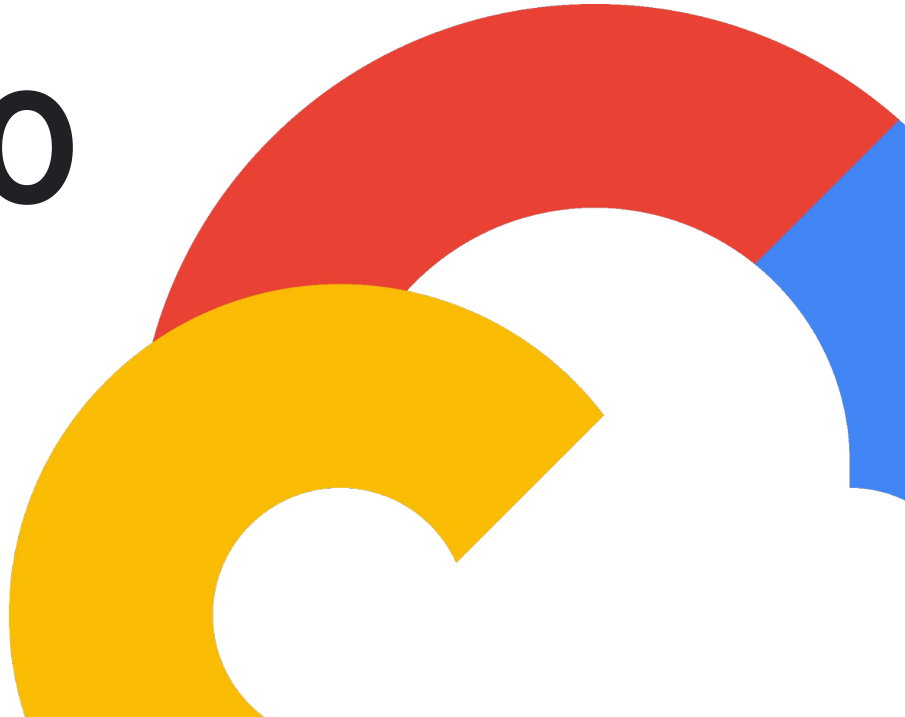
Desarrolla un chatbot en GCP desde 0

Usando BigQuery, Vertex AI y Generative AI



cívica

Google Cloud

Contents

Agenda

9:45h

Introducción & Setup

Conceptos teóricos, presentación del problema y diseño de la solución.

10:30h

1. Indexación

Embedding, creación de índice, llenado de índice y despliegue

11:00h

2. Approximate Nearest Neighbors

Búsqueda de artículos más similares

11:45h

COFFEE BREAK

12:00h

3. Langchain

Combinación de embeddings, llamada a BBDD Vectorial, BigQuery y prompts.

12:45h

4. User Interface

Crea tu propio chatbot a partir de una plantilla en Streamlit.

ADVANCED LEVEL

13:30h

5. Multimodal

Añadimos soporte a fotografías en nuestro chatbot con ayuda de Gemini

Speakers & Support Team

Javier Mayorgas

Head of AI
Cívica Software

Isaac Blázquez

Data & AI Sales Executive Iberia Retail&CPG

Fran Yáñez

Google Cloud Partner Engineer

Con Cívica, la tecnología trabaja para ti.

- Especialización en el desarrollo de soluciones tecnológicas que aportan valor diferencial a nuestros clientes, con especial foco en el mundo del dato.
- Amplio conocimiento y experiencia en Modern Data Stack y servicios Cloud,
- Partner oficial de Google Cloud desarrollando proyectos en grandes clientes.
- Crecimiento anual +30% en facturación y empleados.
- Equipo de alta capacitación tecnológica y baja rotación (+250 personas).
  Premiados como “Great Place to Work” en varias ediciones.

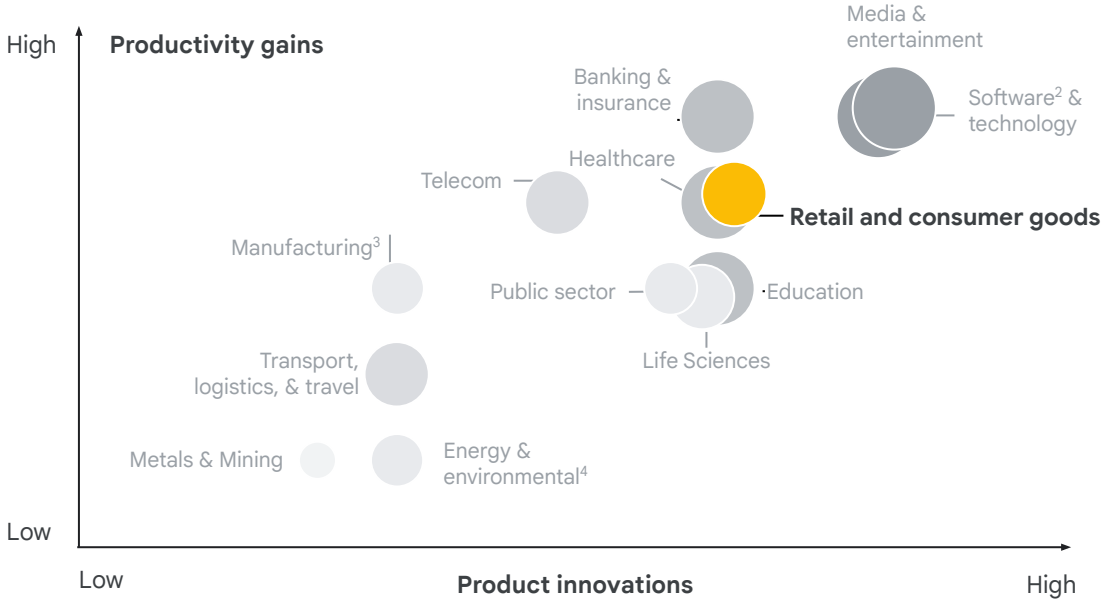
CÍVICA Google Cloud

Conceptos básicos

Introducción a BigQuery e IA Generativa

GenAI will have a transformative impact in retail

Impact of GenAI across industries[1]



1. For each industry, industry-specific software is included 2. Industry-agnostic software 3.Includes aerospace and defense, automotive and assembly, chemicals, semiconductors, basic materials 3. Includes auto retail. 4. Includes oil and gas, power.
Source: McKinsey Global Institute: "The economic potential of generative AI: The next productivity frontier"

28-48%

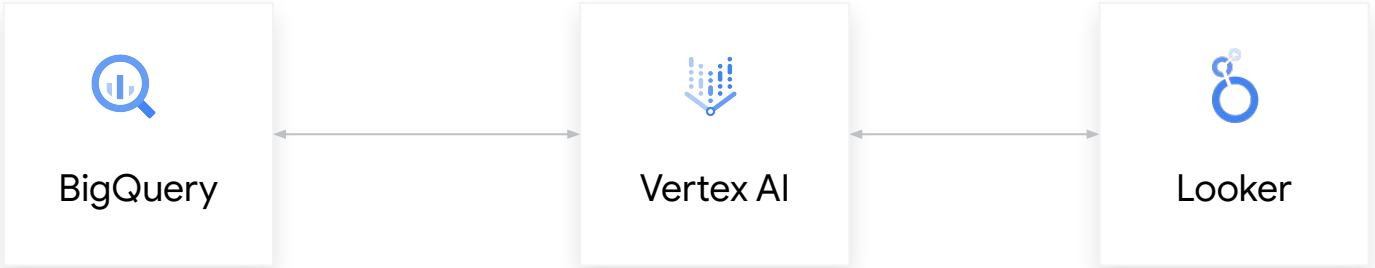
GenAI's impact as a proportion of EBIT in the retail and consumer goods industries

Remove the limits to innovate with data and AI

Start with an AI-ready data platform

Bring AI capabilities to your data

Create AI-driven data experiences



BigQuery

Unified platform from data to AI

60X

BigLake YoY growth to process multimodal and multicloud data

>1 Billion

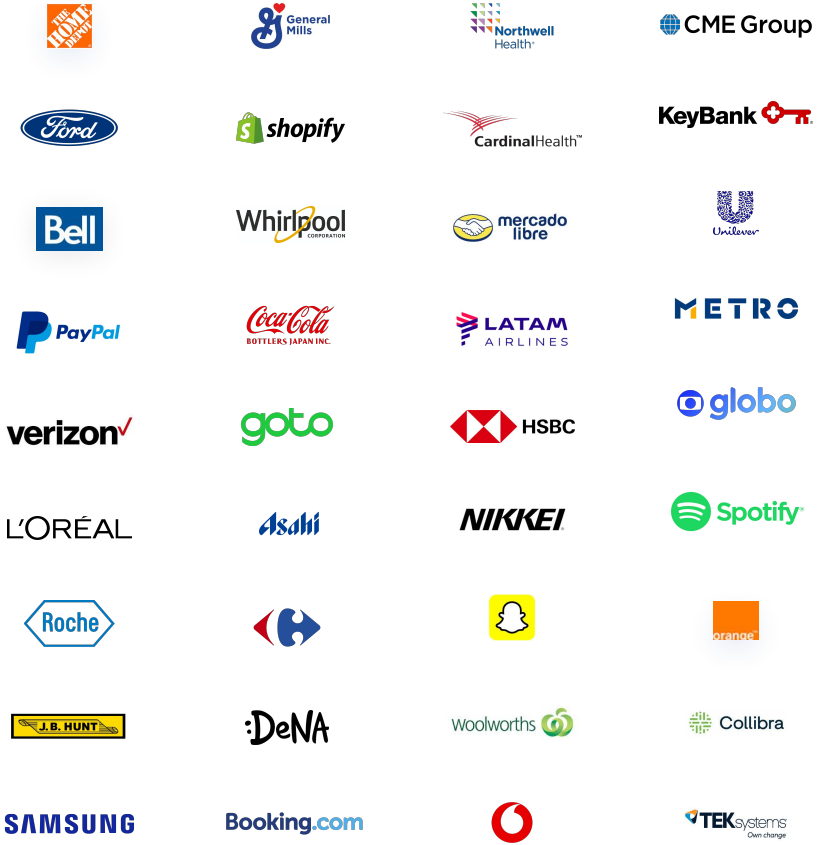
Daily queries across **exabytes** of data by **10s of 1,000s** of customers

