DoiT Workshop

Google Cloud Vertex
AI & Gen AI

Agenda

01 Vertex AI Overview

02 Generative AI Overview

03 Jupyter Notebooks Overview

04 Reasoning Engine

05 Demo

Vertex AI Overview

Building a first ML solution

Identify business problem

1

Develop hypothesis

2

Acquire and explore data

3

Build a model

4

Train the model

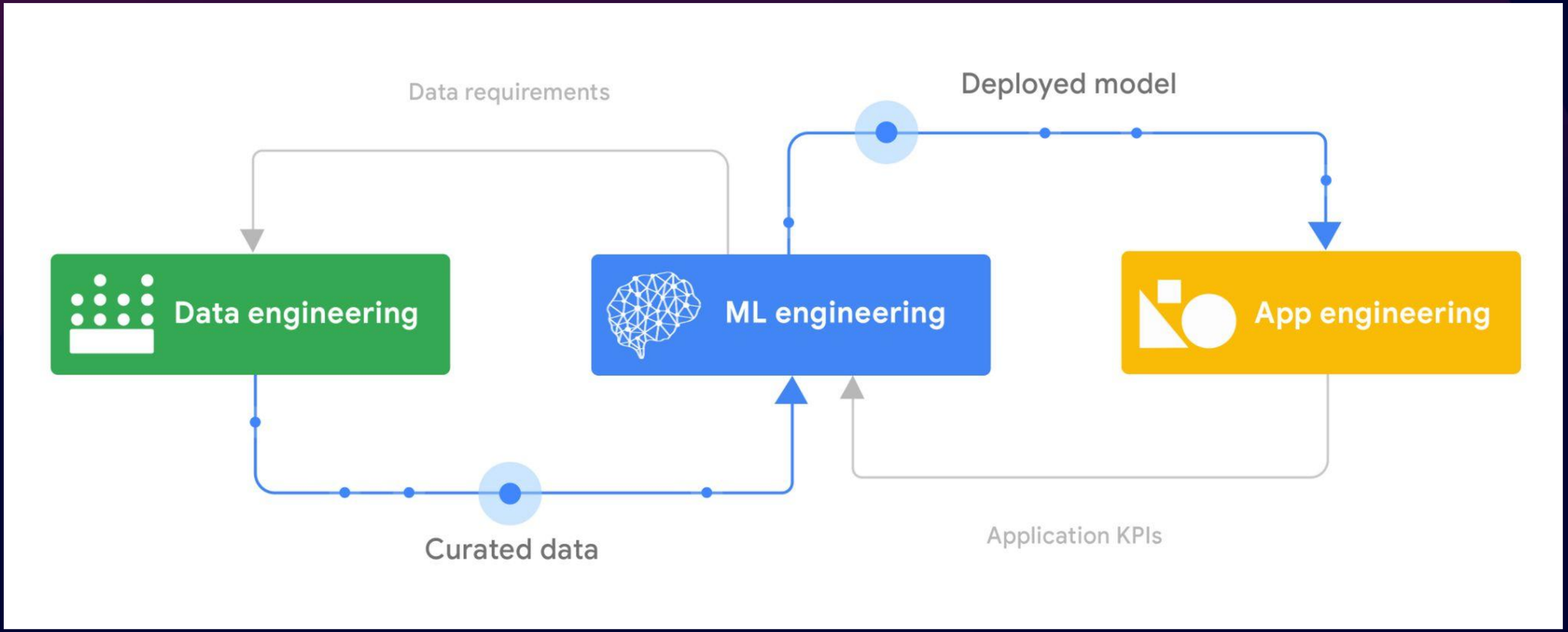
5

Apply and scale

6

Building an ML-enabled system

Is a multifaceted undertaking that combines data engineering, ML engineering, and application engineering tasks



The end-to-end ML journey is long and multidisciplinary

Where can I find training data?

How can I track the results of experiments?

How do I deploy?

Datasets

Feature Store

Workbench

Experiments

Training

Endpoints

Pipelines

Monitoring

And for production?

Where do I start with model experiments?

How can I train at scale?

Vertex AI

A unified ML platform for **solving all business problems**

- **Unified** development and deployment platform for machine learning **at scale**
- Increase **productivity** of data scientists and ML engineers
- Improve **time to value**

No code/low code workflow

- **AutoML**: Vision, Video, Language, Speech, Translation, Tables, Forecast, Tabnet, Workflows
- **BigQuery**: BigQuery ML
- **Generative AI Studio | Generative AI APIs | Model Garden**

Data Science tool kit

- **Workbench**
- **Integration with Data Services**: Cloud Storage, BigQuery, Spark, Spanner, BI

Custom workflow

- **Experiment**: Datasets, Vertex SDK, Experiments
- **Train**: Training, NAS, Vizier
- **Deploy**: Prediction, Matching Engine, Streaming Ingestion

MLOps

- Model Monitoring
- Explainable AI
- ML Metadata
- Feature Store
- Model Registry
- Model Evaluation
- **Pipelines**

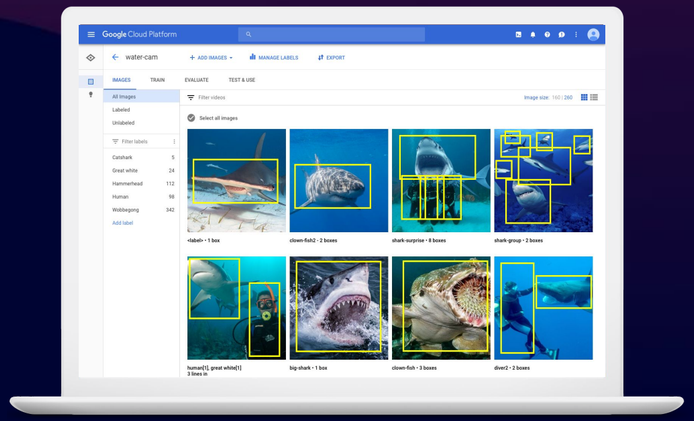
Cloud AI has multiple pathways of consumption

Out-of- the box → DIY

Pre-trained APIs & AI solutions:

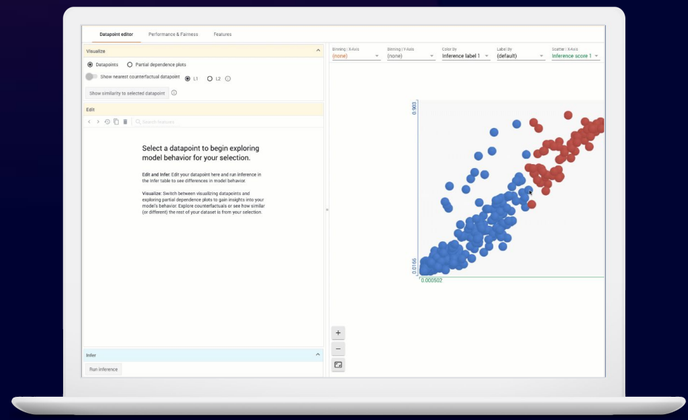
Natural Language
Speech-to-Text
Text-to-Speech
Translation
Vision
Video Intelligence

No training data needed, get started right away!



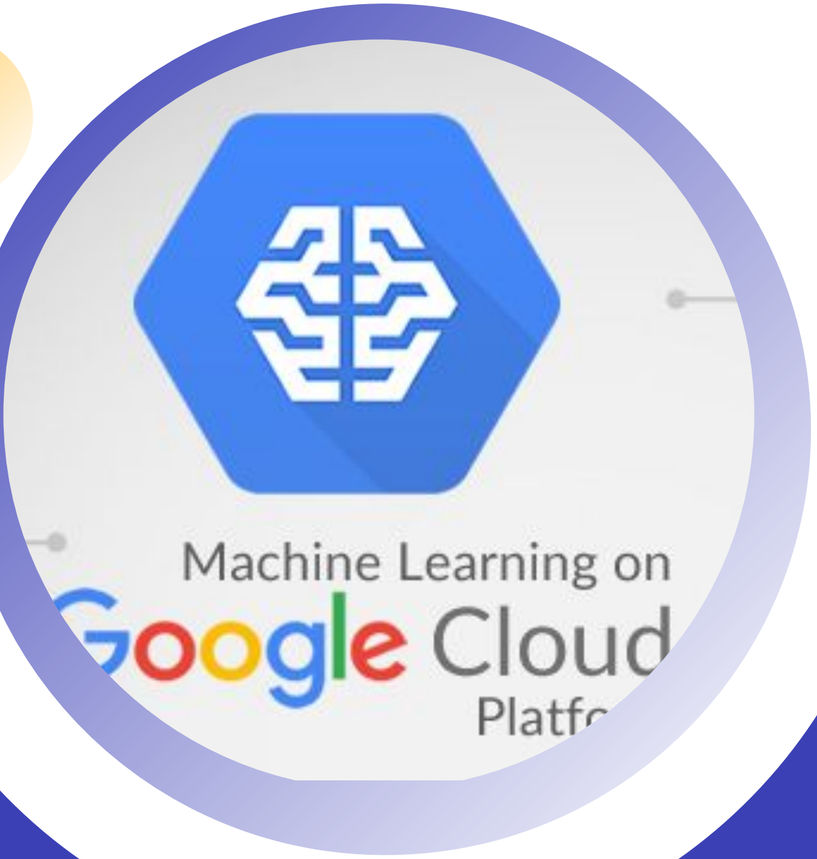
Custom AI with AutoML or BigQuery ML

Easily create custom models (no-code/low-code approach)



