NERDERY®

Automate, Personalize, Analyze:
How to Train Large Language Models (LLM) with Your Own Data

Justin Richie | Vice President of Data and AI | May 2, 2024

VP of Data and AI

Meet
Justin Richie

Justin spearheads Nerdery's data and AI team, expertly crafting and scaling solutions that drive impactful customer outcomes. He is passionate about building cross-functional teams — with a focus on becoming more data-driven.

https://www.linkedin.com/in/justinrichie/

Agenda

01 Welcome

02 The Modern LLM Stack

03 Running LLMs on your Database

04 Expanding Functionality

05 Q&A

Upcoming webinar:
Intro to AI Agents and how to build them

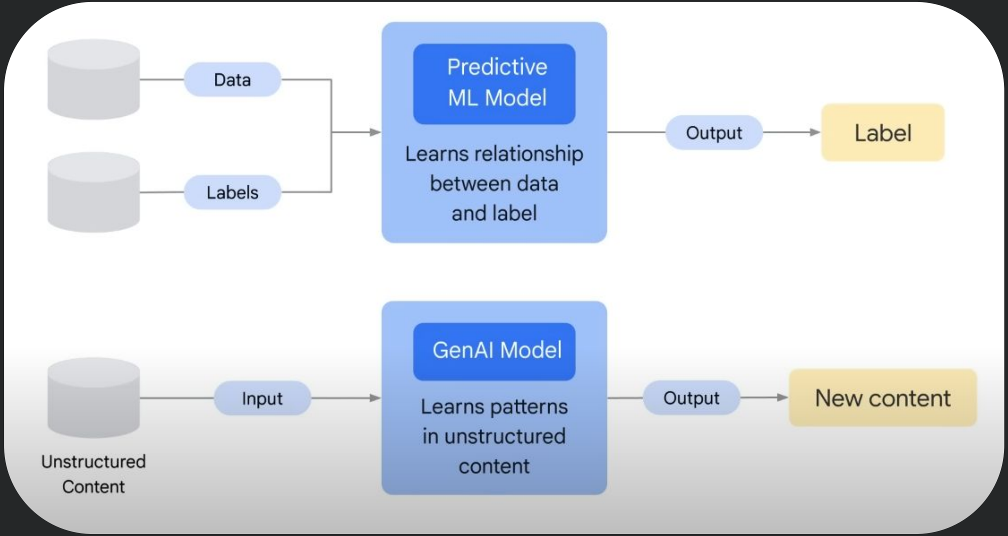
Tuesday, June 18, 2024

Leveraging RAG and LangChain for building AI Apps

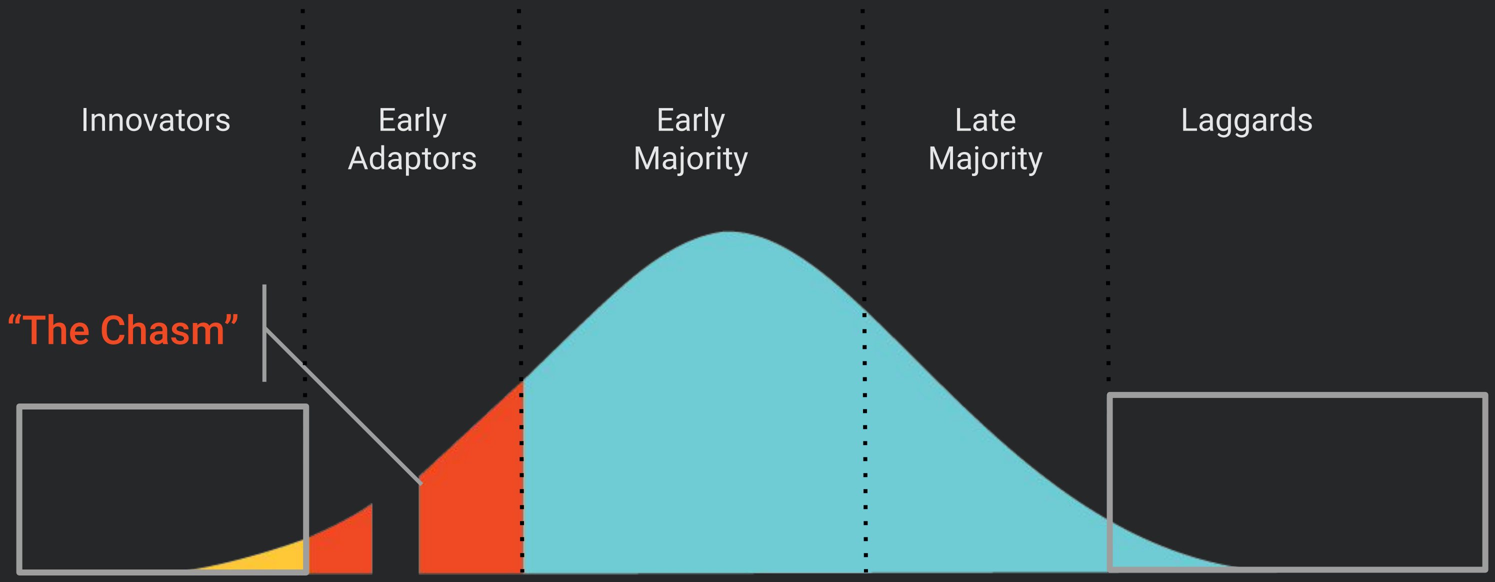
Generative AI: What is it?

