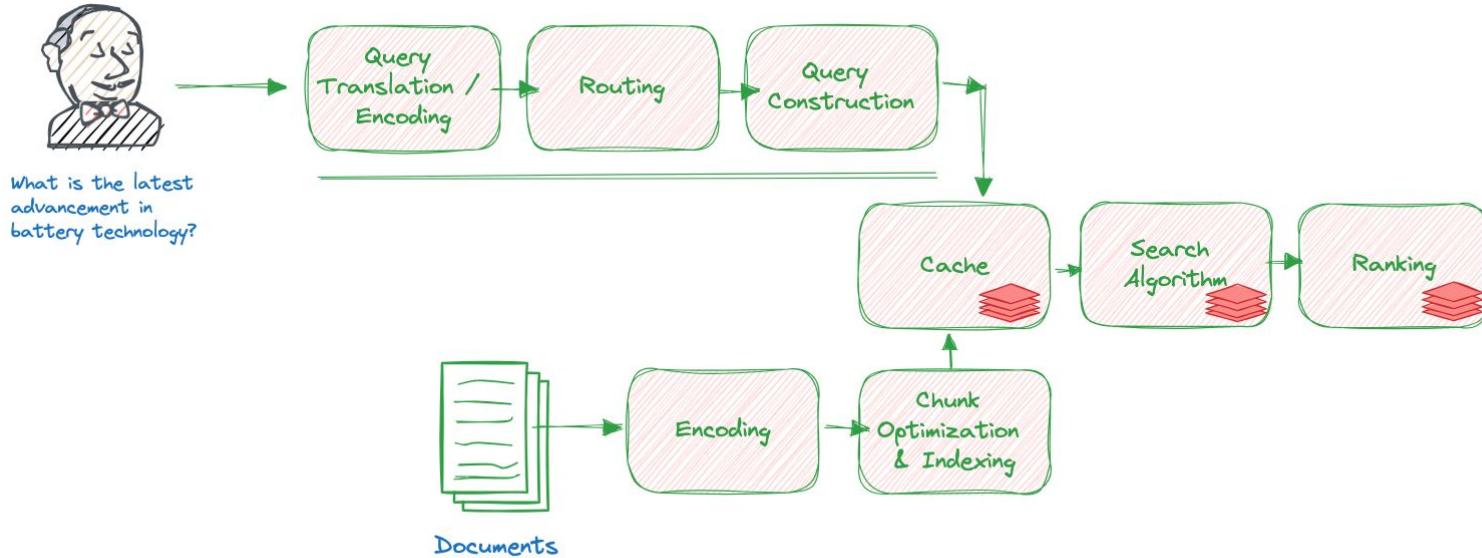
The RAG landscape gets complex quickly...



Today's workshop will touch on



Here are some sub-components our workshop



For the workshop you need following resources:



Redis

Welcome to the joint workshop hosted by AWS and Redis. To get started with your sandbox setup, please follow the instructions below:

1. Visit the registration page at <https://events.oneblink.ai>.

2. Enter the event code **161590**.

3. Verify your email address by checking for a validation code and follow the provided instructions to confirm.

After completing these steps, you'll gain access to the Sandbox environment, which includes AWS services and Redis. The AWS Bedrock service is pre-configured for your convenience.

Should you encounter any issues with sandbox, please don't hesitate to contact our support team at support@oneblink.ai for assistance.

Get started with workshop:

On AWS Console

1. Select AWS Sagemaker
2. On left bar, scroll and look for Notebook
3. Select Notebook instances
4. Select AWSBedrockRedisVSSSageMakerNotebook
5. Select Open JupyterLab



Lab 1:



- Open [redis-py-01.ipynb](#)
- Select Kernel: 'conda_python3'
- Learning Objectives: Redis and RediSearch
 - Text embedding
 - Indexing
 - Search & Ranking

Lab 2:



- Open [langchain-02.ipynb](#)
- Select Kernel: 'conda_python3'
- Learning Objectives:
 - Dataset Preparation
 - RAG with Redis Vector Store and LangChain

Questions?

- 10-15 min

Break

- Antony Prasad Thevaraj

Amazon Bedrock Presentation

- 10-15 min

Break

Hands-on workshop with Redis Vector Store + Amazon Bedrock

- Thank you for joining us today!

Interested in a customer workshop for your team or company? Please fill out the survey providing a few more details, all respondents will receive a \$25 gift card once the survey closed at 5:00pm ET 4/12!