End-to-end AI with core tools

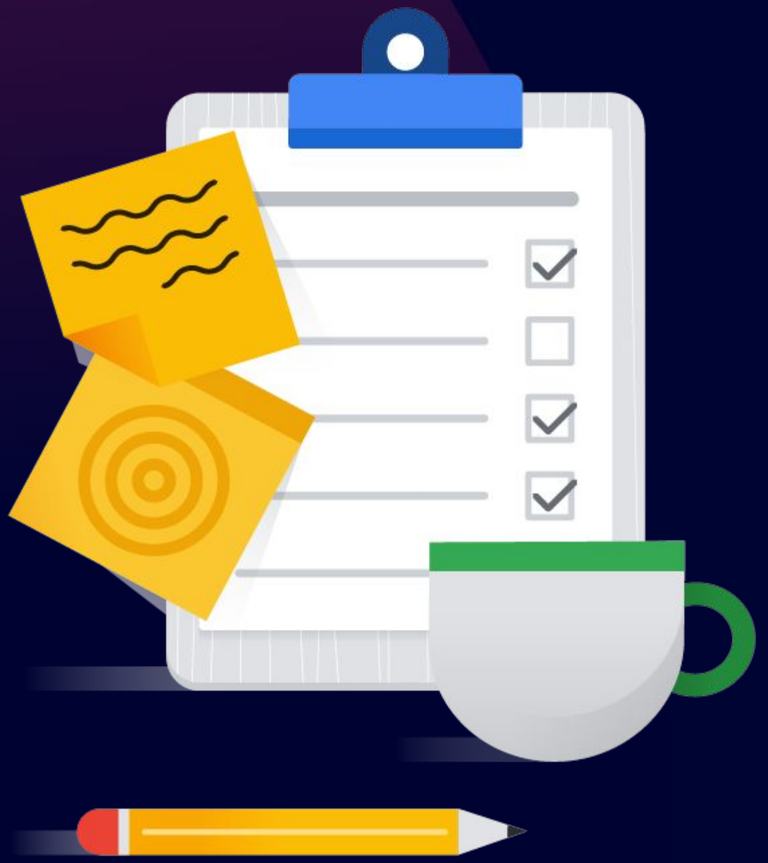
Help data scientists and ML engineers build and deploy AI



Vertex AI Services & Features

Getting Started with Big Data Analytics & Jupyter Notebooks

01	What’s a Notebook?
02	BigQuery magic and ties to Pandas



Vertex AI Workbench

A one-stop surface for Data Science

Fully managed compute with admin control

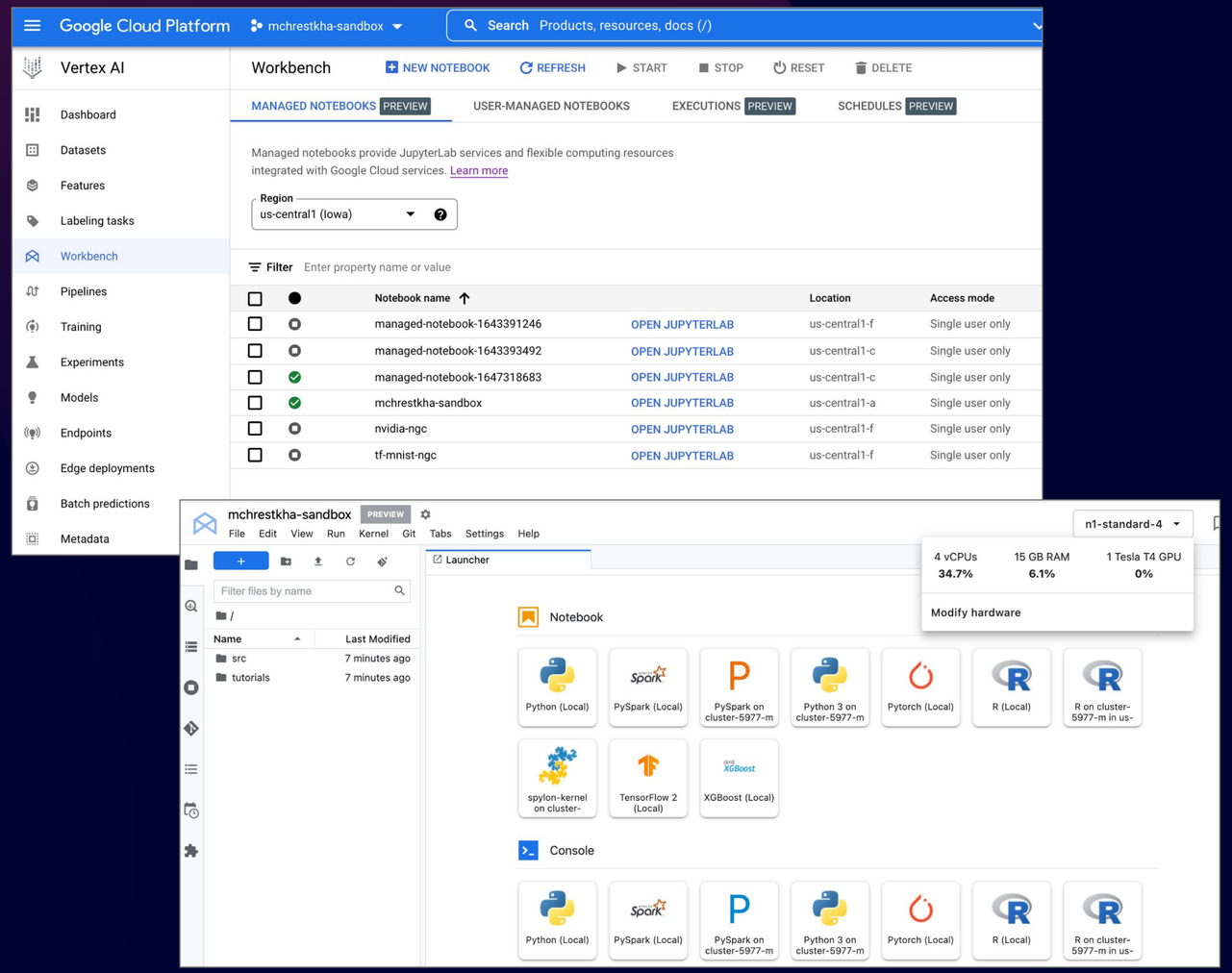
A Jupyter-based fully managed, scalable, enterprise-ready compute infrastructure with easily enforceable policies and user management

Fast workflow for data tasks

Seamless visual and code-based integrations with data & analytics services

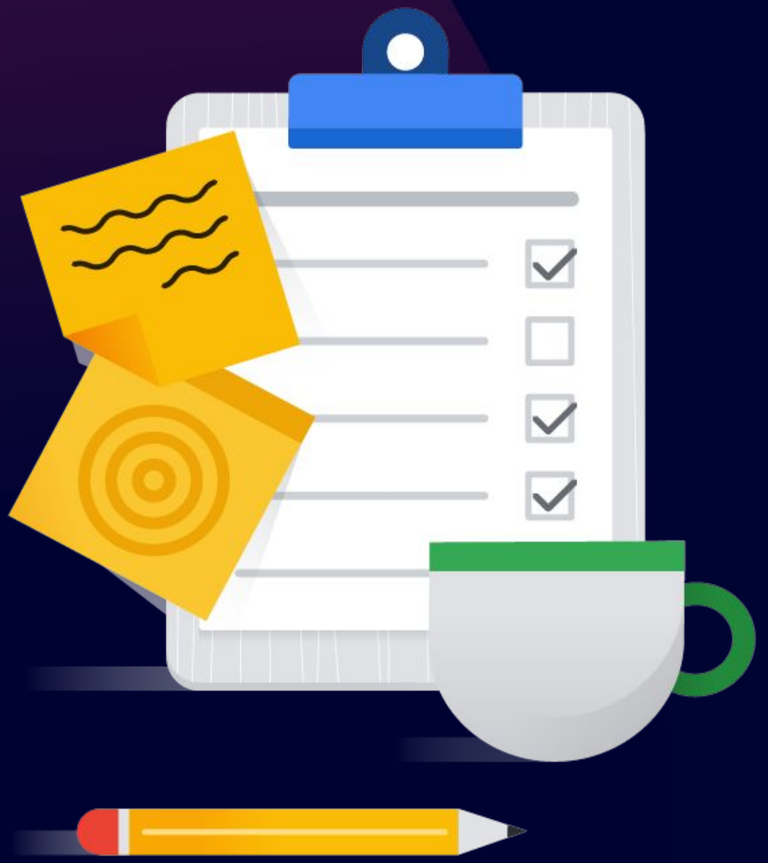
At-your-fingertips integration

Load and share notebooks alongside your AI and data tasks. Run tasks without extra code