GenAI, or Generative Artificial Intelligence, is a type of AI technology focused on creating new, original content or solutions. It analyzes extensive data and learns patterns to generate text, images, ideas, or predictions.

Useful in various sectors like marketing, product development, and decision-making, GenAI enhances creativity, efficiency, and personalization in business processes.



Companies adopting LLMs have a lot of noise



Technology Adoption Lifecycle

The easiest part is the **technology**

- Engage Legal, PR, Compliance and other business teams early in architecture process
- Most users aren't skilled in prompt engineering and need assistance with guided prompts and training
- Engage subject matter experts in the data to develop your chunking strategy
- Manual testing doesn't cut it. Add automated evaluations into testing

The Modern LLM Stack

Where do we start with LLMs on your own data?

Option	Pros	Cons	Analogy
Prompt an LLM	• Lower AI Skills • RAG can augment data freshness • Lowest Cost	• Lower precision on customization ability • Potential security concerns if no governance	Rent an apartment
Tune an LLM	• Specialized for domain specifics • Faster than building your own LLM	• Inaccurate tuning can degrade • Big line between fine tuning and adjusting weights	Build a custom home with a builder
Train Custom LLM	• Specialized to use case • High accuracy to domain specific need	• Deep AI Skills • Training is in the millions of dollars • Need GPUs or TPUs	Build a custom home on a remote island

Difference between tunings in the models

- Fine-tuning versus RAG is typically two paths companies go to customize LLM on their own data
- Tune an LLM for specific needs for your business domain
- Fine tuning is retraining the model every weight in the LLM
- Really big training job and very expensive. Also you have to host it once it's done
- Parameter-Efficient Tuning Methods (PETM) is alternative to train custom data without duplicating the model

Prompting a model can take a variety of forms

Stage 1

Prompt Design with existing LLM

Prompts involve instructions and context passed to a language model to achieve a desired task. This phase assumes no change in the LLM and is seen as “out of the box”

Stage 2

Data Integration and Customization

Prepare your business-specific data and integrate it with Gemini Pro using GCP tools like BigQuery. Analyze the LLM’s responses to refine data inputs, aiming to enhance relevance and accuracy specific to your business context.

Stage 3

Deployment of Retrieval-Augmented Generation (RAG)

Implement a retrieval-augmented generation system by setting up a document store and linking it to Gemini Pro. This allows the LLM to access and utilize stored business information, improving the quality and applicability of its outputs.