250%

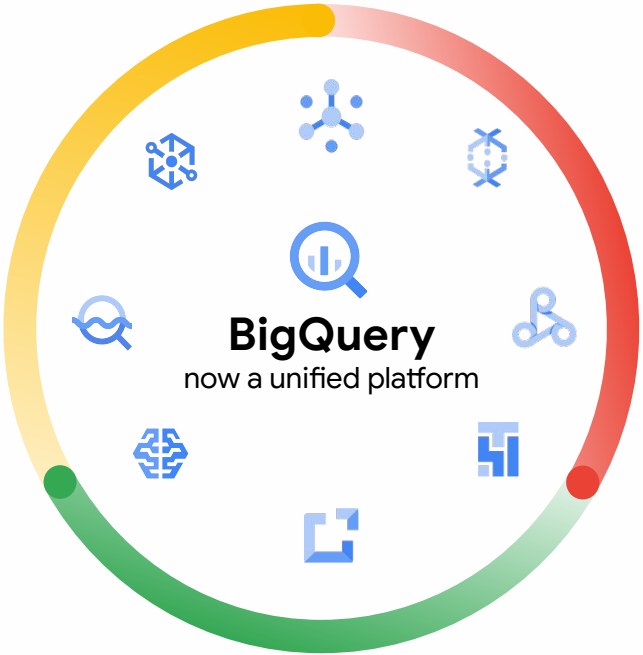
ML queries YoY; **hundreds of millions** of predictions & trainings

54%

Lower **TCO** compared to market alternatives



Simplifying and unifying data analytics



- Multi-engine
- Multi-storage
- Multi-format
- Multi-capability
- AI/ML/LLM integration

Managed AI lakehouse with BigLake

Bringing Google scale to open format data with price-performance & automatic management

GA

Supports open file formats

Query Apache Iceberg, Delta and Hudi formats with built-in fine-grained access control

GA

Performance acceleration

Continuous data optimization for large-scale ingestion of Apache Iceberg tables

Preview

BigLake managed tables

Fully managed experience on Apache Iceberg with support for DML and high throughput streaming

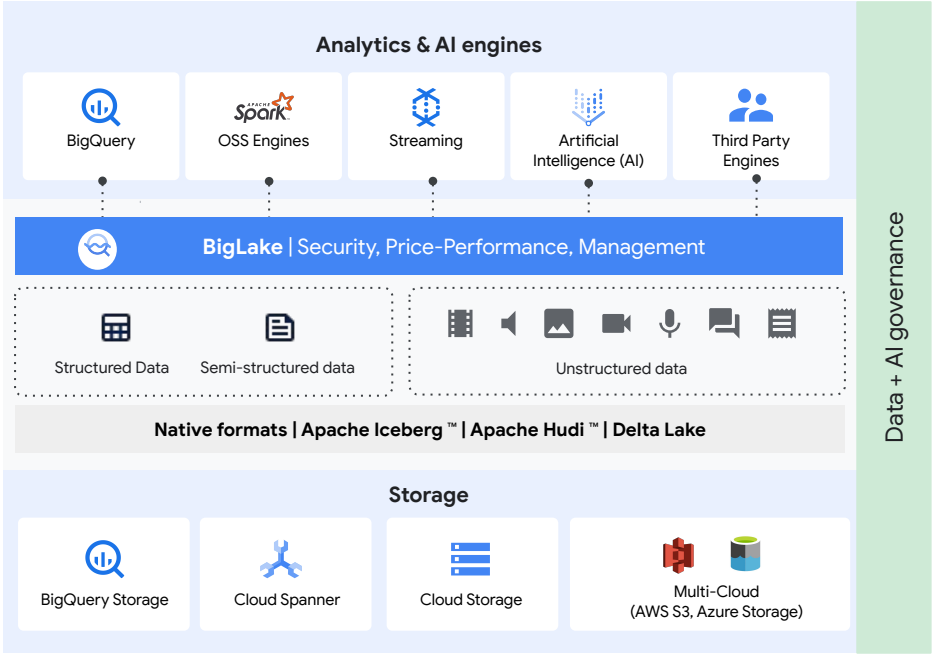
Preview

Native support for Delta More support for open formats and more choice for you

Preview

Spanner external datasets

Use Cloud Spanner schemas as external BigQuery datasets.



Multimodal data and AI with BigLake

Query structured & unstructured data with built-in security and governance across analytics & AI engines

Multimodal data

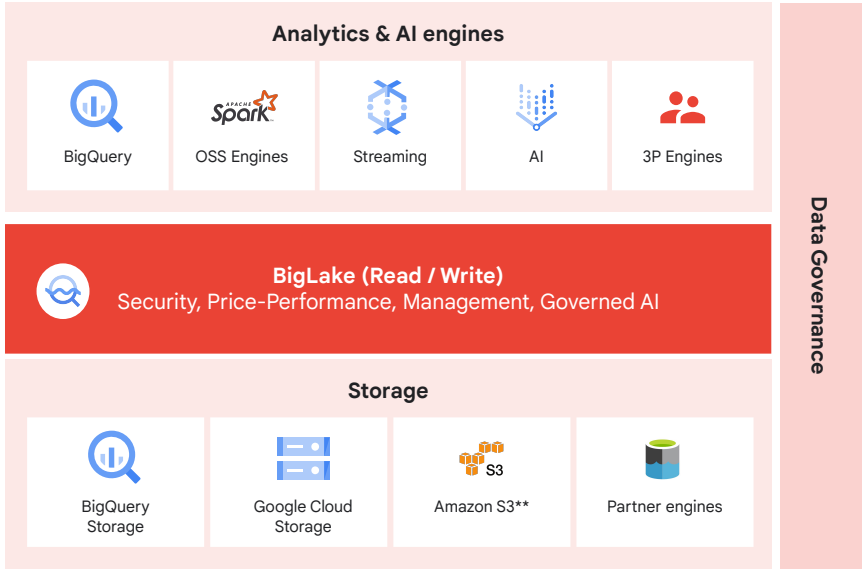
Build a single-copy architecture for all your structured & unstructured data across clouds & open file formats

Unified governance

Unified data management and governance across distributed data

Unified data and access control

Access all data from a single interface with centrally managed security



Query acceleration improvements

Improved price performance with no changes to workloads

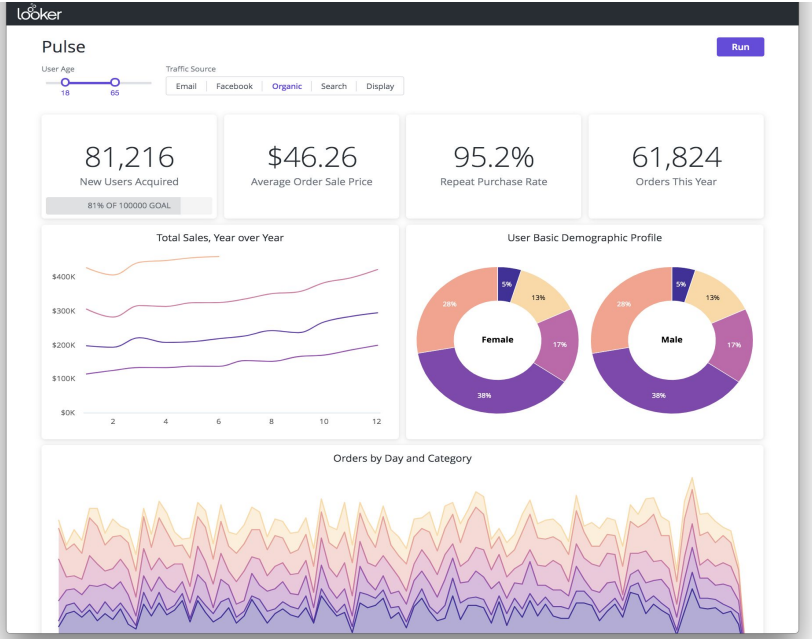
Preview Up to 5x faster query performance with automated, self-learning **history-based optimization**

Preview Up to 50% faster small query (<1 GB scanned) execution. High throughput (‘000s qps) for multi-user workloads.

GA Adaptive vectorized BI engine with **automatic smart caching**. Queries that can be satisfied by the cache will execute in sub-second.

GA **Transparent query acceleration** from any tool: BigQuery Studio, Looker, Looker Studio, Tableau, Power BI, etc.

Preview Improved experience for Looker Studio (job details, monitoring, cost tracking, query cache, information schema, materialized views)



Continuous real-time analytics in SQL

Preview

Unlocking intelligent data pipelines with continuous SQL processing

Continuous analytical processing

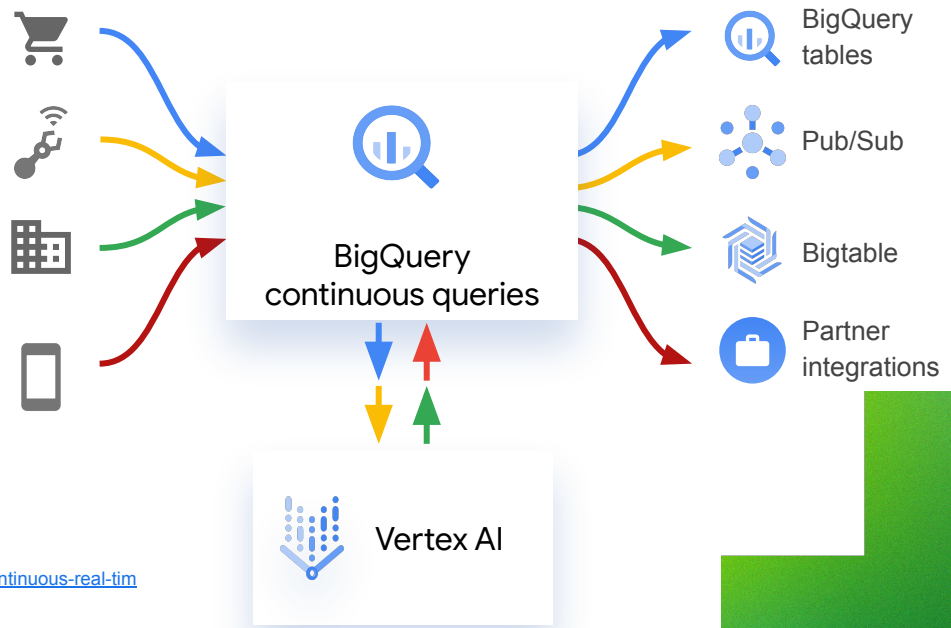
Perform unbounded serverless analytics over streams of incoming data using SQL