Big Data Analytics with Notebooks

01	What’s a Notebook?
02	BigQuery magic and ties to Pandas



Train

Run and scale your code with high availability using Vertex AI Training

Serverless experience, no provisioning

Rapid cluster orchestration

Built-in logging and monitoring

Ephemeral clusters on-demand

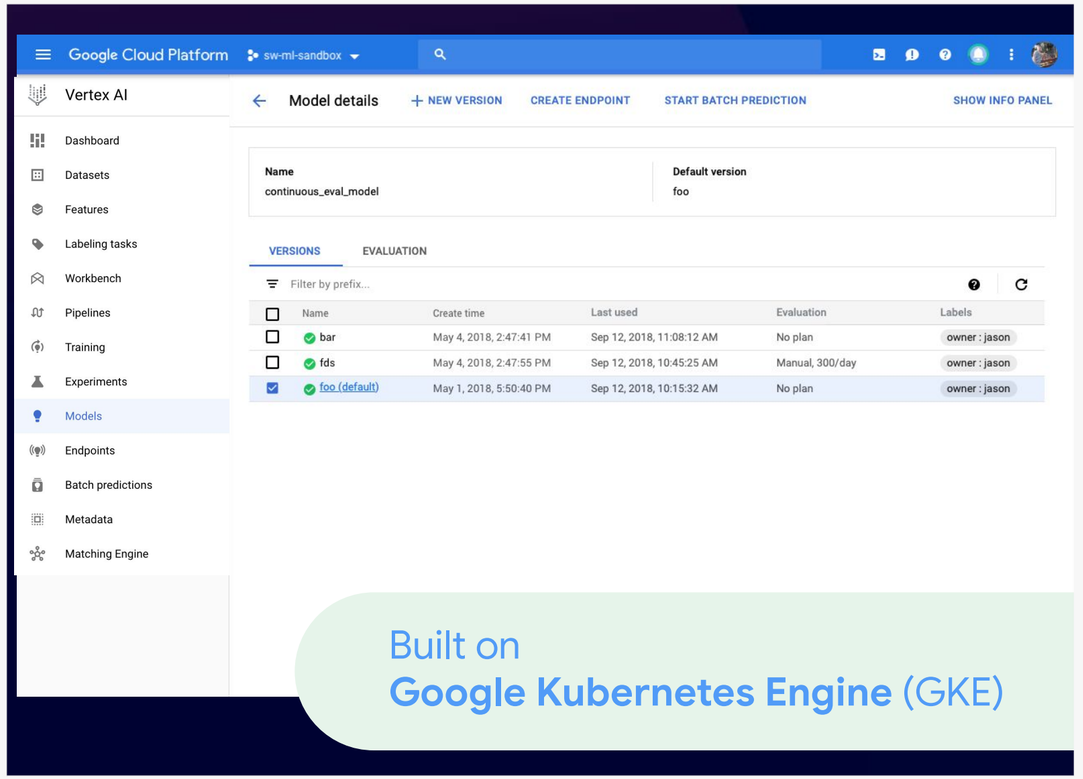
Automatic job queuing

Pay for only what you use

Serve

Robust and reliable model hosting with Vertex AI Prediction

- Serve **online** endpoints for low-latency predictions, or predictions on massive **batches** of data.
- **Built-in security and compliance:** VPC peering and security perimeter. Custom managed encryption keys. Fine-tuned access control.
- **Low TCO:** Scale **automatically** based on your traffic, and alleviate operational overhead.
- **Fast inference on GPUs:** Support for a broad range of machine types specialized for ML, such as GPUs.



Generative AI on Google Cloud

A deep history of research and innovation at Google

2017
Transformer

Google invents Transformer kickstarting LLM revolution

2018
BERT

Google's groundbreaking large language model, BERT

2019
T5

Text-to-Text Transfer Transformer LLM 10B P model open sourced

2020
LaMDA

Google LaMDA model trained to converse

2021
AlphaFold

AlphaFold predicts structures of all known proteins

2022
PaLM

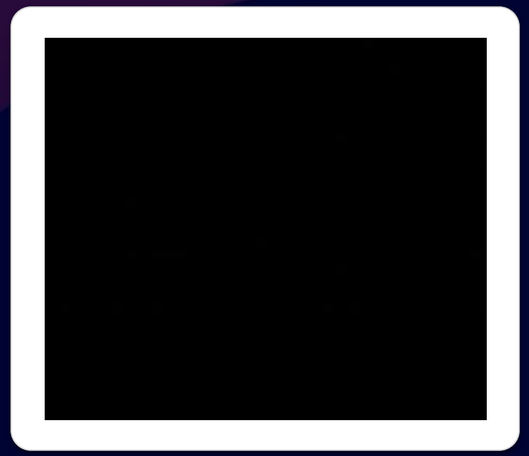
Industry leading large language model

2023
Bard

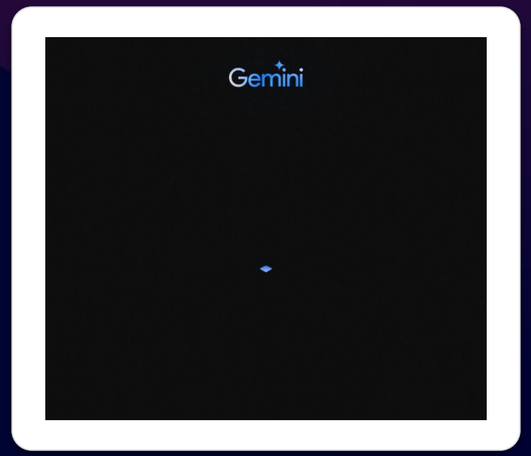
A conversational AI Service powered by LaMDA. (now known as Gemini)

Responsible AI at the foundation

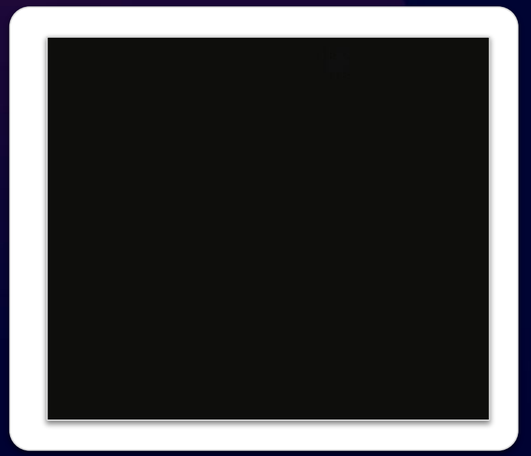
Gemini marks the next phase on our journey to making AI more helpful for everyone



State-of-the-art, natively multimodal reasoning capabilities



Highly optimized while preserving choice



Built with responsibility and safety at the core

Always on AI collaborator for everyone, Gemini

Gemini for Google Workspace

- Helps you write in Gmail and Docs
- Helps you organize in Sheets
- Helps you visualize in Slides
- Helps you connect in Meet

Google Cloud

- Development
- Operations
- Data analysis
- Data science
- Database management
- Security analysis
- Interoperability

- **Democratize generative AI** for all users through **expert assistance** through Google Cloud
- Increase employee productivity in Google Workspace by streamlining **creation, connection, and collaboration**
- Drive **developer efficiency** in the developer IDE, Cloud Console, databases, and security products
- Allow developers to focus on the **most value-add aspects** of their job

Built with responsibility and safety at the core

Gemini includes added new protections to account for multimodal capabilities

The most comprehensive safety evaluations to date

At each stage of development, Google considered potential risks and worked to test and mitigate them. These evaluations applied Google Research's SOTA adversarial testing techniques and included bias and toxicity checks.

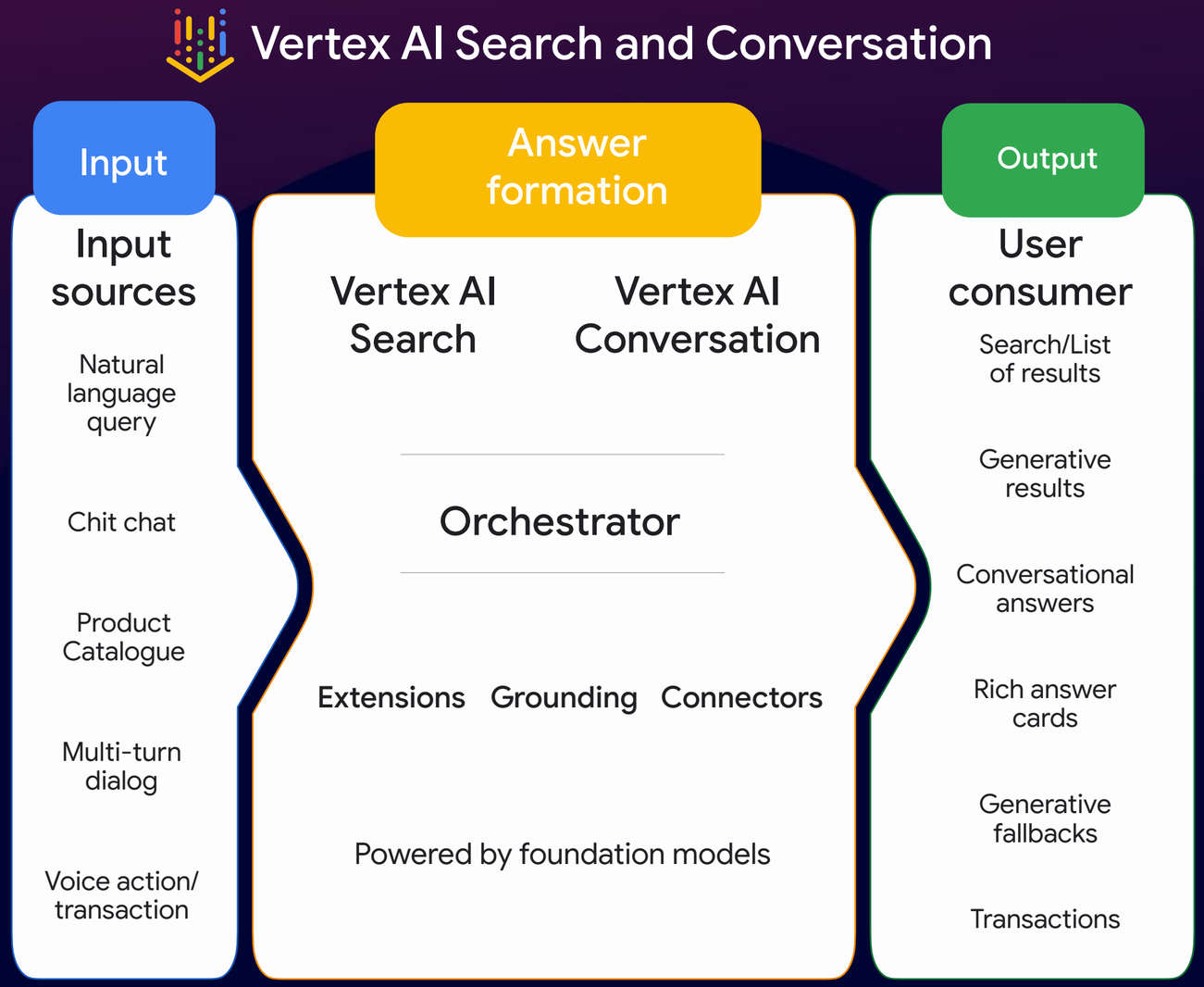
Builds on novel research into potential risk areas