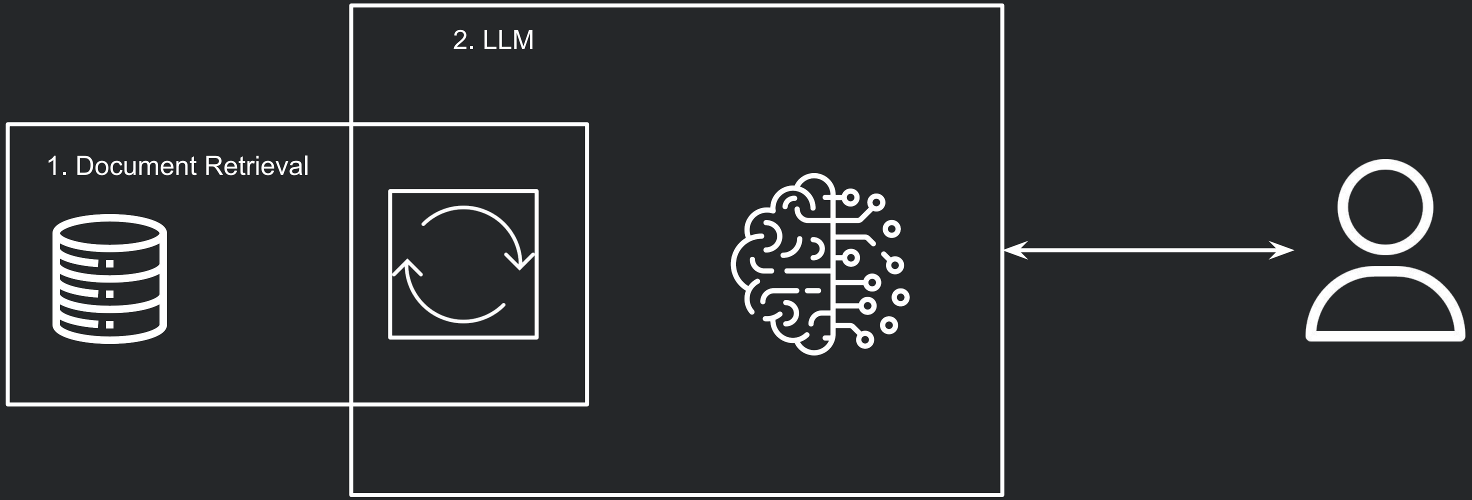
Stage 4

Integration with LangChain for Advanced Applications

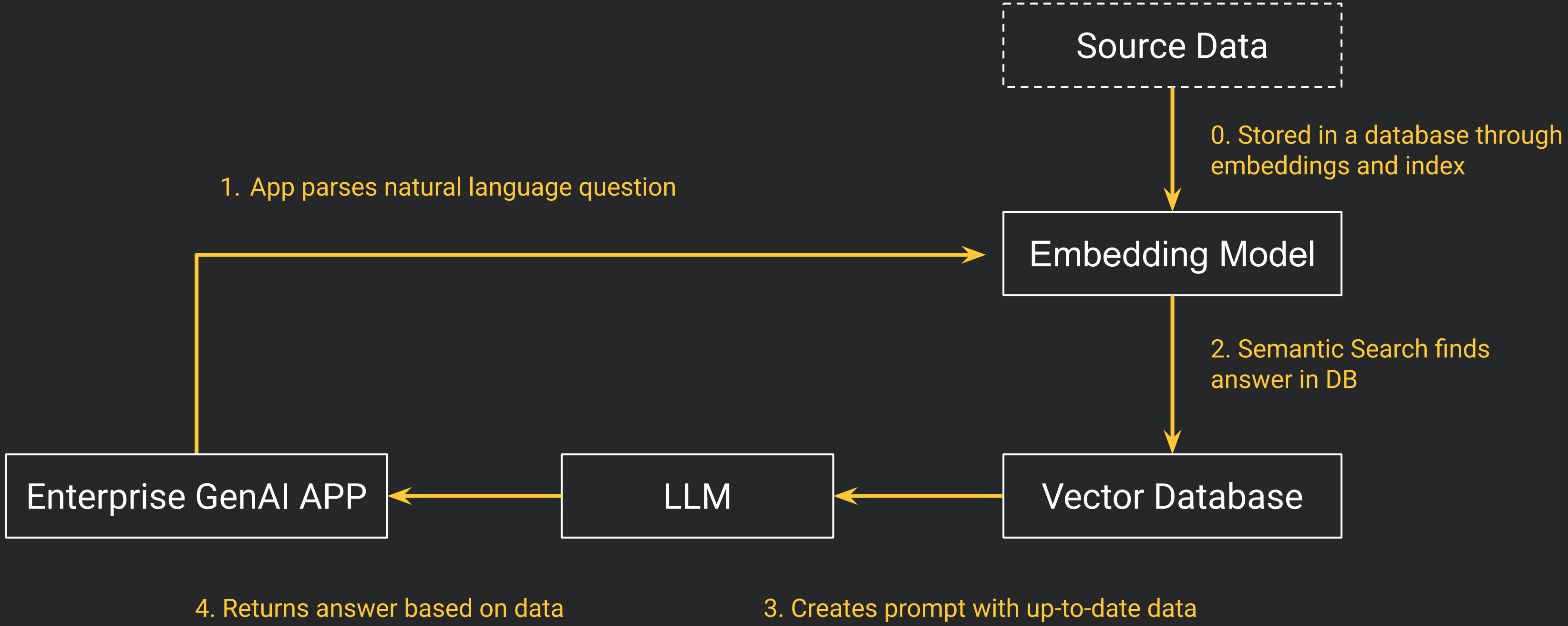
Integrate LangChain to construct complex applications, combining it with Gemini Pro and RAG. Develop applications such as automated customer support or interactive decision aids, continually updating the system with new data to enhance performance.