GenAI for real-time app development

Develop pipelines with AI+ML capabilities, such as real-time anomaly detection, sentiment analysis, and recommendations

Reverse ETL into operational systems

Programmatically transform and replicate data in real-time from BigQuery into Pub/Sub, Bigtable, or another BigQuery table



https://cloud.google.com/blog/products/data-analytics/google-clouds-innovations-for-continuous-real-time-intelligence?e=48754805

BigQuery ML for GenAI

Every data engineer can now be a ML engineer

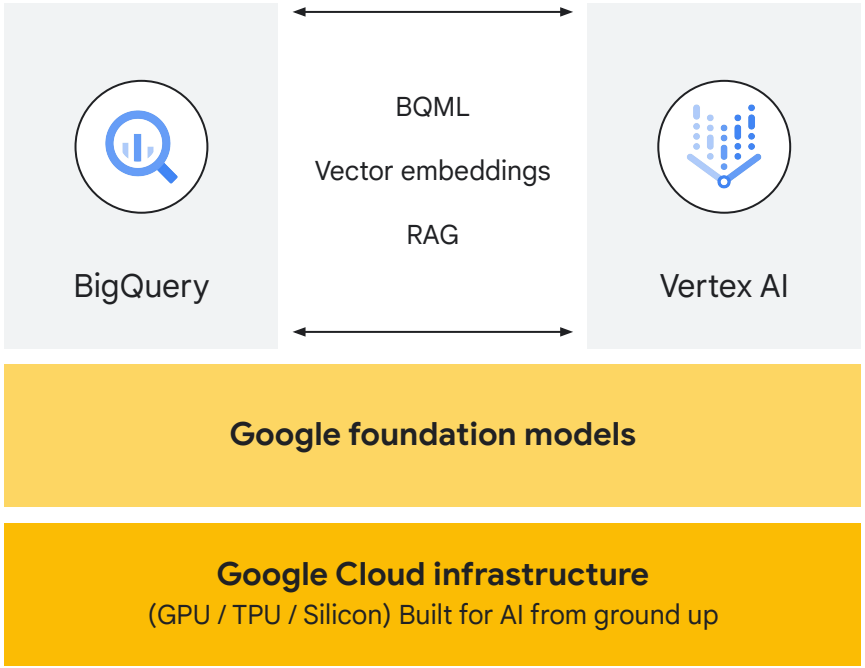
SQL LLM execution in BigQuery

- Summarization
- Sentiment Extraction
- Classification
- Data enrichment
- Entity Extraction
- Translation

Document and Speech AI in BigQuery

Secure and governed Doc & Speech AI over text and audio files directly in BigQuery

Unlock proprietary company data



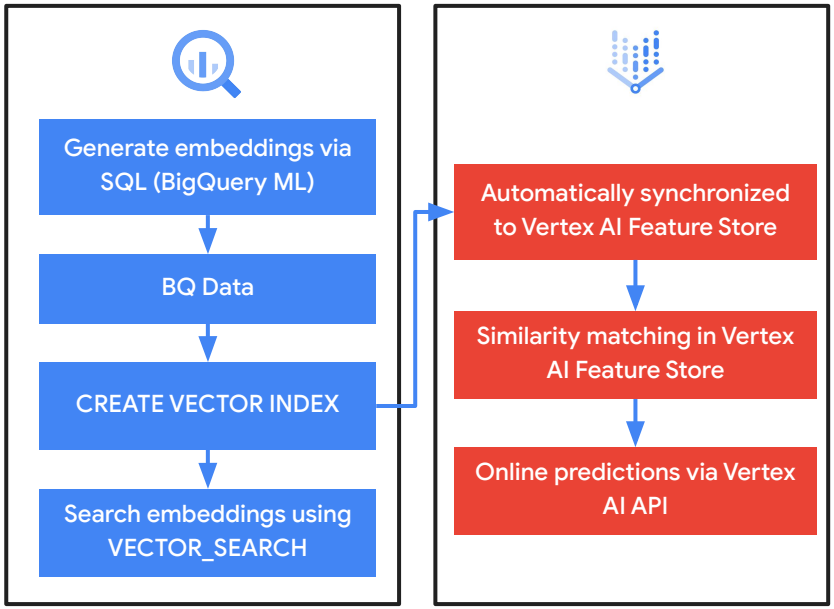
BigQuery vector embeddings and indexes

Preview

- **Collect and standardize on vectors** across multiple databases and Cloud Storage
- **Fully managed vector index** keeps embeddings in sync with Vertex AI for online prediction
- Support for LangChain & Vertex AI extensions

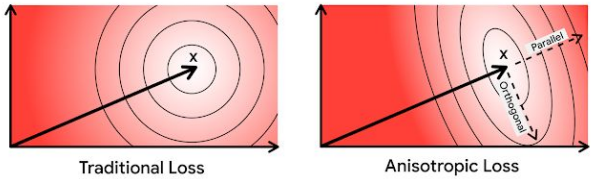
When do you need a vector index?

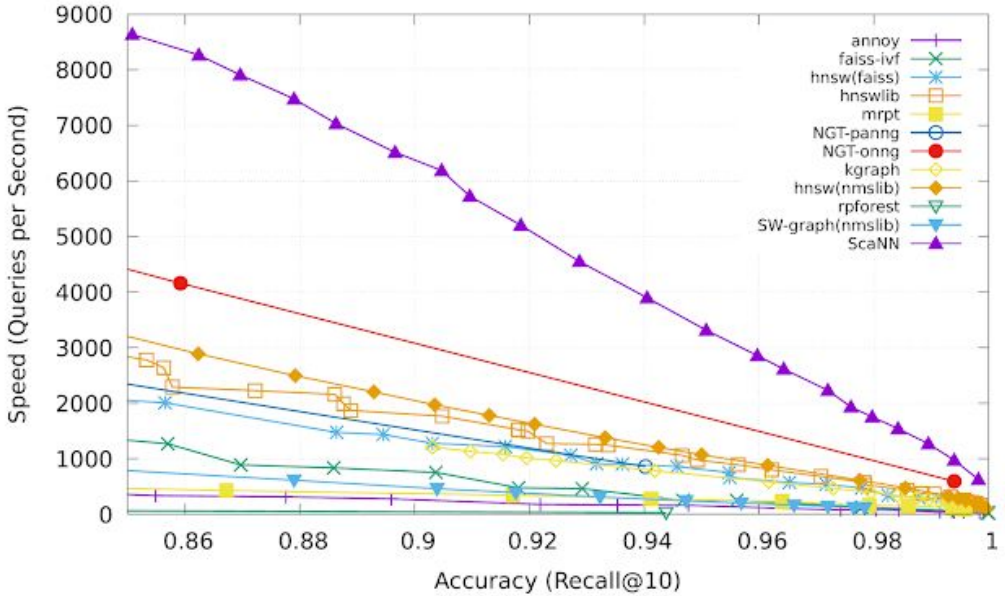
- Retrieval Augmented Generation (RAG) design patterns
- Long-term memory for LLMs
- Semantic Search - search based on meaning vs keyword
- Similarity Search - Identify text similar to other text
- Recommendations



ScaNN: The vector search service for Google Search, YouTube and Play

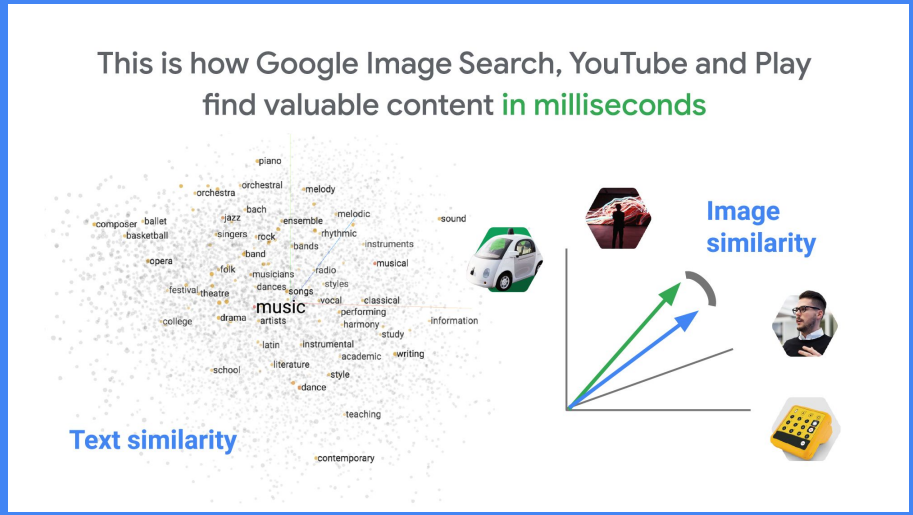
Achieving higher accuracy with shorter latency





This is how Google Search, YouTube, Play and etc searches and recommends relevant content

- Recommendation engines
- Ad / promotion targeting
- Image similarity search
- Semantic text search
- Audio search
- Question answering
- Chat bots
- Anomaly detection
- Data labeling
- and more...



Google's recognized as an industry leader

Gartner, Magic Quadrant for Cloud Database Management Systems, Adam Ronthal, Henry Cook, Rick Greenwald, Aaron Rosenbaum, Ramke Ramakrishnan, Xingyu Gu, December 18, 2023
Gartner, Magic Quadrant for Cloud AI Developer Services, Jim Scheibmeir, Svetlana Sicular, Arun Batchu, Mike Fang, Van Baker, Frank O'Connor, May 22, 2023
Disclaimer: Gartner does not endorse any vendor, product or service depicted in its research publications, and does not advise technology users to select only those vendors with the highest ratings or other designation. Gartner research publications consist of the opinions of Gartner's research organization and should not be construed as statements of fact. Gartner disclaims all warranties, expressed or implied, with respect to this research, including any warranties of merchantability or fitness for a particular purpose.
This graphic was published by Gartner, Inc. as part of a larger research document and should be evaluated in the context of the entire document. The Gartner document is available upon request from Google.