Gemini benefits from the work that Google researchers conducted into novel research into potential risk areas like cyber-offense, persuasion, and autonomy.

Collaboration with a diverse group of external experts and partners

Gemini is evaluated by a diverse group of external experts and partners to help to identify blind spots and ensure that the model is as safe as possible for everyone.

Low code developer platform for gen AI-powered search and conversation experiences



- Empower developers to easily build **personalized** conversational experiences
- Ground foundation models with **fresh and factual data** to reduce hallucinations
- Connect foundation models to **real-time actions and real-time data** to scale helpfulness
- Leverage years of **information retrieval and deep retrieval** experience from Google Search
- Simplify known use cases with **pre-built tasks, playbooks, and orchestrators**

130+ enterprise-ready foundation models in Model Garden

Model Garden on Vertex AI

Gemini foundation models

- Gemini 1.0 Nano
- Gemini 1.5 Pro
- Gemini 1.0 Ultra
- Gemini 1.5 Flash

Google foundation models

- PaLM 2
- Imagen 2 & 3
- Chirp
- Codey
- Embeddings
- Veo

Google task specific models

- Speech-to-Text
- Text-to-Speech
- Cloud Natural Language API
- Translation API
- Enterprise Document OCR
- Occupancy analytics
- Vision API
- Cloud Video Intelligence API

Google domain specific models

- MedLM — Life Science and Healthcare
- Sec-PaLM — Cybersecurity

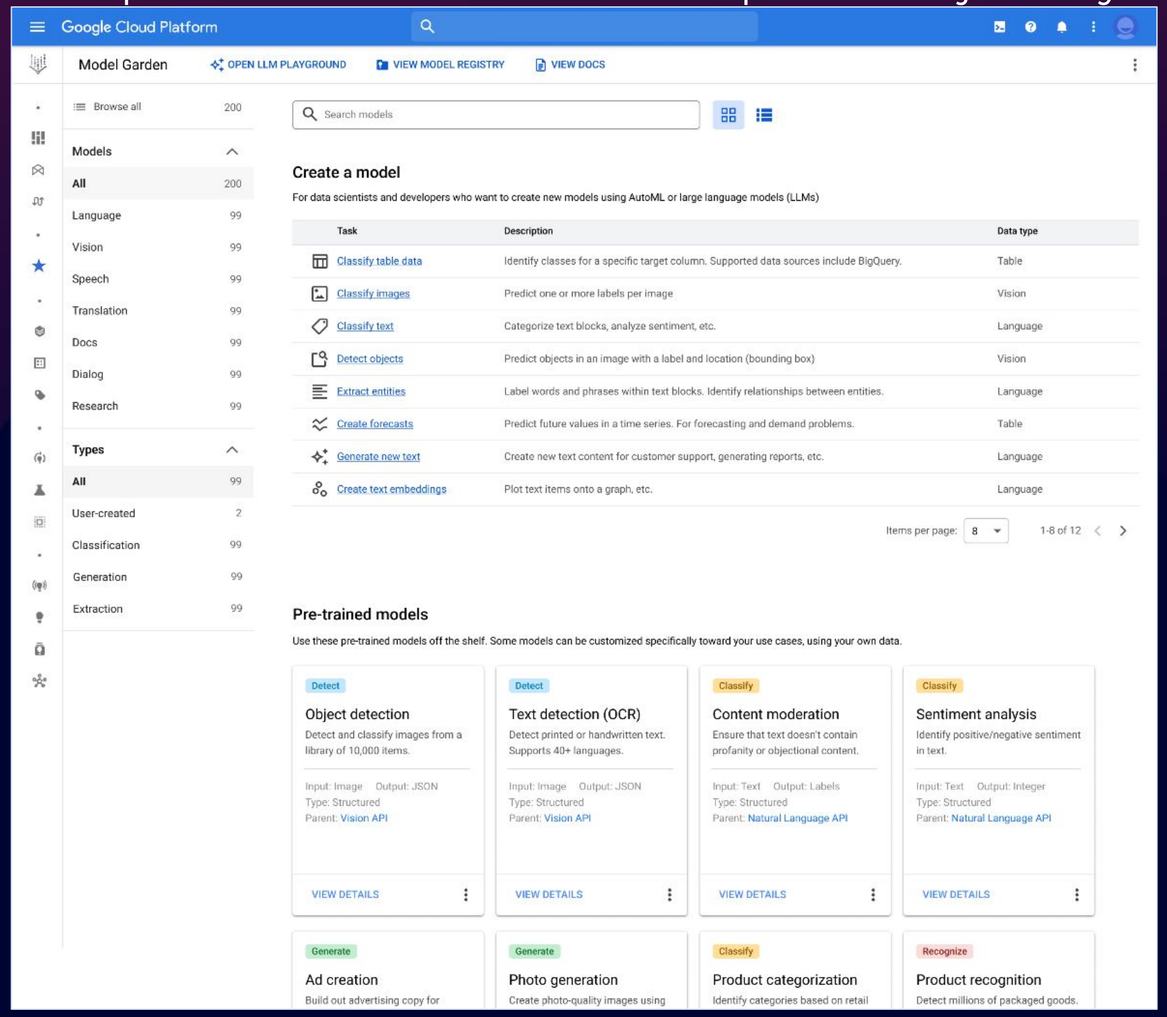
Partner & Open Ecosystem

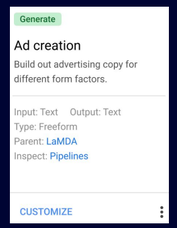
- Llama 2
- Code Llama
- TII Technology Innovation Institute — Falcon
- Claude 2 Pre-announce
- MISTRAL AI_
- Gemma

- **Choice and flexibility** with Google, open source, and third-party foundation models
- **Multiple modalities** to match every use case
- **Multiple model sizes** to match cost and efficacy needs
- **Domain-specific models** for specialized industries
- Enterprise ready with **safety, security, and responsibility**
- Decrease time to value with **fully integrated platform**

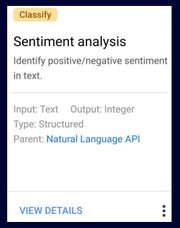
Vertex Model Garden

One place to launch various enterprise user journeys

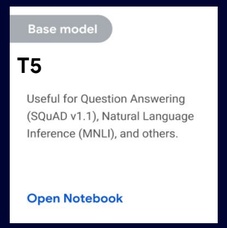




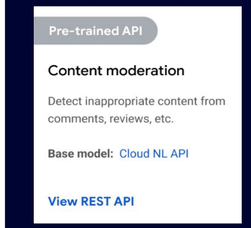
Use LLM directly for large variety of use cases out of box.



Finetune LLM with additional data for targeted industry or use cases.



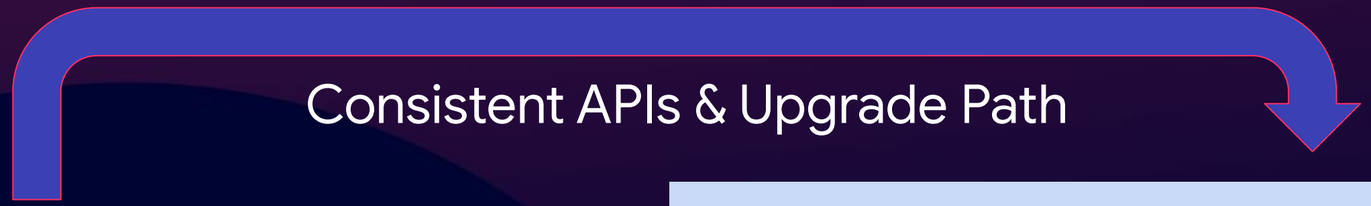
Customize open source structured data models via Data Science Notebooks.



API access to closed sourced models inc. AutoML and pre-trained APIs

One stop shop to find 1st party, open source and 3rd party models.