RAG 101

RAG, or Retriever-Augmented Generation, combines information retrieval with generative models to enhance AI's ability to provide relevant responses

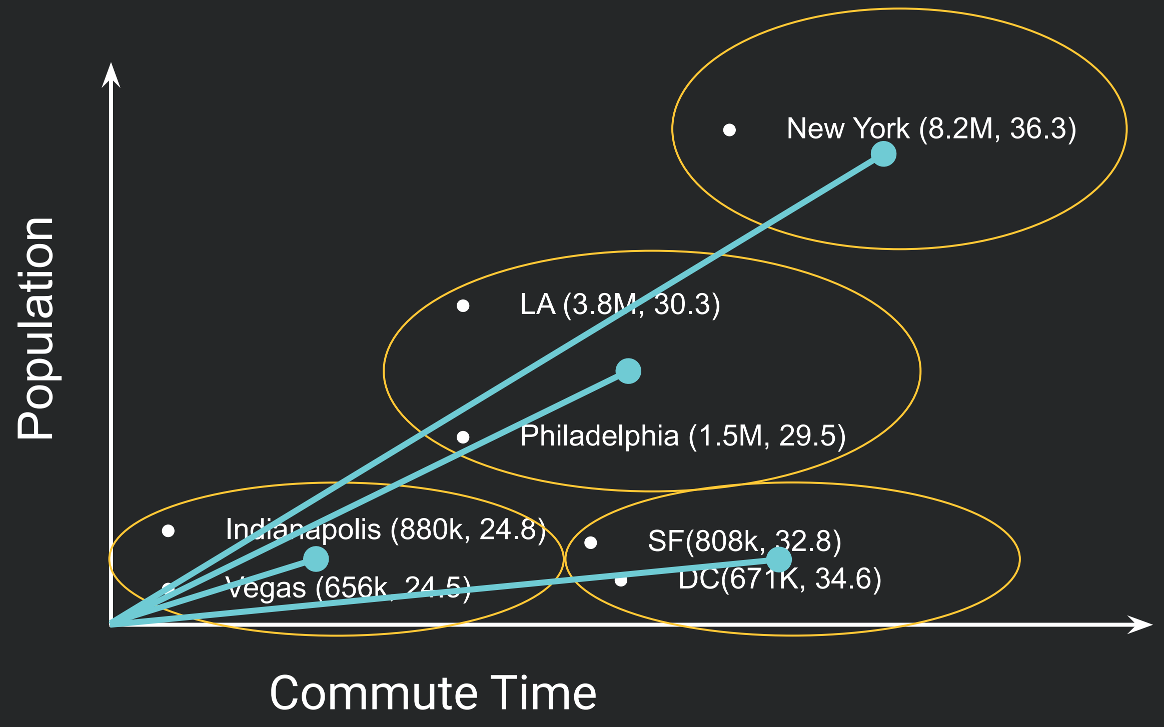


RAG architecture



A word on embeddings

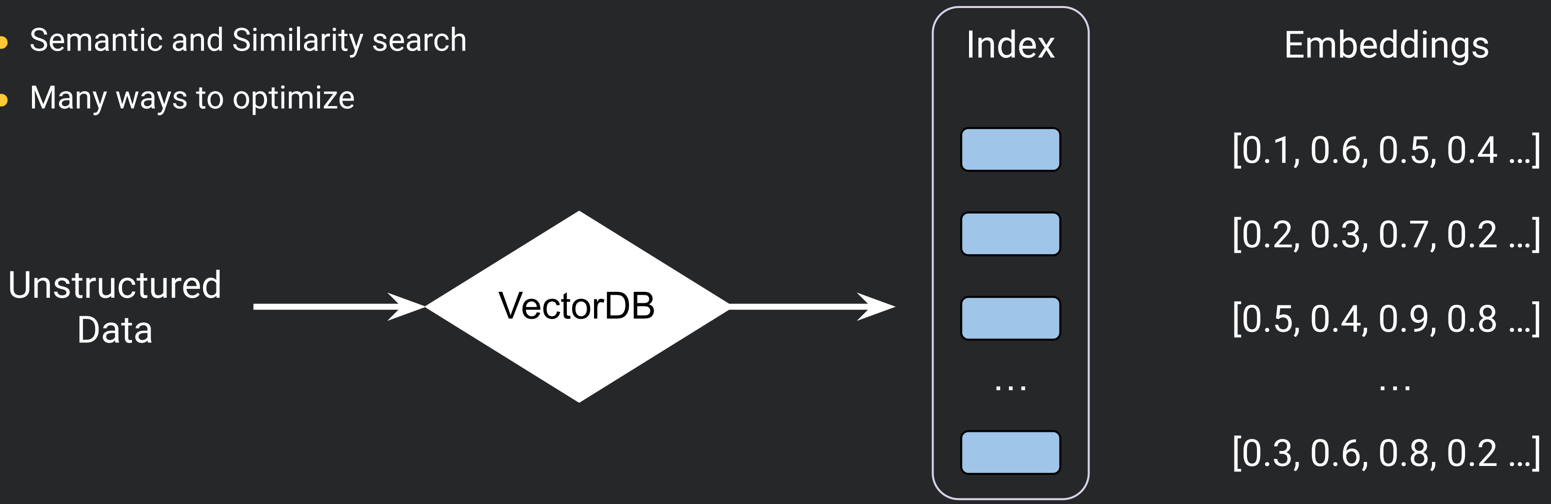
Index is a data structure to add in search process, embeddings are the distance in related items (ANN)



A note on vector databases

Index is a data structure to add in search process, embeddings are the distance in related items (ANN)

- Benefits include: long-term memory for LLMs
- Semantic and Similarity search
- Many ways to optimize



5 reasons your project will fail

- Not thinking about cost ramifications of LLM API calls
- Hiring data scientists when you really need cloud, ML or software engineers