The Forrester Wave™: AI Infrastructure, Q4 2021, The Forrester Wave™: Data Security Platforms, Q1 2023, The Forrester Wave™: Streaming Data Platforms Q4 2023, The Forrester Wave™: Cloud Data Warehouse Q2 2023 , The Forrester Wave™: Data Management for Analytics, Q1 2023, The Forrester Wave™ is copyrighted by Forrester Research, Inc. Forrester and Forrester Wave™ are trademarks of Forrester Research, Inc. The Forrester Wave™ is a graphical representation of Forrester's call on a market and is plotted using a detailed spreadsheet with exposed scores, weightings, and comments. Forrester does not endorse any vendor, product, or service depicted in the Forrester Wave. Information is based on best available resources. Opinions reflect judgment at the time and are subject to change.

Google Cloud

A leader / the most visionary
Cloud Database Management Systems — Gartner

A leader
Cloud AI Developer Services — Gartner

A leader
Streaming Data Platforms — FORRESTER®

A leader
Cloud Data Warehouse — FORRESTER®

A leader
Data Management for Analytics — FORRESTER®

A leader
AI Infrastructure — FORRESTER®

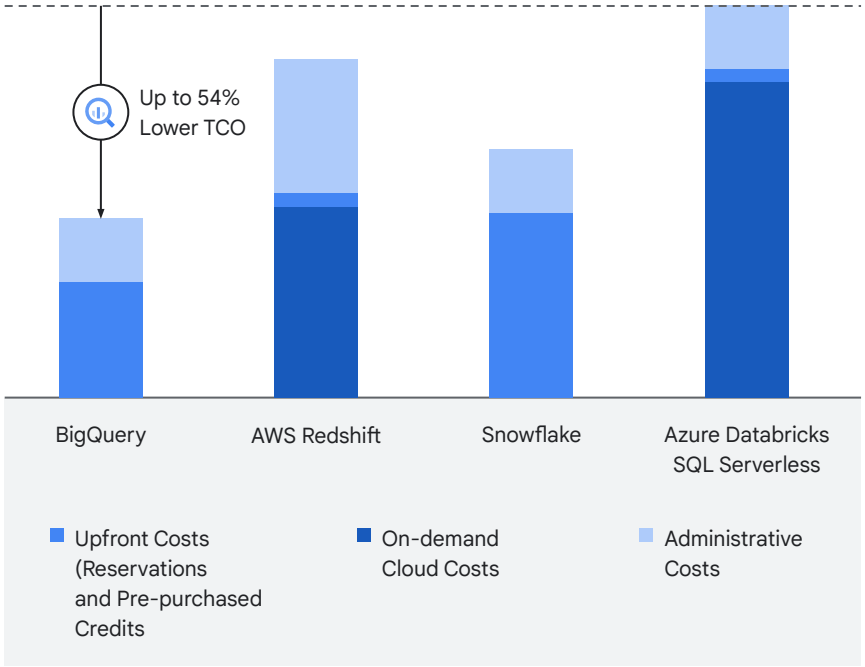
A leader
Unstructured Data Security Platforms — FORRESTER®

A leader
Data & Analytics Platforms Asia Pacific — IDC ANALYZE THE FUTURE

BigQuery offers up to 54% lower TCO over alternative cloud-based EDW solutions

- Upto **54% lower** TCO vs alternative solution
- Lower cost and complexity of running **ML workloads** and **GenAI** use cases
- Eliminate upfront investment and **reduce operational overhead**

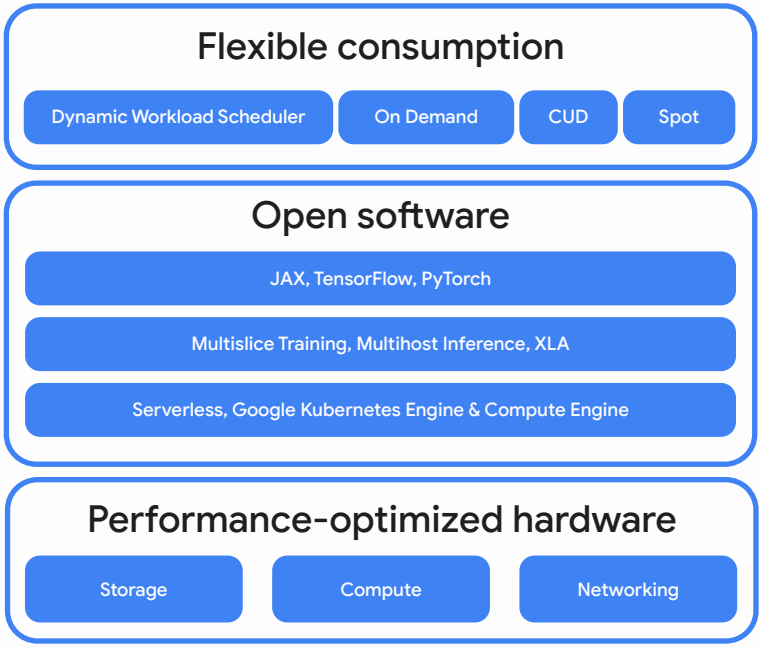
> **BigQuery is the only native cloud solution that was designed from day one with true serverless and streaming capabilities and has been delivering and refining the tools and processes across 10 years of customer deployments."**



Source: ESG 2024, Economic Advantage of Google Cloud BigQuery

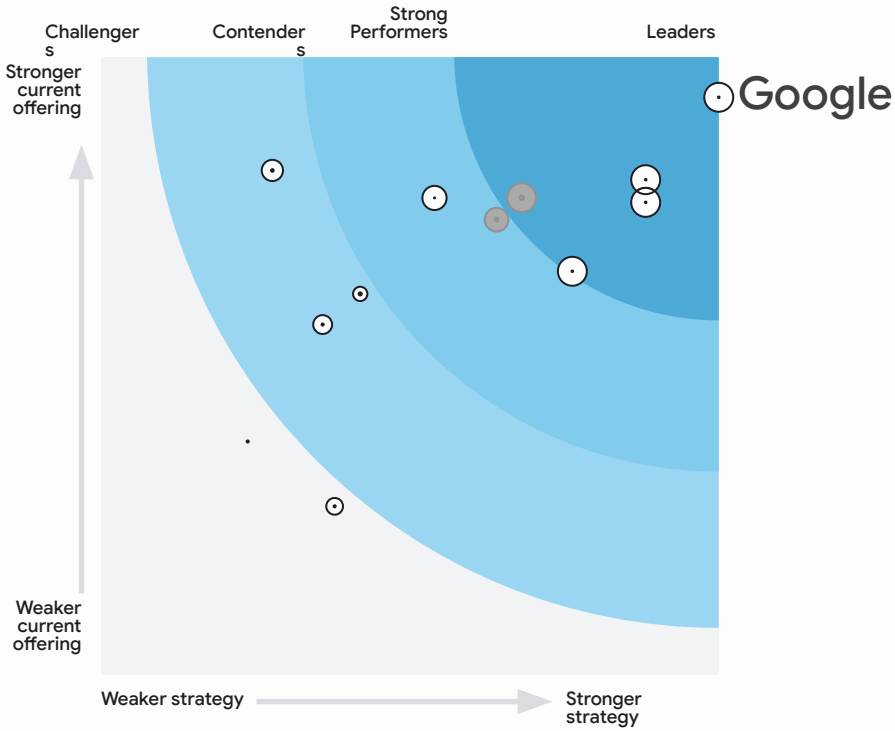
AI Hypercomputer

Next-generation AI supercomputing architecture



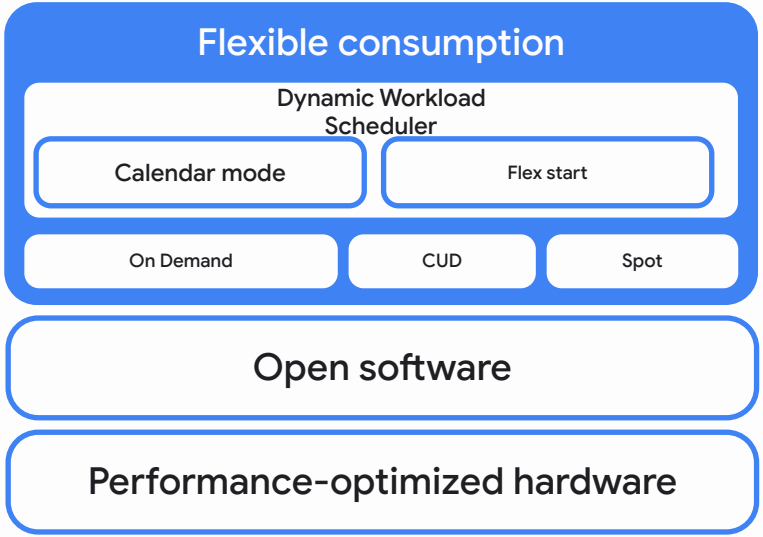
The Forrester Wave:

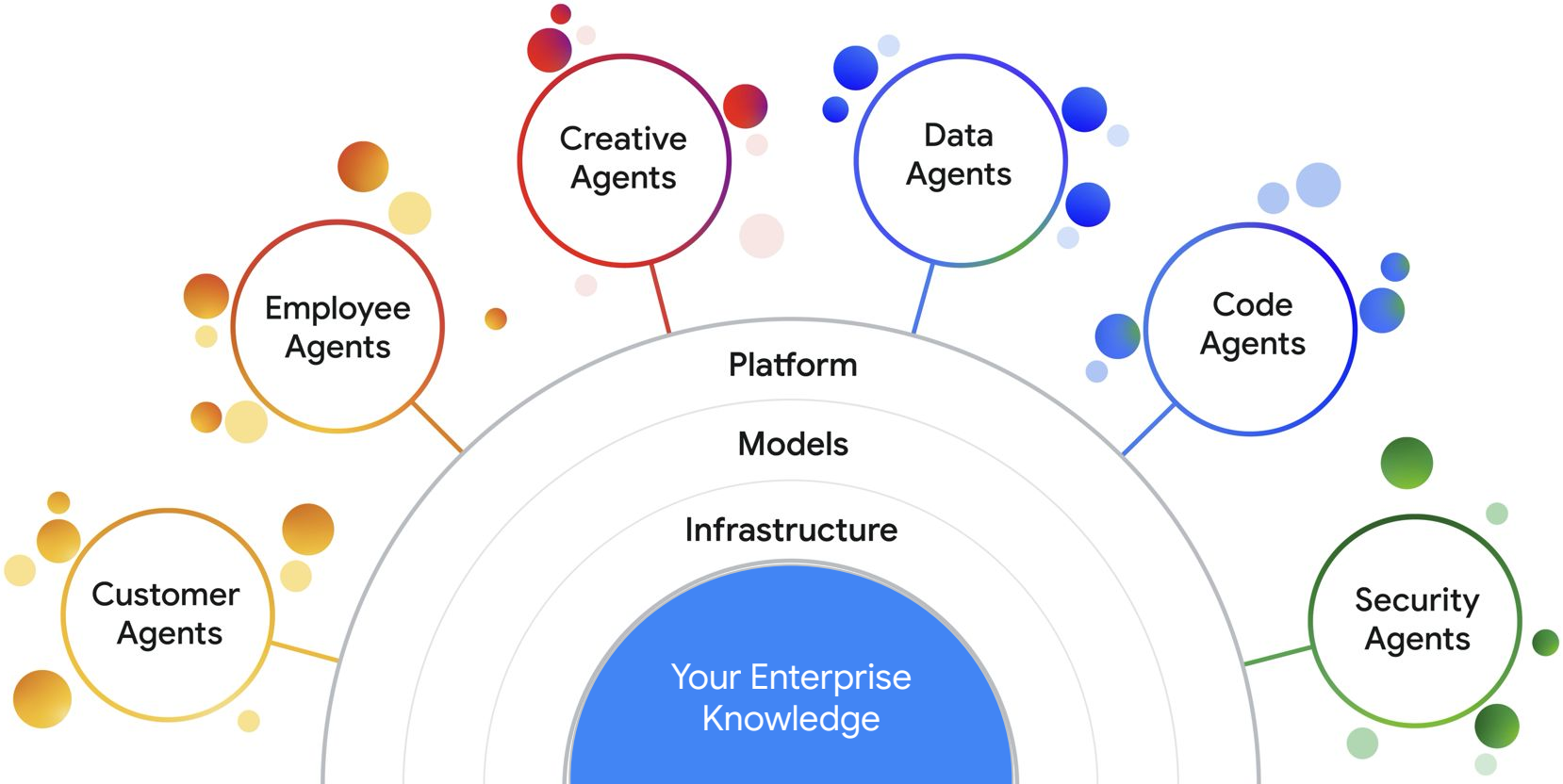
AI Infrastructure Solutions Q1 2024



AI Hypercomputer

Next-generation AI supercomputing architecture




Creative Agents
Data Agents
Employee Agents
Code Agents
Customer Agents
Security Agents
Platform
Models
Infrastructure
Your Enterprise Knowledge

MLOps for Large Models Compare Prompts | Compare Models

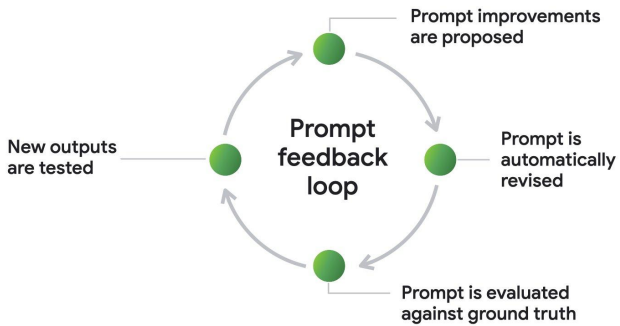
Prompt Management

Prompt Version History & Management: Version and fork prompts and make comparisons with previous edits

Prompt Notes: Custom note fields supported for individual prompts to track progress

Side-by-Side Comparison: Compare prompts side-by-side to vary models or vary prompts.

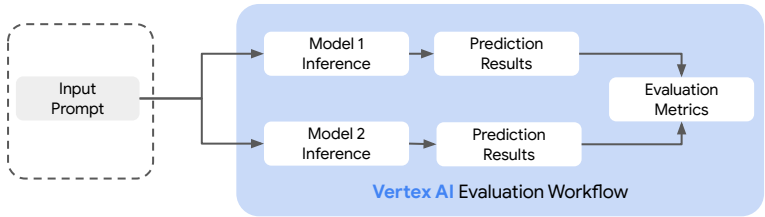
Coming Soon: Custom Prompt Tags for status updates or categorization and AI-Assisted Prompting that provides feedback to rewrite the prompt.

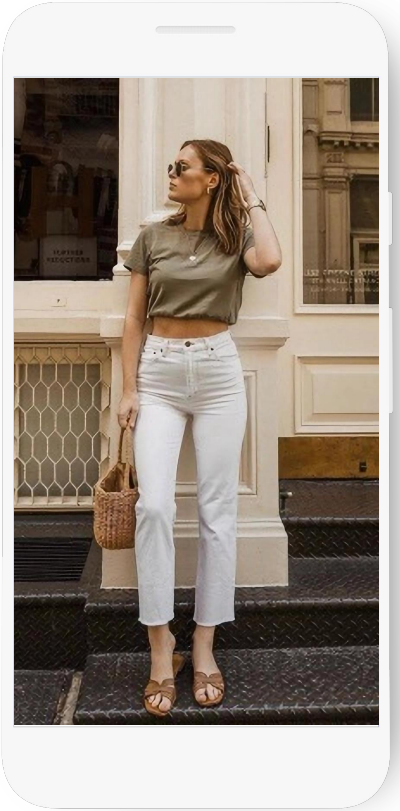


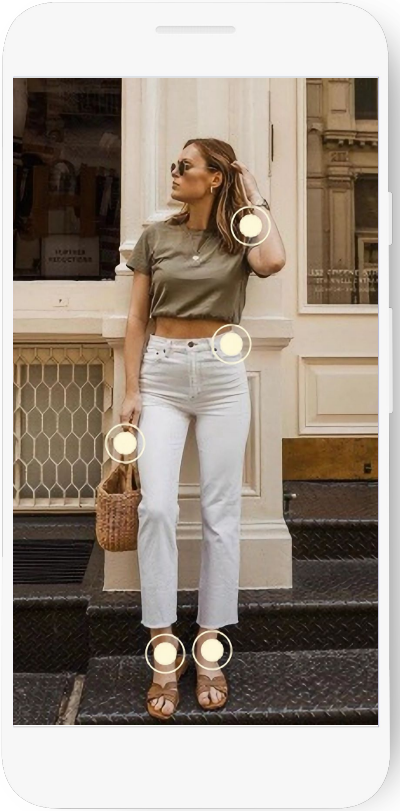
Gen AI Evaluation

Rapid evaluation, now in Preview, lets developers evaluate model performance in seconds based on a small data set

Auto SxS, now Generally Available, can assess the performance of two different models using a large language model, and provides explanations of why one response outperforms another and certainty scores, which helps users understand the accuracy of the evaluation







24x

Click Through Rate

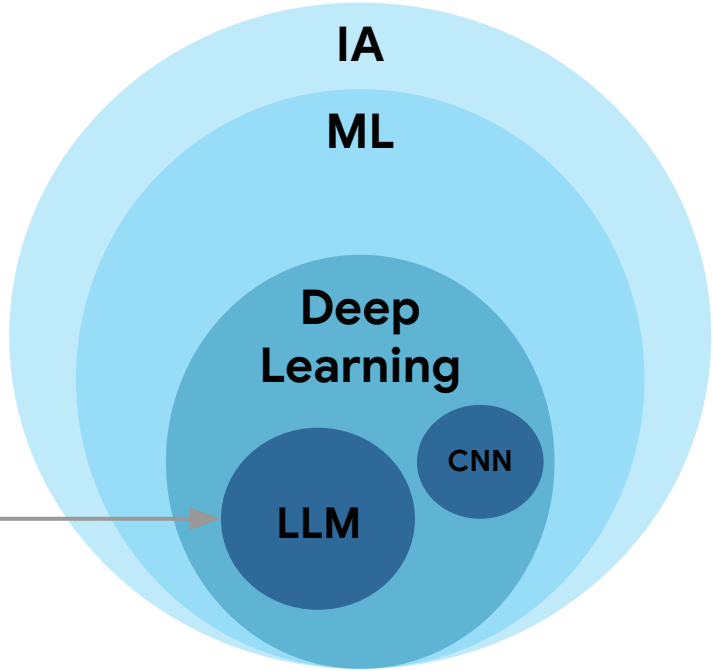
The retailers and the new search experiences in the ecommerce

The technology property of search is moving to retailers rapidly

https://ai-demos.dev/

¿Qué es la IA Generativa?