Gemini Pro is available on Google AI Studio & Vertex AI Studio



Google AI Studio

The fastest way to build with Gemini

- Free-web-based developer tool
- Google/Gmail account
- Prototype and launch apps quickly with an API key
- Data sharing with Google for model & service improvement

Vertex AI Studio (on Google Cloud)

For developers who want scale in production

- Enterprise-ready AI platform
- Google Cloud account
- Data privacy, indemnity protection, VPC Service Controls, CMEK, AXT, & data residency
- Tooling to customize, augment, deploy, and govern models

Document AI Workbench

For creators (UI), app developers (API), and AI practitioners (fine-tuning)

Custom Document Extractor powered by gen AI

Public Preview

Extract data from documents with no training

Use out of the box to extract fields from documents

No training required
Extract data from docs up to 200 pages

Save time by auto-labeling

Use foundation model to prepare dataset prior to training a custom model

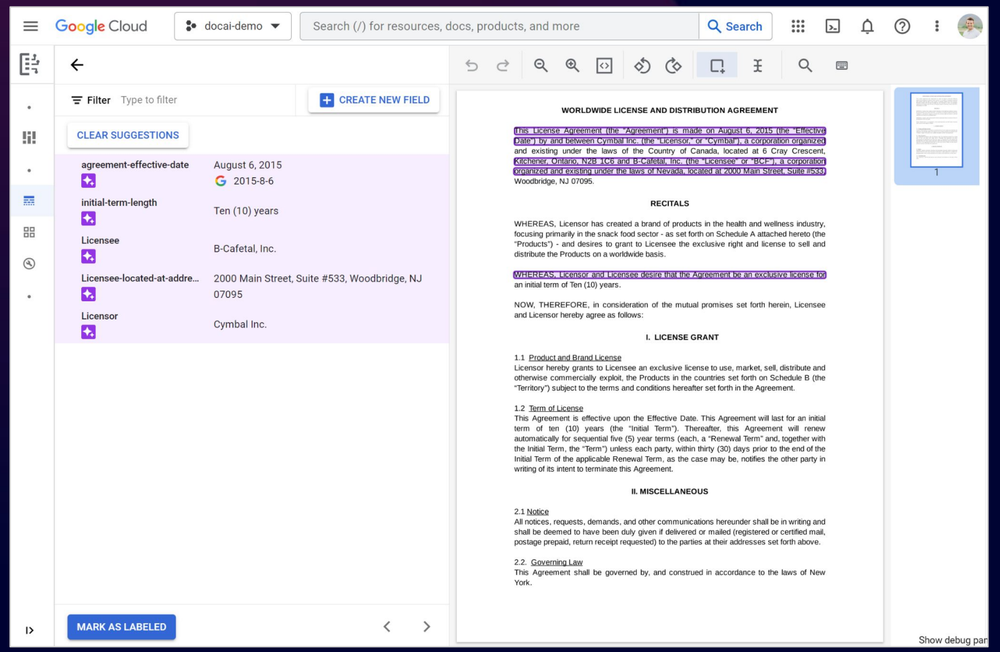
Typical use cases

Free-form, dense text documents (ex: contracts)
Semi-structured documents with layout variation
When little or no training data is available

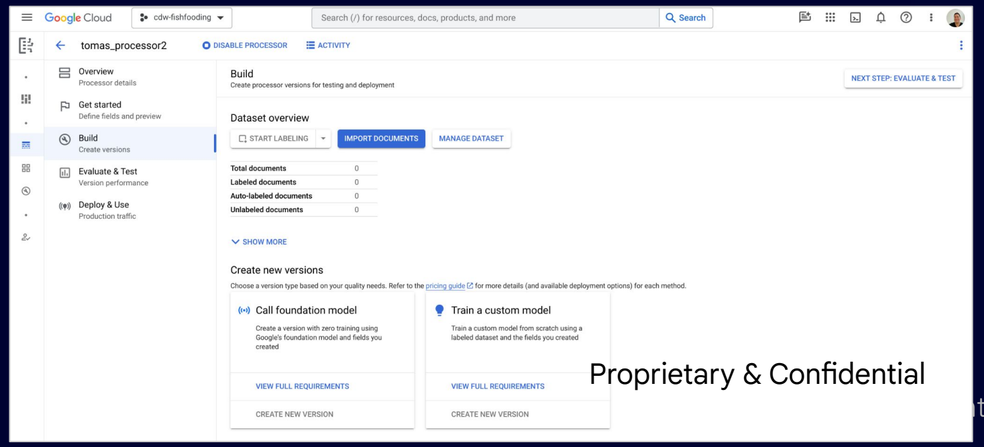
Get started

Landing page, Cloud Console

Create schema and test extraction



Chose either foundation model or train a custom model



Gemma at a Glance

Gemma is a family of lightweight, state-of-the art open models built from the same research and technology used to create Gemini

Gemma **2B and 7B**, the first models in the family, redefine portability with state-of-the-art performance at each size. Available as a base and instruction-tuned model.

Responsible by design
With class-leading safety built-in from datasets through tuning, you can build responsible, modern AI solutions.

Breakthrough performance
Gemma models achieve exceptional benchmark results, even outperforming some larger open models.

Multi-framework
TPU and GPU optimized, multi-framework support, seamless device compatibility, and model customization empower AI development.

Community-driven innovation
Open access to models, world-class partners and extensive ecosystem fosters collaboration and AI advancements.

Firebase Genkit

Framework designed to help build AI-powered applications & features on Node.js & Go

Unified API for AI generation

Use one API to generate or stream content from various AI Models include, but not limited to, OpenAI, Gemini, Claude, etc.

Structured Generation

GEnerate or stream structured objects with build-in validation. Simplify integration with your app and concert unstructured data into usable format.

Tool calling

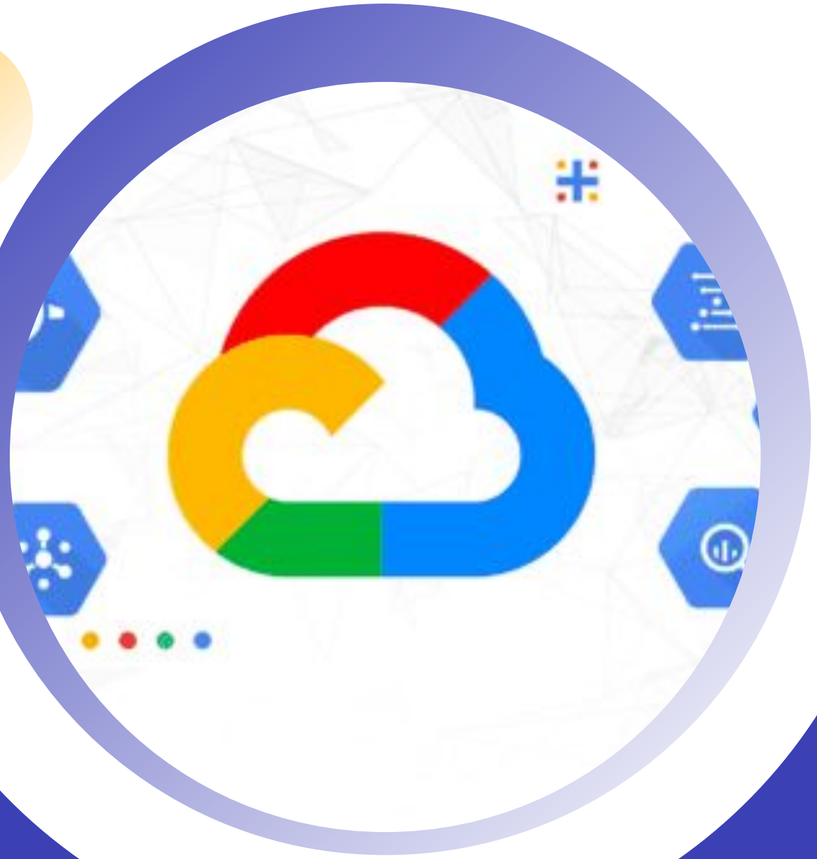
Allow AI models to call functions and APIs as tools to complete tasks. Models decide when and which tool to use.

Retrieval-augmented generation (RAG)

Improve accuracy & relevance of generated output by integrating your data.

Prompt templating

Create effective prompts that include rich text templating, model settings, multimodal support, and tool integration within a compact runnable prompt file.



Other Google Cloud AI Services

Document AI solution

Portfolio to help customers implement intelligent document processing solutions for their business - available with both API and UI

