Driving Innovation with Generative AI

Built with Pega and Google Cloud

Uri Mariash
Key Account Executive
Google Cloud

Hakim Graia
Principal Cloud Architect
Google Cloud

Rob Smart
Principal Solutions Consultant
Pega

Building Blocks

Vertex AI

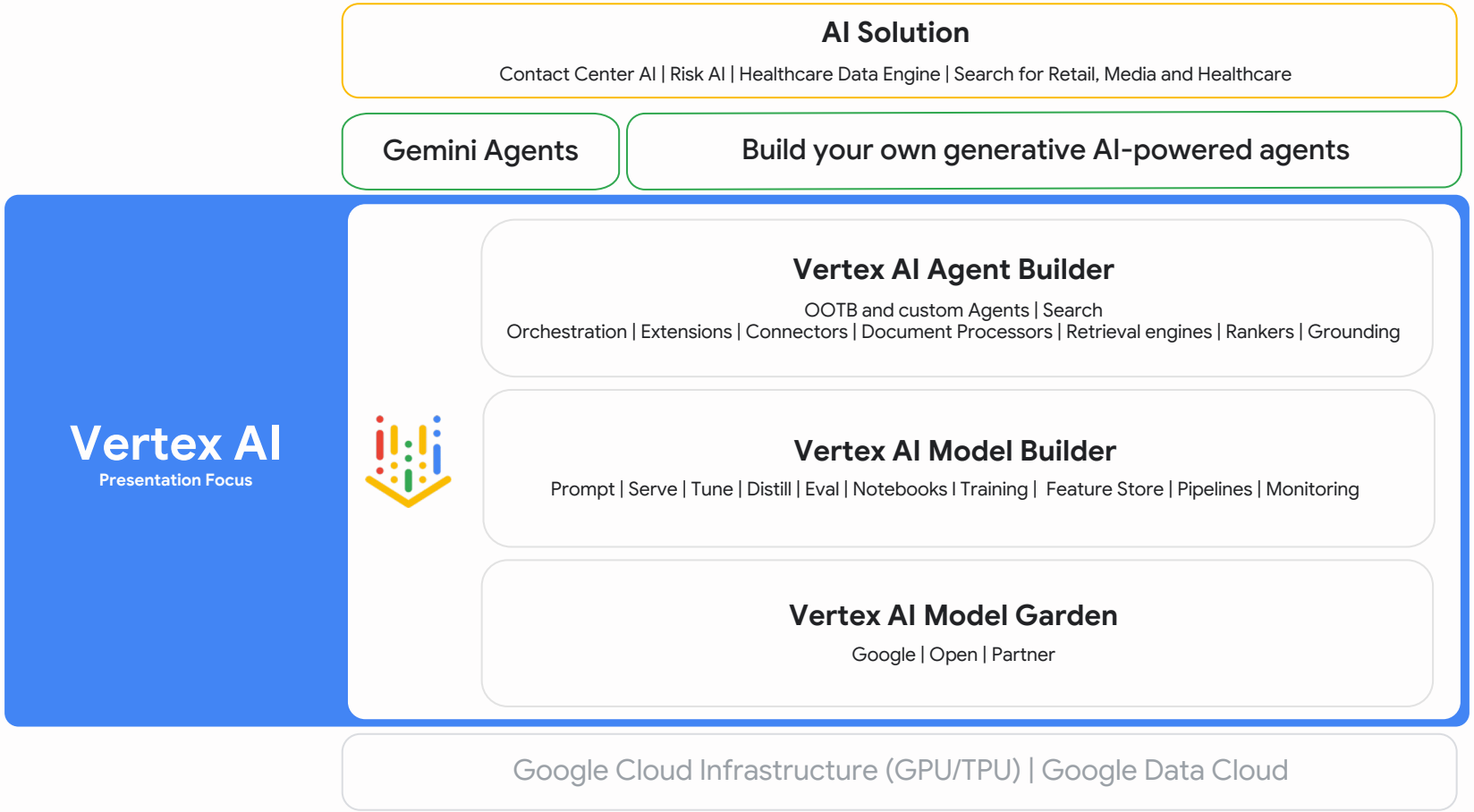
Enterprise-ready generative AI for builders