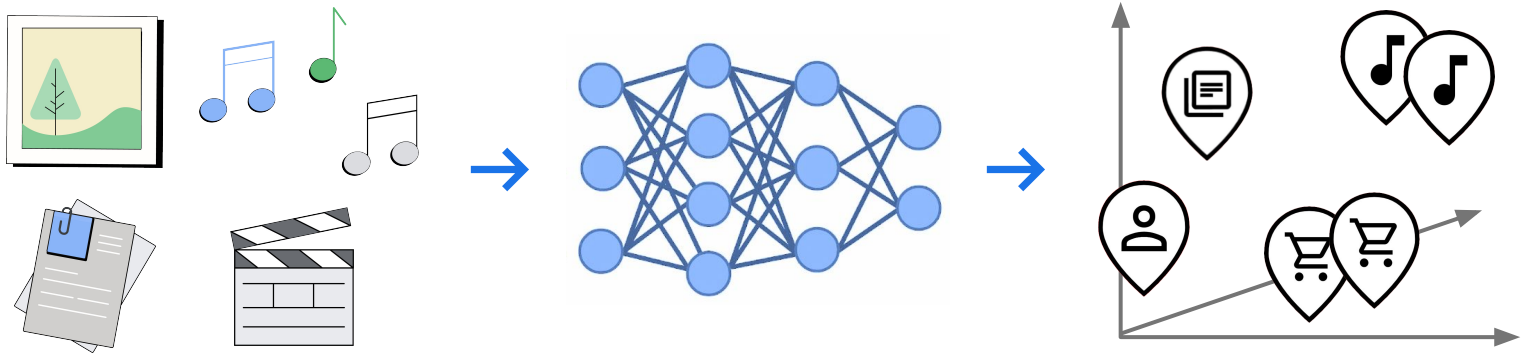
Large Language Models (LLMs) son un subcampo del Deep Learning



Características

- Entrenados con una gran cantidad de datos
- Capaces de generar contenido
- Múltiples usos

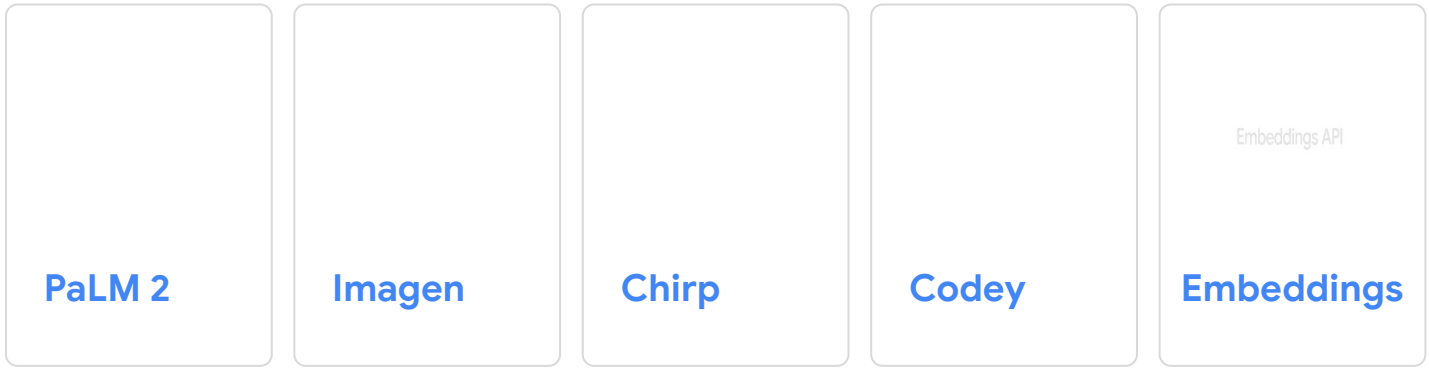
¿Qué es un embedding?



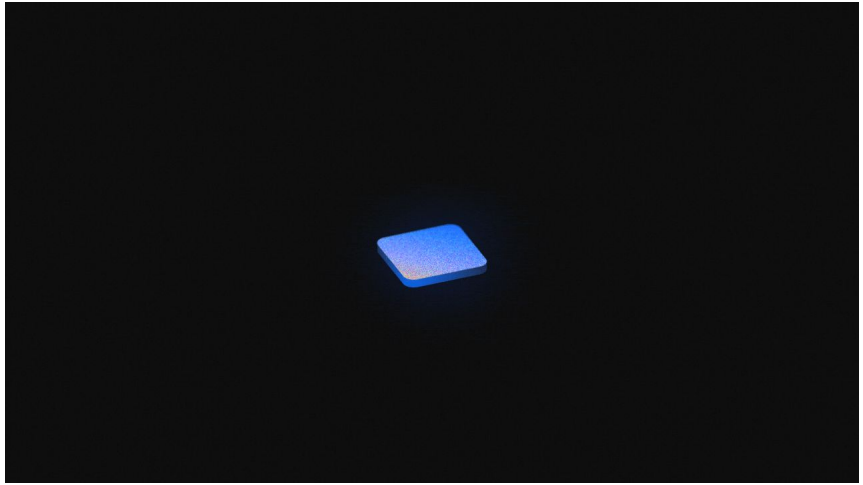
Data (10^4~10^6 dims) **DL models** **Embs (10^2~10^4 dims)**

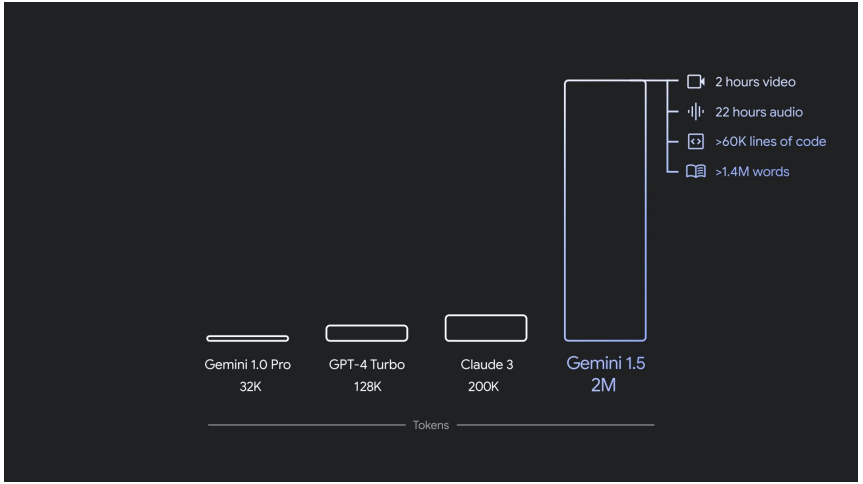
"An embedding is a relatively low-dimensional vector into which you can translate high-dimensional vectors. Ideally, an embedding captures some of the semantics of the input by placing semantically similar inputs close together in the embedding space."

Google Foundation Models



... And Gemini




2 hours video
22 hours audio
>60K lines of code
>1.4M words
Gemini 1.0 Pro
32K
GPT-4 Turbo
128K
Claude 3
200K
Gemini 1.5
2M
Tokens

Vertex AI - One AI platform, every ML tool you need

Predictive AI

Regression & Classification
Forecasting
Sentiment Analysis
Entity Extraction
Object Detection

Generative AI

Text, Image & Code Generation
Text & Code Rewriting & Formatting
Summarization
Extractive Q&A
Image & Video Descriptions

Multimodal Generative AI

Natural Image Understanding
Video Question Answering
Automatic Speech Recognition & Translation
Spatial Reasoning and Logic
Mathematical Reasoning in Visual Contexts

Train	Serve	MLOps	Prompt	Tune	RAG

Experiment Tracking	Pipelines	Model Registry & FeatureStore	Prompting	Tuning & RLHF	Grounding
Explainability	Eval & Monitoring	Embeddings & VectorDB	Function Calling & Extensions	RAG & Agents	Punting & Safety

AI Solution

Contact Center AI | Risk AI | Healthcare Data Engine | Search for Retail, Media and Healthcare

Gemini for Google Cloud

Gemini for Google Workspace

Build your own generative AI-powered agent

Vertex AI

Vertex AI Agent Builder

OOTB and custom Agents | Search
Orchestration | Extensions | Connectors | Document Processors | Retrieval engines | Rankers | Grounding

Vertex AI Model Builder

Prompt | Serve | Tune | Distill | Eval | Notebooks I Training | Feature Store | Pipelines | Monitoring

Vertex AI Model Garden

Google | Open | Partner

Google Cloud Infrastructure (GPU/TPU) | Google Data Cloud

Google Cloud AI

New Stack

AI Solution

Contact Center AI | Risk AI | Healthcare Data Engine | Search for Retail, Media and Healthcare

Gemini Agents

Build your own generative AI-powered agents

Vertex AI

Presentation Focus

Vertex AI Agent Builder

OOTB and custom Agents | Search

Orchestration | Extensions | Connectors | Document Processors | Retrieval engines | Rankers | Grounding

Vertex AI Model Builder

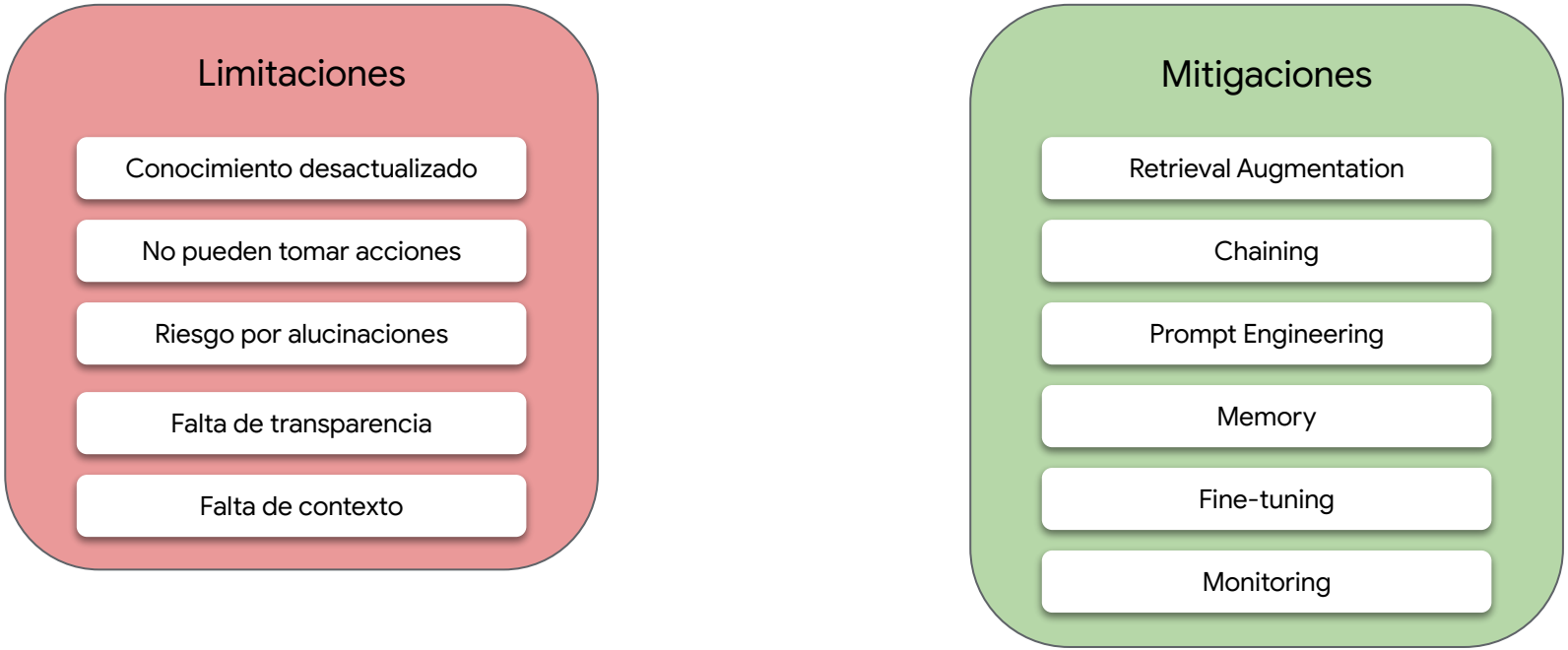
Prompt | Serve | Tune | Distill | Eval | Notebooks | Training | Feature Store | Pipelines | Monitoring

Vertex AI Model Garden

Google | Open | Partner

Google Cloud Infrastructure (GPU/TPU) | Google Data Cloud

El lado oscuro de los LLMs



What is LangChain?