- Not understand devops tools like distributed applications (i.e. Ray) and containers/ gke/cloud run
- Not creating ROI metrics from project inception
- Domain specific knowledge limitations leading to poor results

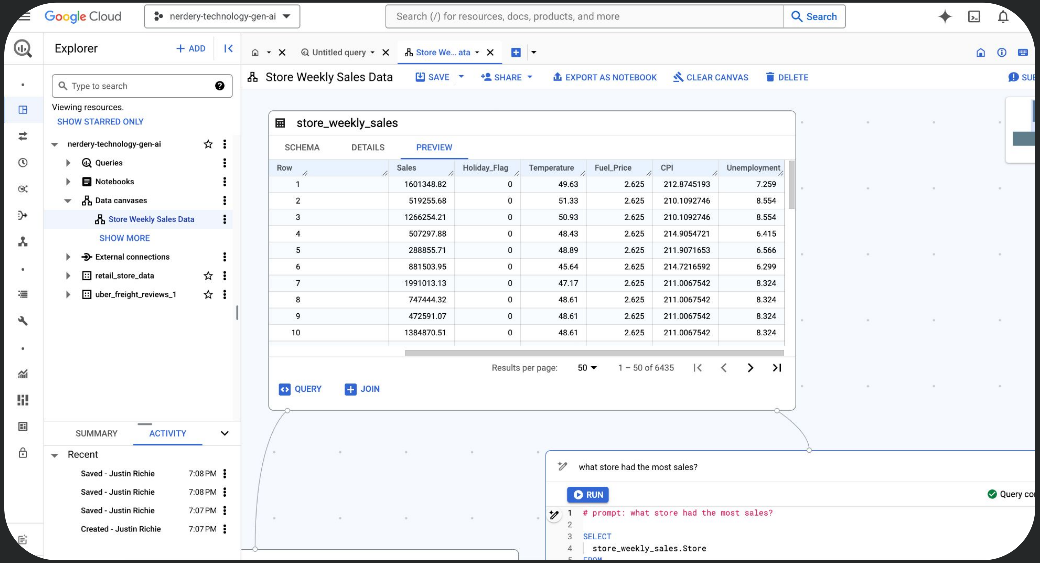
Glossary of terms

- **Gen AI Token:** Tokens used in generative AI models for processing text
- **Chunking:** Segmenting text or data into smaller, manageable pieces
- **Vector DB Index:** Indexing mechanism for fast retrieval in a vector database
- **Vector DB Embedding:** Storing data as vectors in a database for similarity searches
- **RAG (Retrieval-Augmented Generation):** AI model enhancing responses by retrieving relevant data
- **Langchain:** A tool for building language model applications using LLMs