Differentiated with generative AI

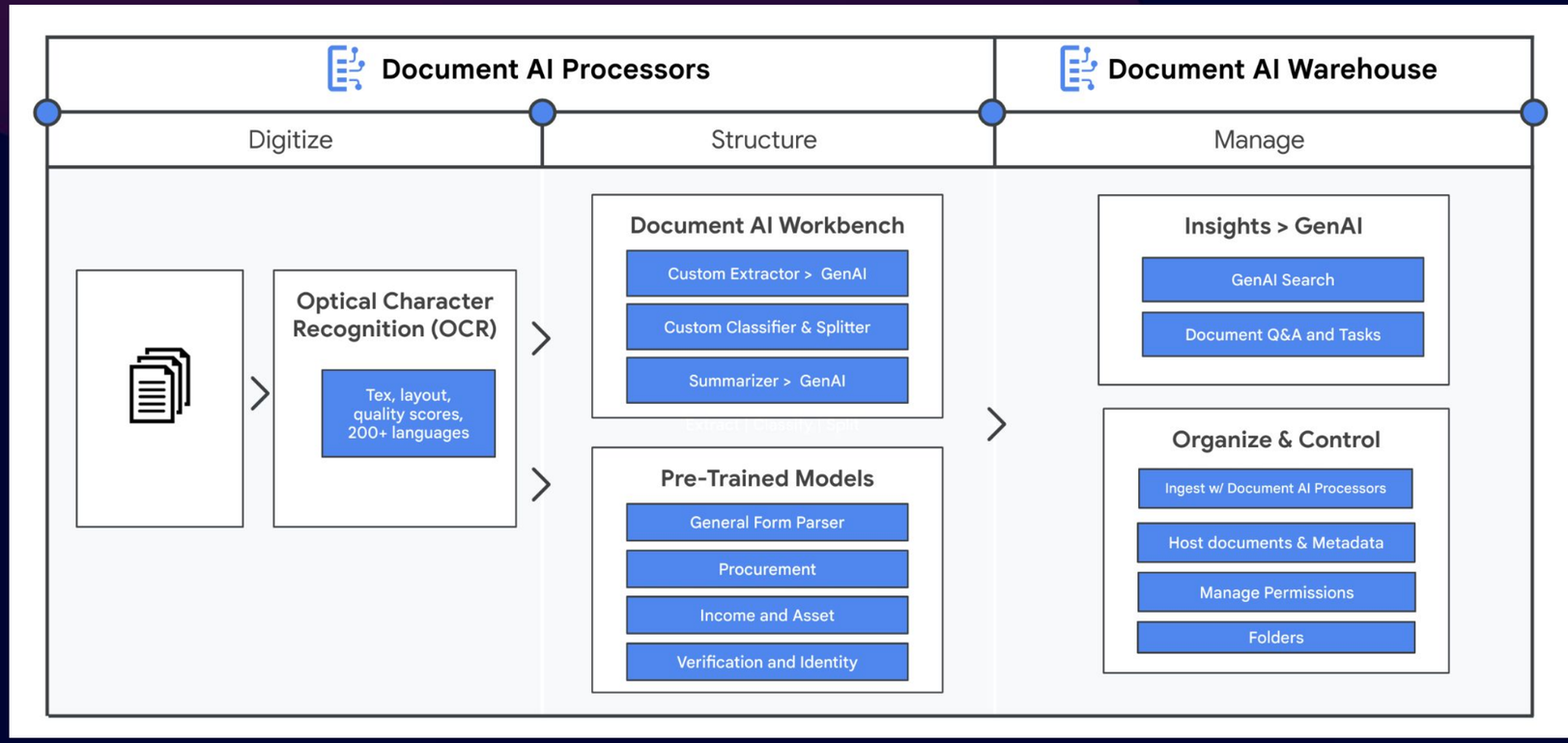
- Zero-shot extraction
- Document Insights and tasks

End-to-end solution

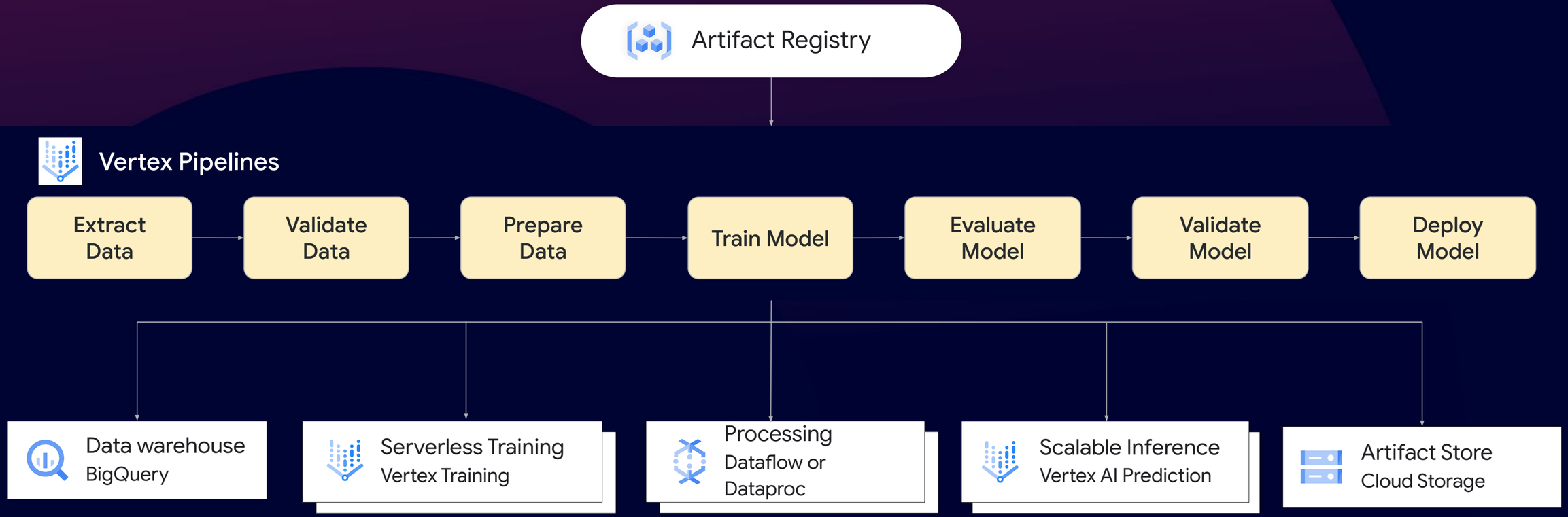
- Pre-trained, customizable, and generative AI
- OCR supports over 200+ languages
- Document management and insights with Warehouse

Driving business value

- Automate document processing
- Lower manual reviews
- Easy path to customization



Simplify MLOps with Vertex AI Pipelines



Technology: Google Cloud

Pre-built components

The Google Cloud Pipeline Components (GCPC) SDK provides a set of prebuilt Kubeflow Pipelines components that are production quality, performant, and easy to use.

Load and share an ecosystem of components

```
from kfp import components

loaded_comp = components.load_component_from_file('component.yaml')
```

```
loaded_comp =
components.load_component_from_url('https://raw.githubusercontent.com..'
```

Simplify MLOps with Vertex AI Pipelines

Easy-to-use Python SDKs: Build your Pipelines using the battle-tested and easy-to-use KFP SDK and TFX SDK

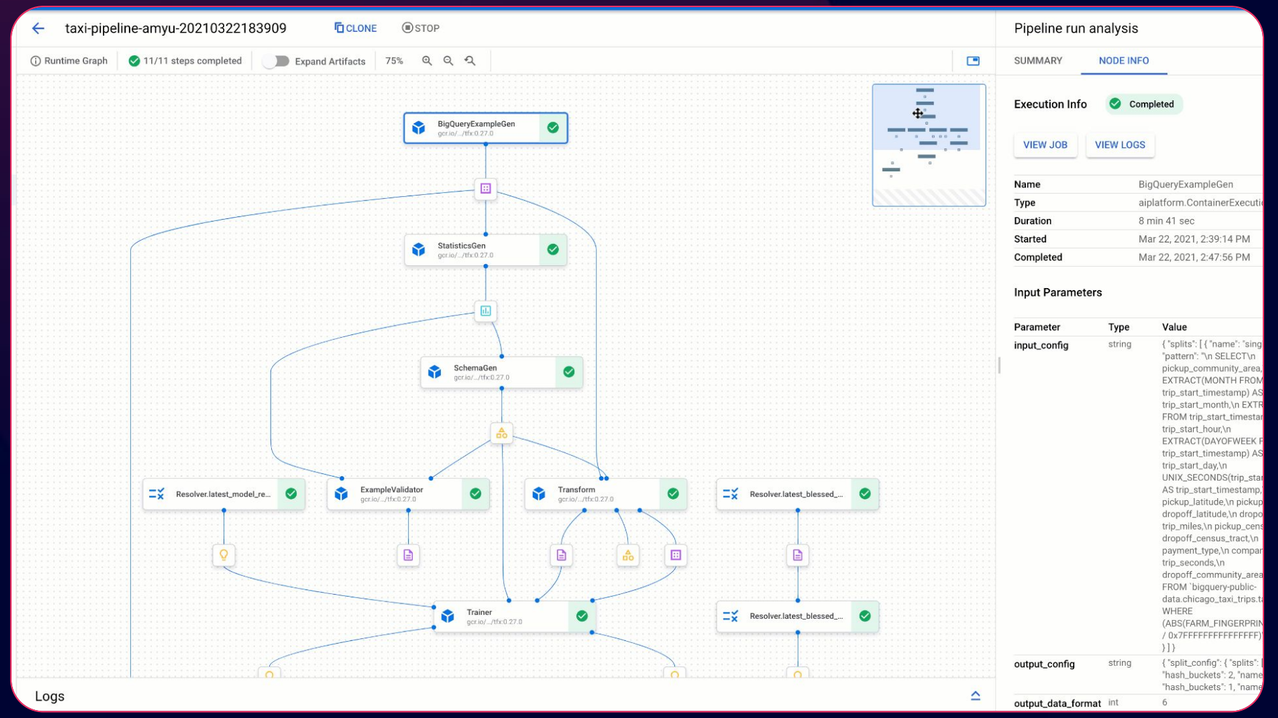
Scalable: Run as many pipelines on as much data as you want without having to worry about compute resources

Cost-effective: Pay for the pipelines you run and the resources they use.

Secure: Integrated with Google Cloud security features like IAM, VPC Service Controls, and CMEK.

Metadata Tracking and Experiments: Automatically store metadata about every artifact and experiment produced.

Portable: Your pipelines can be ported anywhere where Kubernetes runs with Kubeflow Pipelines.