Github Resource

- Dedicated training materials for building semantic search and RAG applications with Redis in Python.
- Runnable on Jupyter Notebook or Google Collab.
- Client options:
 - Redis Py
 - RedisVL
 - LangChain

<https://github.com/Redislabs-Solution-Architects/redis-aws-bedrock-workshop/>

The screenshot shows a GitHub repository page for 'redis-aws-bedrock-workshop'. The repository has 1 branch and 0 tags. The main file list includes 'main', 'resources', '.gitignore', 'LICENSE', 'README.md', 'langchain-03.ipynb', 'redis-py-01.ipynb', 'redisvl-02.ipynb', and 'requirements.txt'. The 'About' section states: 'This is a workshop designed for you to learn about Redis and Amazon Bedrock.' It includes links for 'Readme', 'MIT license', 'Activity', 'Custom properties', '0 stars', '6 watching', '0 forks', and a 'Report repository' button. The 'Releases' section says 'No releases published' and has a 'Create a new release' link. The 'Packages' section says 'No packages published' and has a 'Publish your first package' link. The 'Languages' section shows 'Jupyter Notebook 100.0%'.

Popular Redis Client Libraries for AI Use Cases

Redis Python



Base client libraries offer the most configurability and performance, at a cost of more complexity.

<https://github.com/redis/redis-py>

RedisVL



Purpose-built frameworks wrap base clients for specific use cases. Focused on ease of development.

<https://github.com/RedisVentures/redisvl>

LangChain, LlamaIndex, et al



AI orchestration frameworks that natively integrate with LLMs, Storage, APIs, Vector DB, and more.

<https://python.langchain.com/docs/integrations/providers/redis>

A few more hands-on examples

- Document search application: <https://github.com/RedisVentures/redis-arXiv-search>
- Chatbot application:
<https://github.com/RedisVentures/gcp-redis-llm-stack/tree/main/examples/chat-your-pdf>
- OpenAI Cookbook:
https://github.com/openai/openai-cookbook/tree/main/examples/vector_databases/redis

- Results

model = 'amazon.titan-tg1-large'

```
[108]: query = "What was Nike's revenue last year compared to this year??"
res=qa(query)
res['result']

[108]: '$40 million less'

[109]: query = "How many products does Nike offer? What is the industry that Nike is part of?"
res=qa(query)
res['result']

[109]: Nike offers consumer products and the industry that Nike is part of is the athletic footwear, apparel and equipment industry'

[110]: query = "Is Nike an ethical company?"
res=qa(query)
res['result']

[110]: No

[111]: query = "How many employees work at Nike???"
res=qa(query)
res['result']

[111]: 44,538'
```

model = 'amazon.titan-text-lite-v1'

```
[74]: query = "What was Nike's revenue last year compared to this year??"
res=qa(query)
res['result']
```

```
[74]: ' 42,293 44,538'
```

```
[75]: query = "How many products does Nike offer? What is the industry that Nike is part of??"
```

```
res=qa(query)
res['result']
```

```
[75]: ' Nike offers 42 products. Nike is part of the athletic footwear, apparel and equipment industry.'
```

```
[76]: query = "Is Nike an ethical company?"
res=qa(query)
res['result']
```

```
[76]: ' Yes'
```

```
[77]: query = "How many employees work at Nike???"
```

```
res=qa(query)
res['result']
```

```
[77]: ' 44,538'
```



model = 'OPENAI_API'

```
[92]: query = "What was Nike's revenue last year compared to this year??"
res=qa(query)
res['result']

[92]: " As of May 31, 2023, Nike's revenue was $42,293 million, which is a 10% increase from the previous year's revenue of $44,538 million (as of May 31, 2022)."

[93]: query = "How many products does Nike offer? What is the industry that Nike is part.of?"
res=qa(query)
res['result']

[93]: ' Nike offers a wide range of products including athletic footwear, apparel, and equipment. It is part of the athletic footwear, apparel, and equipment industry, which is highly competitive both in the United States and worldwide.'

[94]: query = "Is Nike an ethical company?"
res=qa(query)
res['result']

[94]: " I'm sorry, I cannot determine if a company is ethical. This information is not provided in the context given."

[95]: query = "How many employees work at Nike???"
```

res=qa(query)

res['result']

```
[95]: ' As of May 31, 2023, Nike had approximately 51,217 employees.'
```

- Open discussion

Q&A

- Next steps

Thank you for joining us today

Please fill out the survey to
receive a Redis swag bundle:

- Next steps

Redis resources

- Start building your Apps today for [FREE with our 14-day trial](#)
- [Redis Cloud on AWS Marketplace](#)
- Checkout our Developer Hub at [developer.redis.com](#)
- Try Fraud Detection: [Git Repo](#)
- [Bedrock + Redis Getting Started Resources](#)
- Hands-on-labs: [Redis VSS](#)
- Want to connect? Email Redis at aws@redis.com

AWS resources

- [AWS Reference Architecture: Real-time Fraud Detection](#)
- [Migrate Redis to Redis Enterprise Cloud in AWS - Prescriptive Guide](#)



→ **Start your free trial today!**

thank you.

Redis