LangChain is a framework for developing applications powered by large language models (LLMs).

It makes **easier** to work & build systems with language models.

Integration: bring external data such as your files or APIs

Agents: build systems capable of interacting with its environment and take decisions.

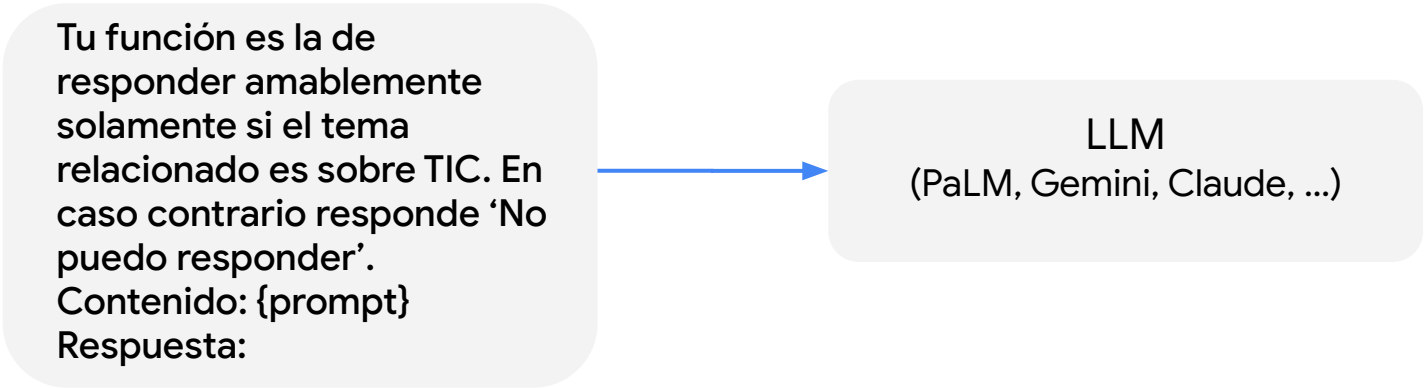

LangChain
Chains
Generic Functionality (LLM, Sequential)
Index Related (QA, Summarization)
Custom Chains
Agents
Agent Types (zero-shot-react)
Tools & Toolkits
Agent Executors
Memory
ConversationBufferMemory
ConversationBufferWindowMemory
ChatMessageHistory
VectorStore-Backed Memory
Models
LLMS
Chat
Embeddings
Prompts
Templates
Output Parsers
Example Selectors
Indexes
Document Loaders
Text Splitters
Vector Stores
Retrievers

Chains en Langchain

Los Chains es una de las piezas más importantes de Langchain para construir pipelines reusables.

Un Chain es una secuencia de llamadas a componentes, que pueden llegar a ser otros Chains.

Ejemplo sencillo Chain



Agents en Langchain

Un Agent es una entidad autónoma capaz de tomar acciones para alcanzar una tarea u objetivo. Sirven para manejar workflow complejos, de múltiples pasos e interacción continua como chatbots.

Chain vs Agent

Son conceptos similares. Ambos usan LLMs para conseguir su objetivo/tarea.

Chain

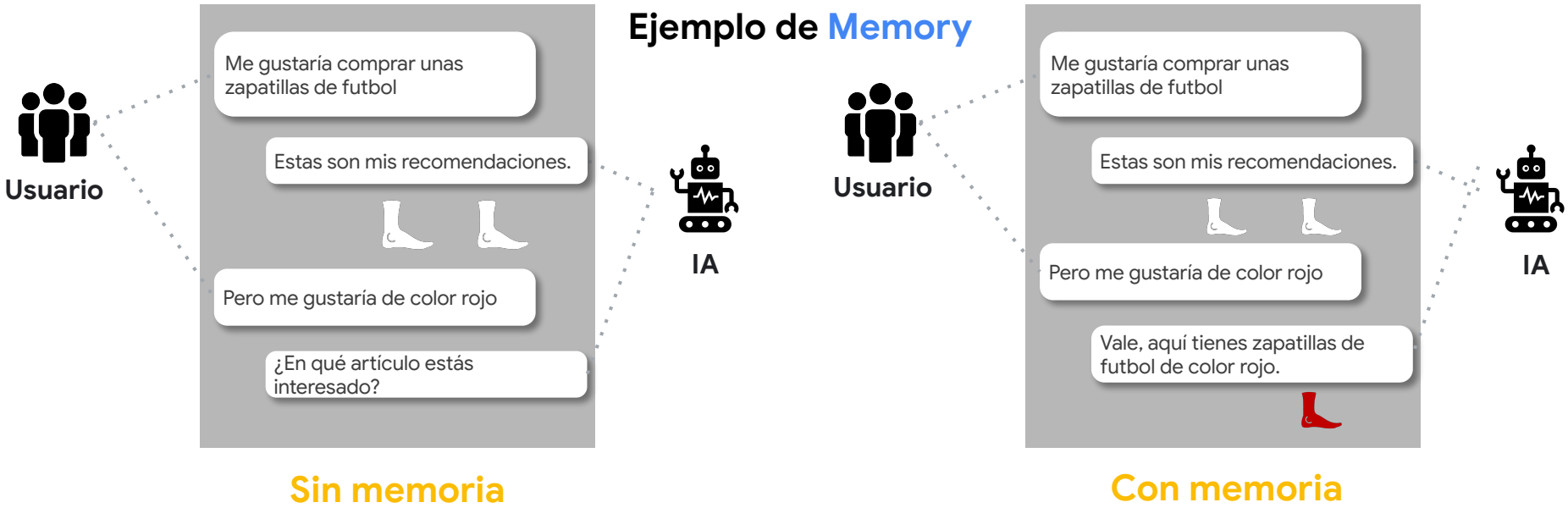
Definen lógica reusable con componentes de manera secuencial **orquestando módulos de bajo nivel**.

Agent

Los Agents **combinan y orquestan Tools (que pueden ser Chains)** para conseguir el objetivo. Observan el entorno y deciden de manera autónoma qué Tool ejecutar y cuándo basado en la observación.

Memory en Langchain

Memory permite persistir la información entre ejecuciones de un Chain o Agent para que nuestra LLM App pase a tener de statelessness -> statefullness.



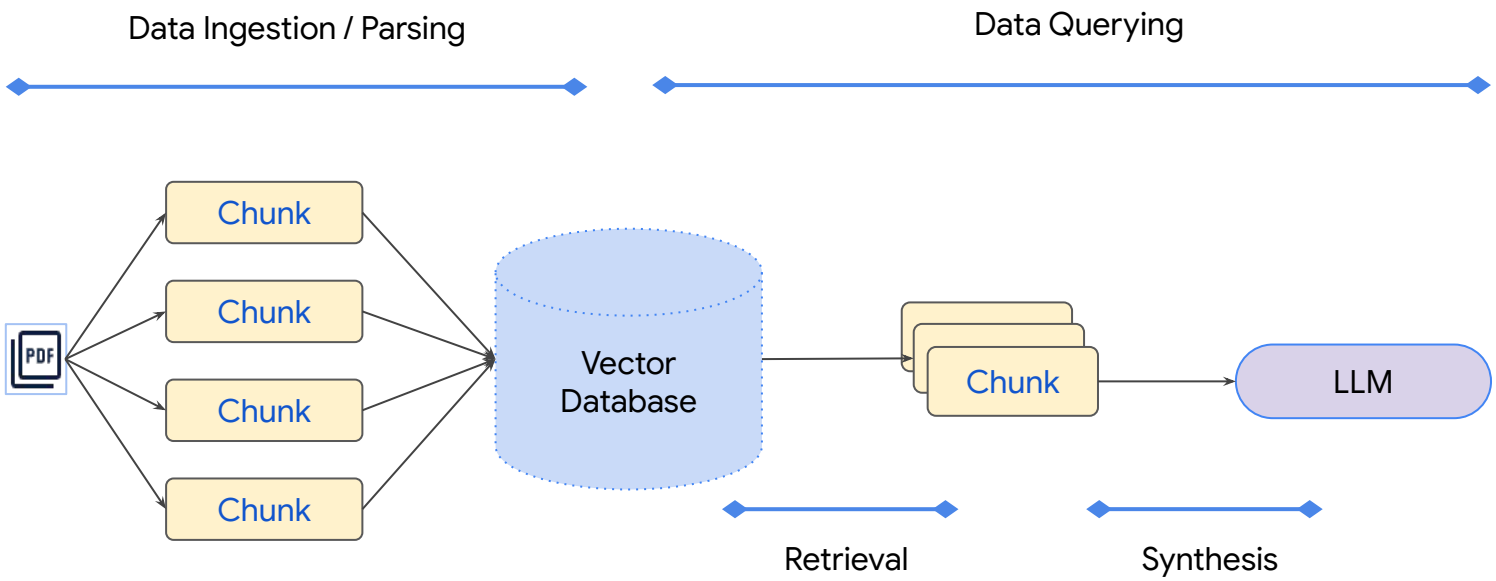
Tools en Langchain

Una Tool da la capacidad de integrar servicios externos como BBDDs, APIs con agentes.