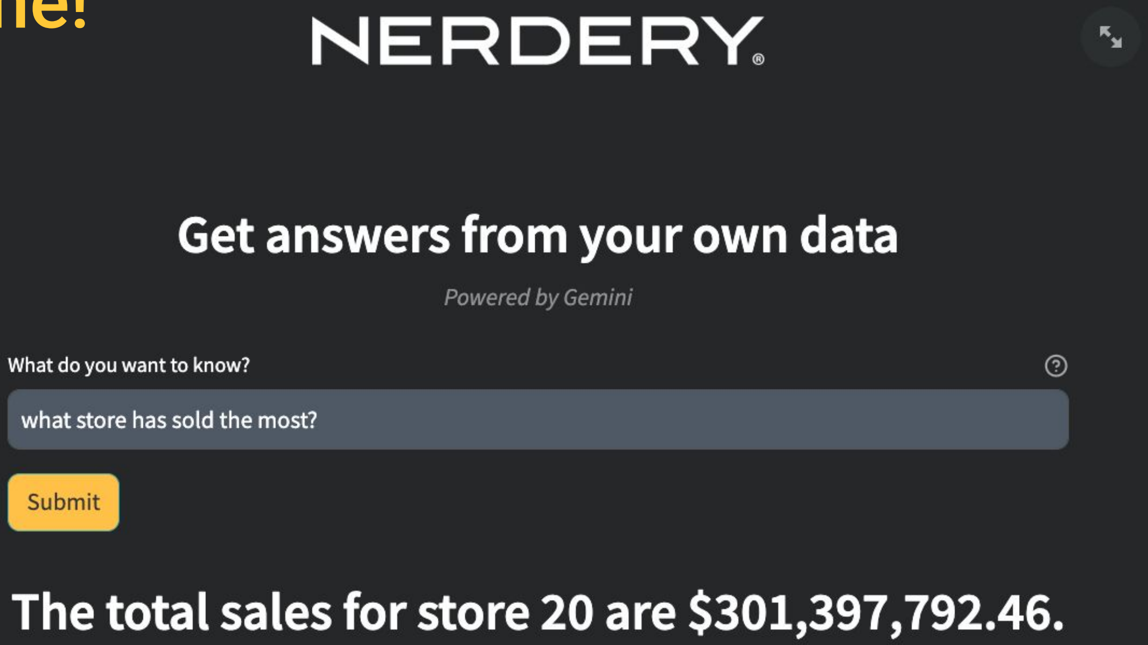
Running LLMs on your database

What is coming from Next (Data Canvas)

- Google Cloud's new BigQuery Data Canvas product is a drag-and-drop tool to query data with natural language
- Data Canvas complex data analytics through natural language commands
- Automates SQL queries to export directly to sheets and Looker
- Product is limited but will gain more availability in the future



Demo time!


NERDERY
Get answers from your own data
Powered by Gemini
What do you want to know?
what store has sold the most?
Submit
The total sales for store 20 are $301,397,792.46.