Data Storage, Migration & Processing Tactics

Preview

Simplify data movement by leveraging various Google Cloud product & services

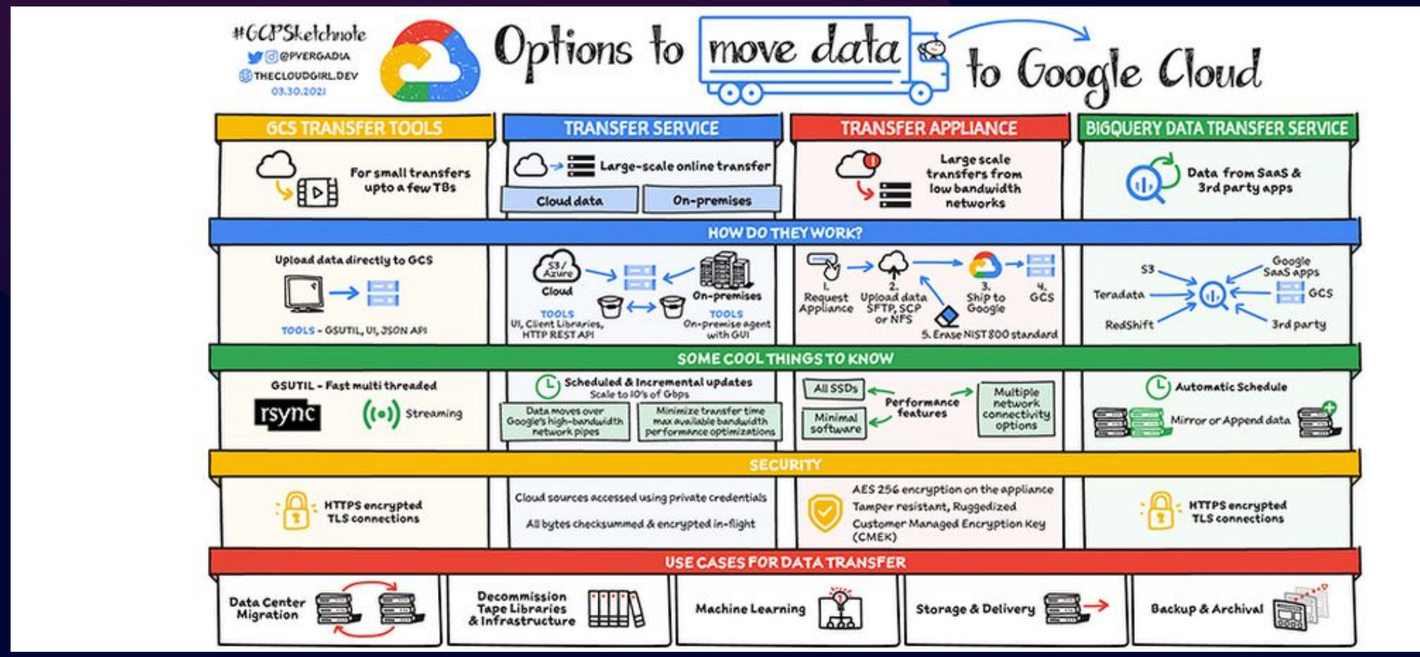
BigQuery Data Transfer Service: Simplifies and automates data movement to BigQuery

Datastream: Serverless platform that supports change streams from databases such as Oracle, MySQL, Postgres into Google Cloud databases

Cloud Storage as mounted file system: Leverage Cloud Storage FUSE to stream data from a Cloud Storage bucket(s) to your instance.

BigQuery Omni: Run BigQuery analytics on data stored in Amazon Simple Storage Service (Amazon S3) or Azure Blob Storage using BigLake tables.

Storage Transfer Service: Automate data transfer to/from/between object and file storage systems, including Google Cloud Storage, Amazon S3, Azure Storage, on-premises data, and more



Enhance LLM responses with RAG architecture

combining this extra knowledge with its own language skills, the AI can write text that is more accurate, up-to-date, and relevant to your specific needs.

Access to updated information

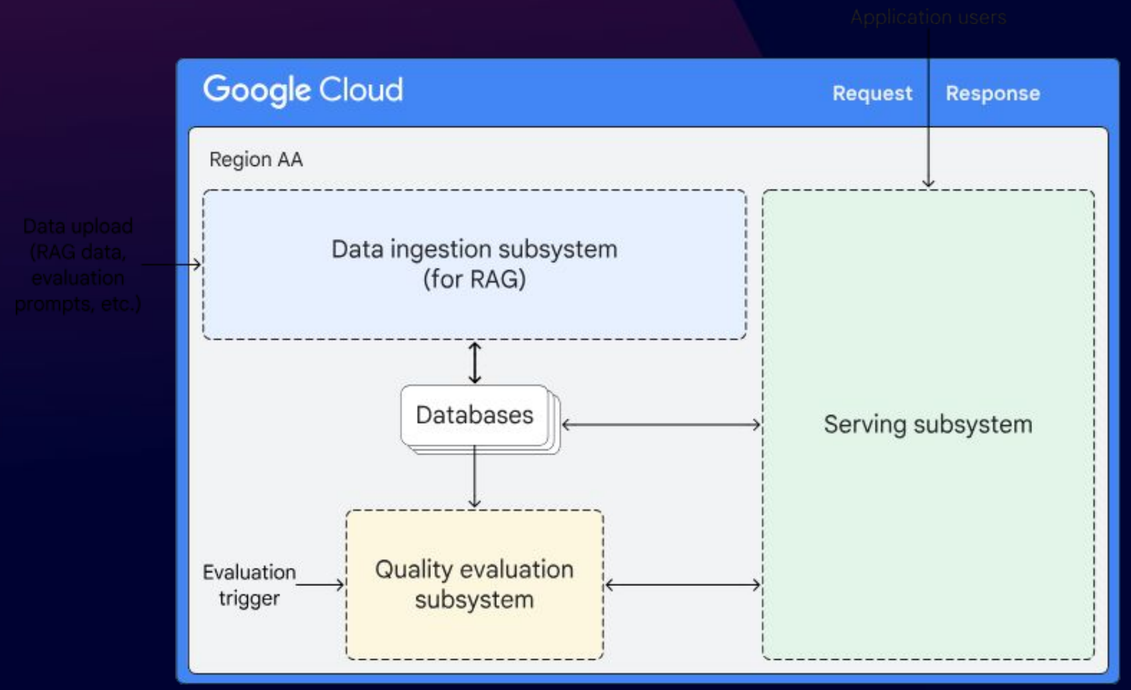
Traditional LLMs are often limited to their pre-trained knowledge and data. This could lead to potentially outdated or inaccurate responses. RAG overcomes this by granting LLMs access to external information sources, ensuring accurate and up-to-date answers.

Factual grounding

helps address factual inaccuracy issue by providing LLMs with access to a curated knowledge base and may also assist in preventing hallucinations being sent to the end user.

Contextual relevance

This contextual grounding helps to reduce the generation of irrelevant or off-topic responses.



Vertex Reasoning Engine

Preview

A managed runtime for your customized agentic workflows in generative AI applications

Customizable:

By utilizing LangChain's standardized interfaces, LangChain on Vertex AI can be adopted to build different kinds of applications

Simplified deployment:

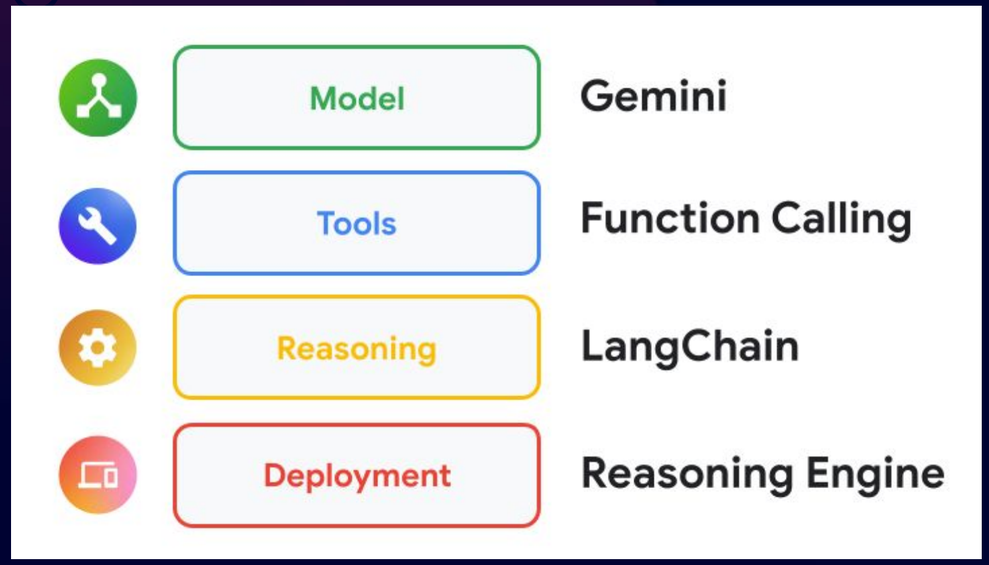
Leverages same APIs as LangChain & simplifies deployment via single click deployment

Integration with Vertex AI ecosystems:

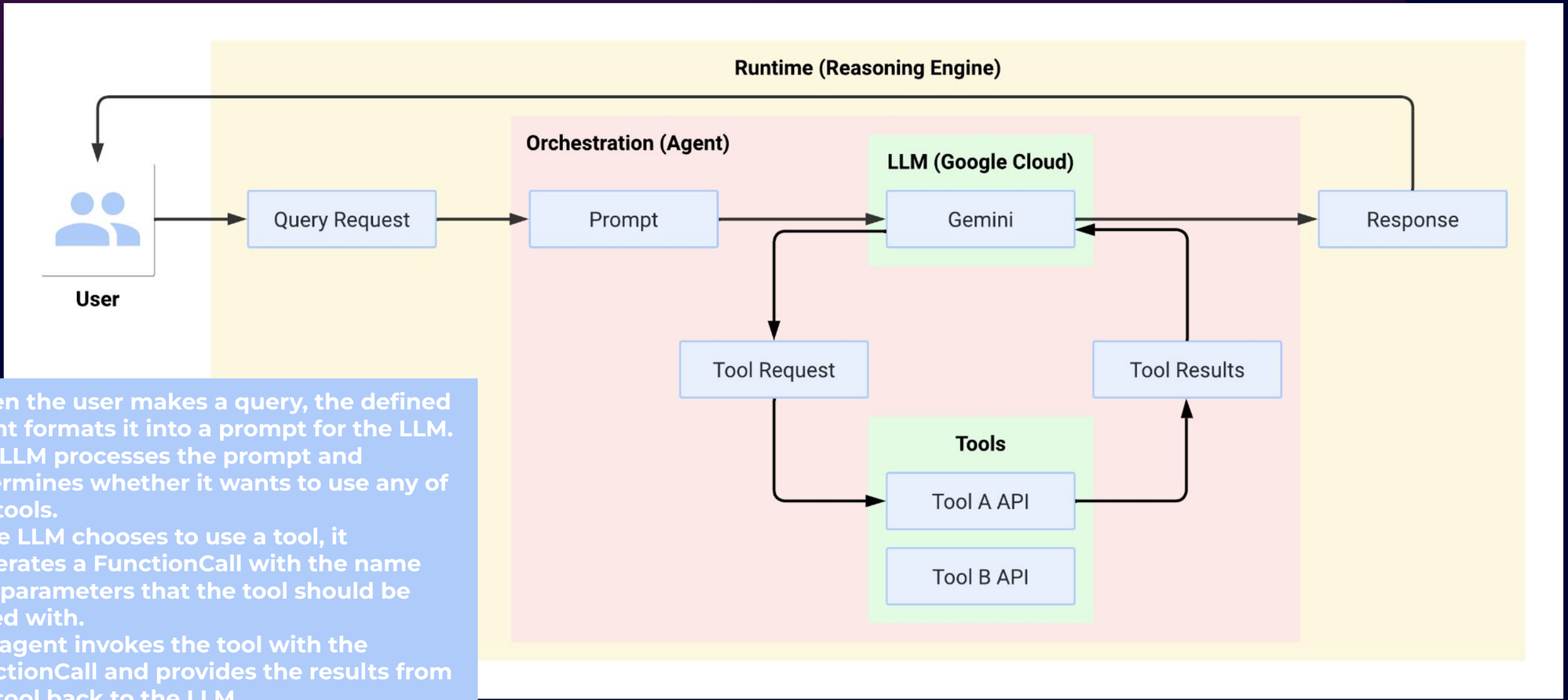
Uses Vertex AI's infrastructure and prebuilt containers to help deploy LLM applications. API can be to integrate with Gemini models, Function Calling, and Extensions.

Secure, private & scalable:

Allows you to use a single SDK call instead of managing the development process on your own.
Frees you from tasks such as application server development, container creation, and configuration of authentication, IAM, and scaling.
Vertex AI handles autoscaling, regional expansion, and container vulnerabilities.



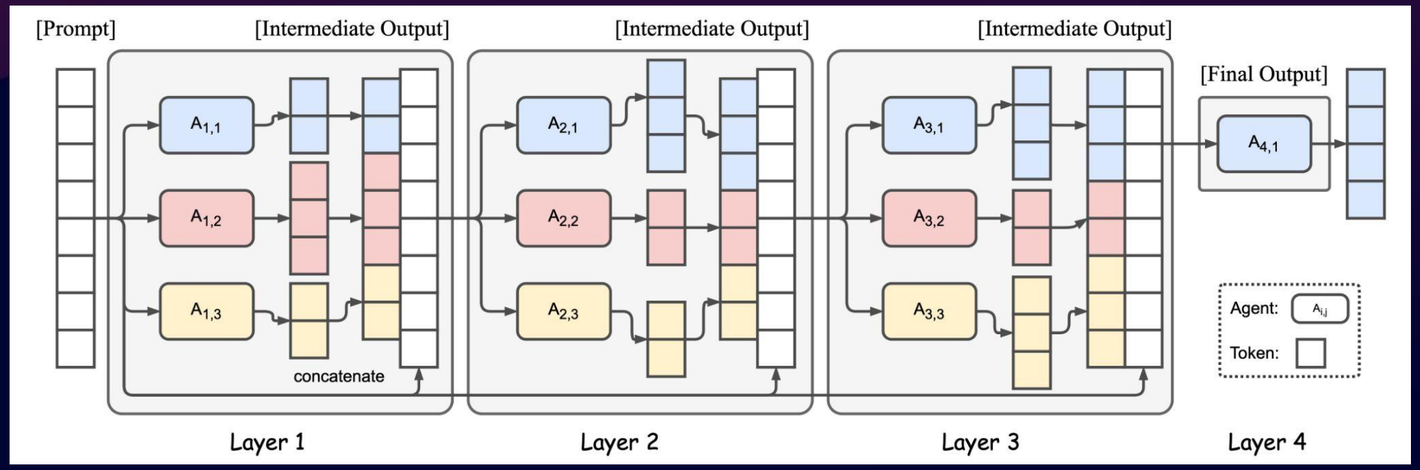
Reasoning Engine System Flow

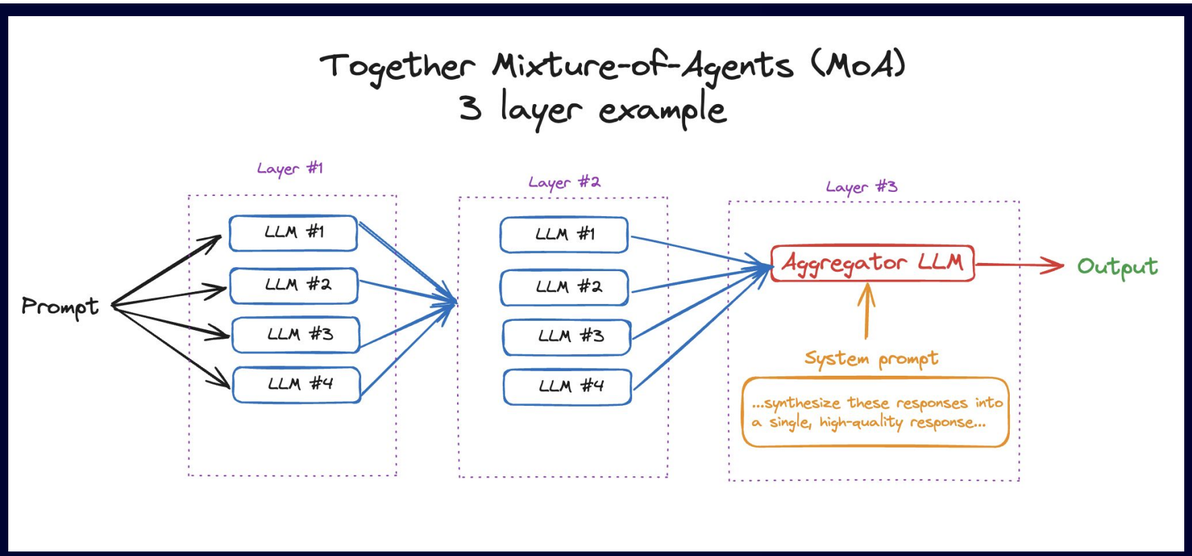


1. When the user makes a query, the defined agent formats it into a prompt for the LLM.
2. The LLM processes the prompt and determines whether it wants to use any of the tools.
3. If the LLM chooses to use a tool, it generates a FunctionCall with the name and parameters that the tool should be called with.
4. The agent invokes the tool with the FunctionCall and provides the results from the tool back to the LLM.
5. If the LLM chooses not to use any tools, it generates content that is relayed by the agent back to the user.

Mixture of Agents via Reasoning Engine

leverage the collective strengths of multiple LLMs to enhance performance, achieving state-of-the-art results by employing a layered architecture where each layer comprises several LLM agents





Section break

Thank you

Resources

Jupyter Notebooks

- Getting Started with Gemini (Python)
- Getting Started with Gemini
- Vertex AI Python SDK Gemini Docs
- Getting Started with BigQuery Datasets
- BigQuery Local Training via PySpark
- Get Started with Vertex AI Training LightGBM
- TensorFlow Serving Functions with Raw Predictions
- Workbench Executor
- Gemini API Prompting Quickstart
- Anomaly Detection in Security Logs with BQML
- Intro to Kubeflow
- End-to-End Journey for Each Model
- Mixture of Agents
- Google Maps API Agent via Reasoning Engine
- Tutorial on Anomaly Detection via Autoencoders

Additional Resources:

- Hugging Face - Repository for Building Machine Learning Models
- Papers with Code - State-of-the-Art Research Papers on Machine Learning & Datasets
- Octopus-v4 by Nexus AI: Demonstration of an MoA

Clone Repositories:

git clone https://github.com/GoogleCloudPlatform/vertex-ai-samples.git

git clone https://github.com/google-gemini/cookbook.git