Ejemplos de Tools

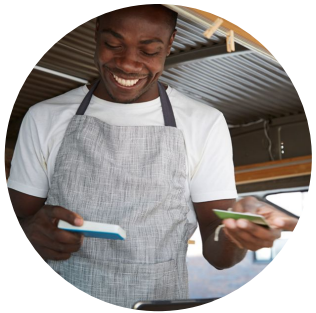
- Search Engine
- Wikipedia
- Stocks API
- Calculator
- Maps
- Wheater
- Custom Tool

Workflow para construir un agente con datos propios



Problema a resolver

Problema a resolver



Product Owner

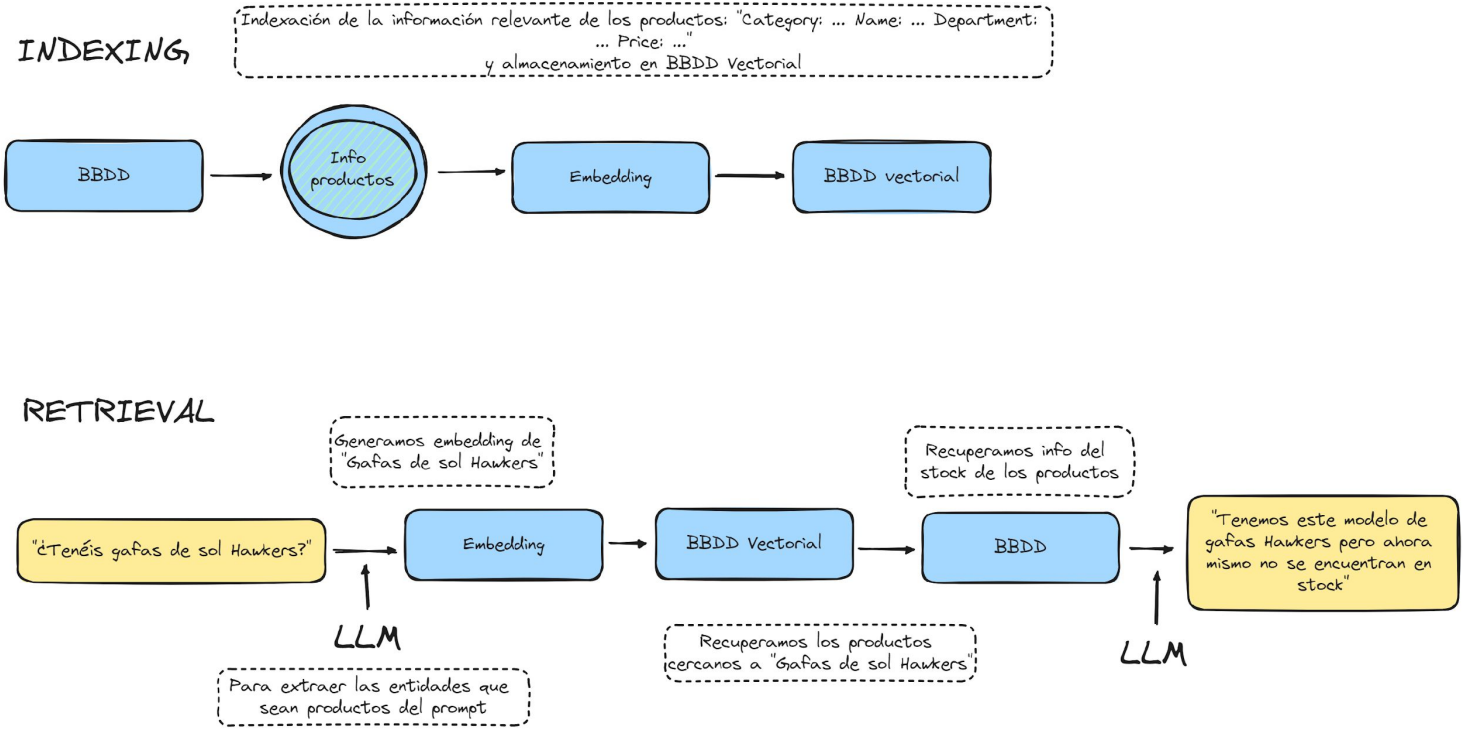
Data & AI Team

Formamos parte del equipo de IA de una empresa que vende productos online a través de su página de e-commerce. La empresa está comenzando a tener **problemas por la carga de trabajo que supone el servicio de atención al cliente**. Además de reducir la intervención humana, necesitamos una manera de dar soporte inmediato y continuo. La información ofrecida debe ser coherente y consistente con los productos que tenemos en ese momento

"Se me ocurre incluir un chatbot en nuestra página de e-commerce donde los clientes puedan entrar y conversar igual que si estuvieran con un dependiente de nuestras tiendas. Sería genial que les sugiriese productos, hable de nuestra marca y que incluso puedas pasarle fotos para obtener productos relacionados."

Diseño de la solución

Diseño Lógico de la solución

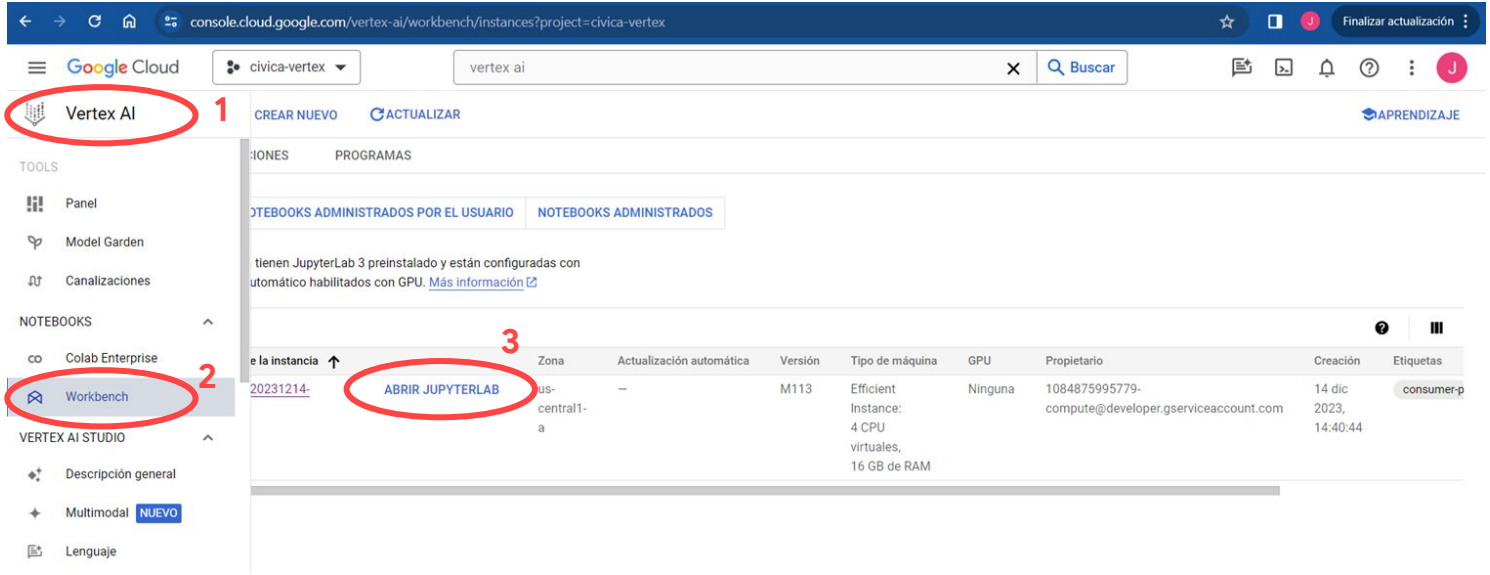


Diseño Técnico de la solución

Hands on Lab!

Pasemos a la acción

Setup



Bloque 1: Indexación

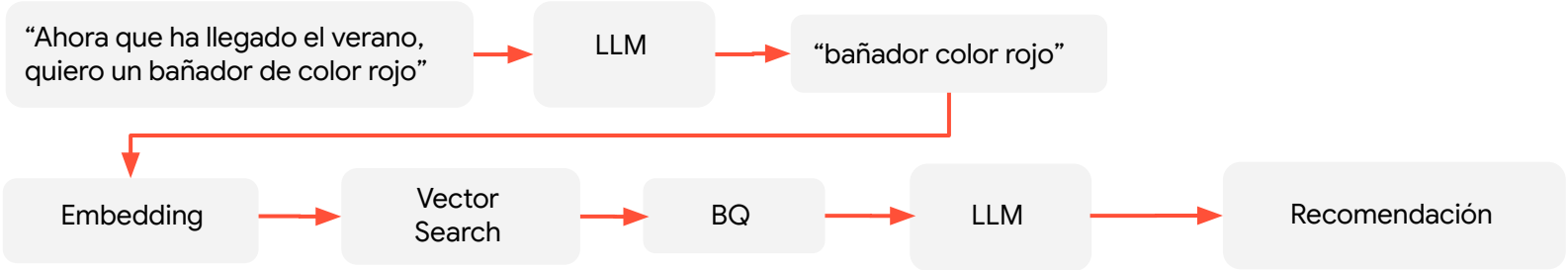
- Creamos embeddings de los productos en BigQuery
- Creamos un índice
- Insertamos los embeddings en el índice
- Creamos un endpoint para el índice

Bloque 2: Búsqueda de los items más cercanos

Coffee Break

Bloque 3: Langchain

- Langchain: obtener producto del prompt del usuario para generar el embedding
- Langchain: obtener producto más similar en el índice, recuperar más información de Big Query y recomendarlo



Bloque 4: User Interface

Implementar función *bot_answer*

Invocar al agente generado anteriormente

Bloque 5: Multimodal

¡Gracias!

Para más información:

comercial@civica-soft.com

www.cívica-soft.com

www.linkedin.com/company/cívica-soft

cívica

Google Cloud