Expanding Functionality

Realtime data protection

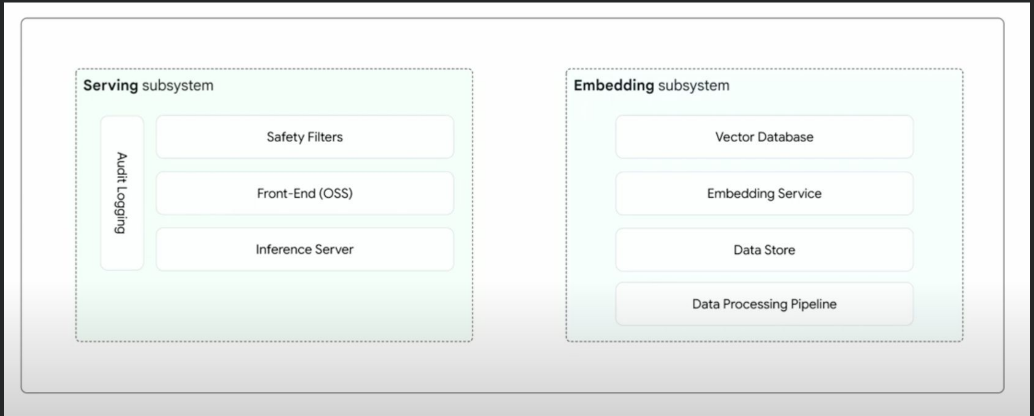
- GCP offers Sensitive Data Protection (SDP) to identify, block, and mask over 150 different data elements
- PII and financially sensitive data
- Can also filter on potentially harmful data
- De-identification, masking, tokenization, and bucketing

Hi my name is Justin, my SSN is 408-659-5555 and my DOB is 1/1/1984

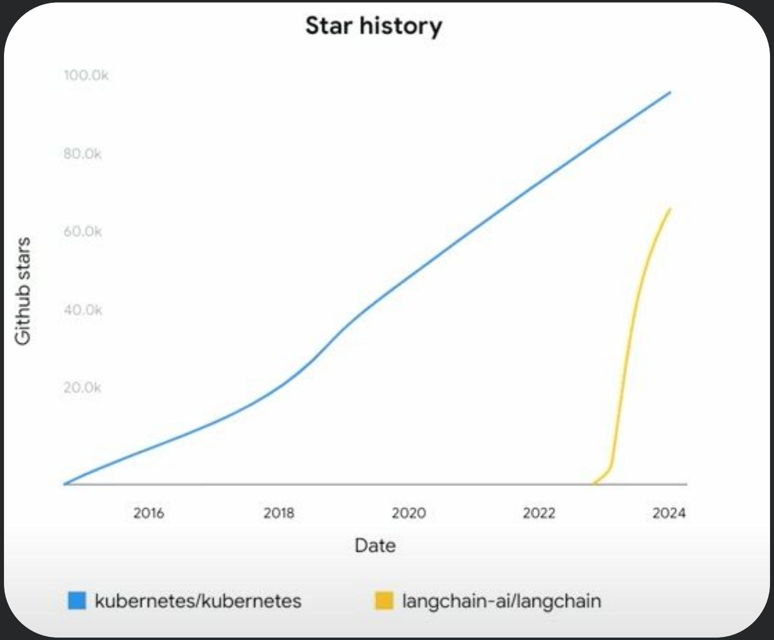
Hi my name is [PERSON_NAME], my SSN is [SSN] and my DOB is [DATE_OF_BIRTH]

How to protect your own data

- IAM will be vital to implement to secure access at user and product level
- API Gateways can help secure APIs
- Data encryption is essential with SSL/TLS
- GCP Cloud Audit Logs will be vital to track activity



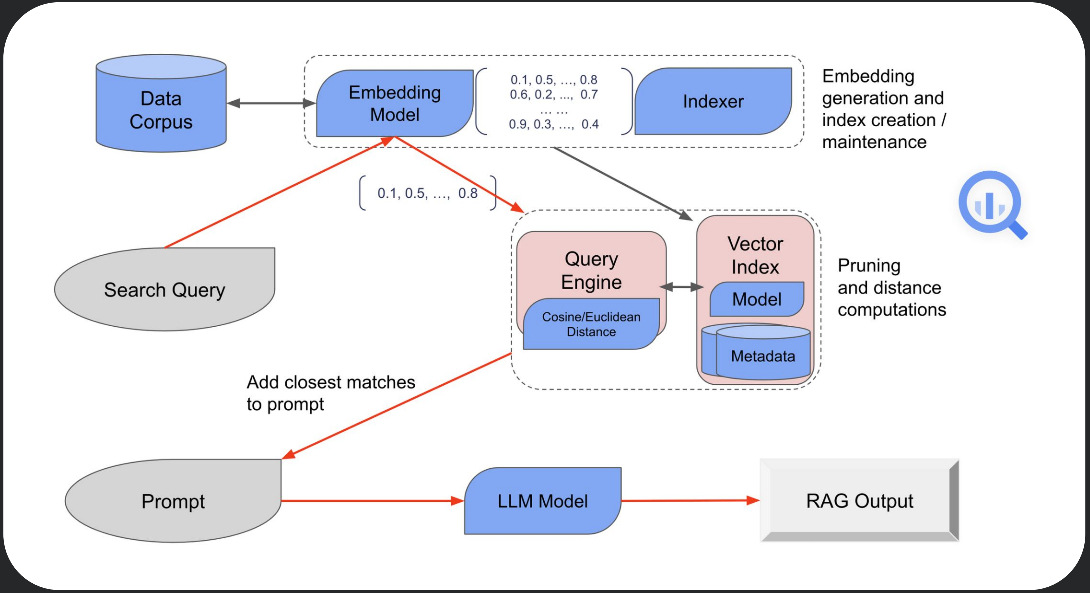
Where does LangChain fit into this?



- Connecting LLMs to your company's private sources of data and APIs
- Offering a complete set of interoperable and interchangeable building blocks
- Customizability and control with a durable runtime baked in
- Open sourced and powered by a community of 2K+ contributors

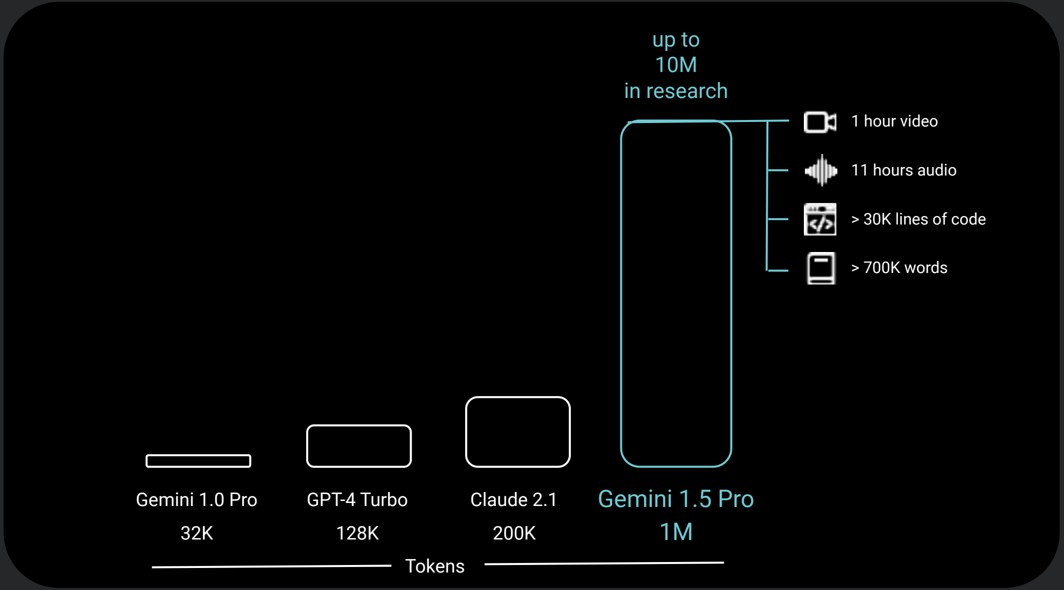
BigQuery vector support in public beta

- It offers a simple and intuitive CREATE VECTOR INDEX and VECTOR_SEARCH syntax that is similar to BigQuery's familiar text search functionality
- It works with BigQuery's embedding generation capabilities
- BigQuery vector indexes are automatically updated
- The LangChain implementation simplifies Python-based integrations with other open-source and third-party frameworks



Where will RAG go?

- TL DR; It's not going anywhere anytime soon
- Gemini 1.5 Pro's 1M token context window is a game-changer, enabling the model to process vast amounts of information in a single go
- Gemini 1.5 Pro leverages the Mixture of Experts (MoE) architecture, enhancing its ability to process and respond to complex queries efficiently



Recap + Summary

- It’s easy to get started with your own data in LLMs, RAG makes it accessible
- GCP tools make it easier than ever to get started
- Vector databases are really the way to get the best performance for RAG architectures
- It’s better to have something to 80% of desired functionality than 100% and not used

Q&A

NERDERY®

Nerdery Data + AI

info@nerdery.com
877.664.6373
www.nerdery.com

Nerdery HQ

7700 France Ave, Suite 285
Edina, MN 55435

NERDERY®

Thank You