Best models from Google and the industry

End-to-end model building platform with choice at every level

Develop and deploy agents faster, grounded in your enterprise truth

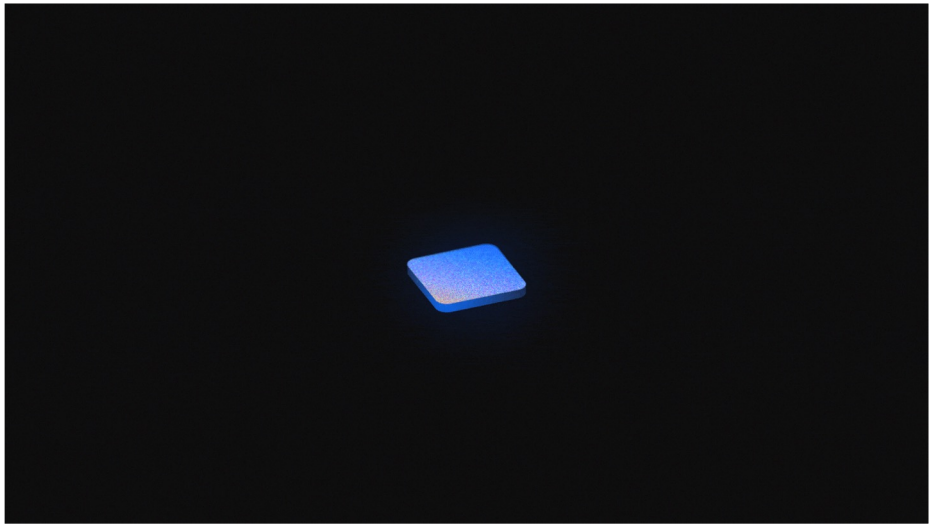
Built on a foundation of enterprise readiness


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

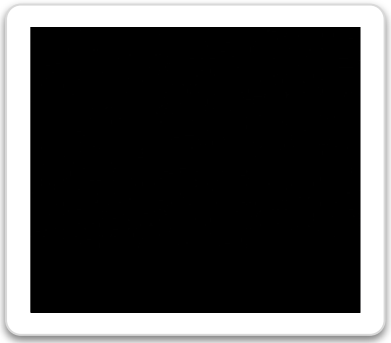
The next chapter of Generative AI innovation

Gemini

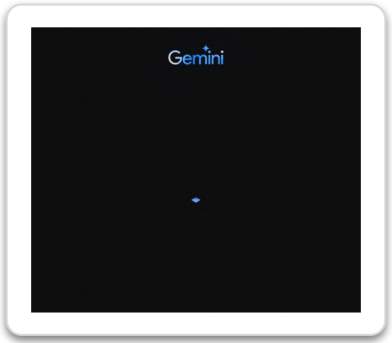
Gemini is the most capable and general model we've ever built, and is the result of a large-scale collaborative effort by teams across Google, including Google DeepMind and Google Research.



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

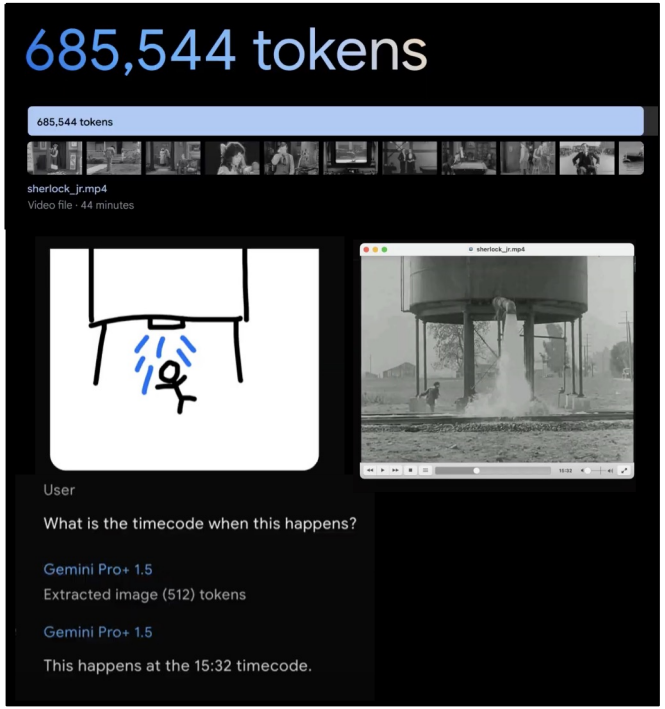
Built with responsibility and safety at the core

Gemini 1.5 Pro

Mid-size multimodal model with breakthrough long-context understanding

Gemini 1.5 Pro delivers dramatically enhanced performance and represents a step change in our foundation model approach, including:

- A **new Mixture–of-Experts (MoE) architecture** that provides more efficient training and serving, while increasing model performance
- An **expanded context window** (up to 1 million tokens) for complex reasoning across vast amounts of information
- **Better understanding and reasoning across modalities** including text, code, image, audio and video
- **Extensive ethics and safety testing** that builds on novel research on safety risks and leverages red-teaming techniques to test for a range of potential harms



Vertex AI

- Unified development and deployment platform for data science and machine learning
- Increase productivity of data scientists and ML engineers

No code / low code workflow

Data Science tool kit

Custom workflow

MLOps

Model Garden

Pre-built/ Pre-trained Models

- Vision
- Video
- Language
- Speech
- Translation
- Tables
- Forecast
- TabNet

BigQuery

- BigQuery ML

Vertex AI Studio

Workbench

Integration with Data Cloud

- Cloud Storage
- BigQuery
- Spark
- Spanner
- AlloyDB
- BI

Experiment

- Datasets
- SDK
- Experiments

Train

- Training
- NAS
- Vizier

Deploy

- Prediction
- Vector Store
- Streaming Ingestion

Model Monitoring

Explainability

ML Metadata

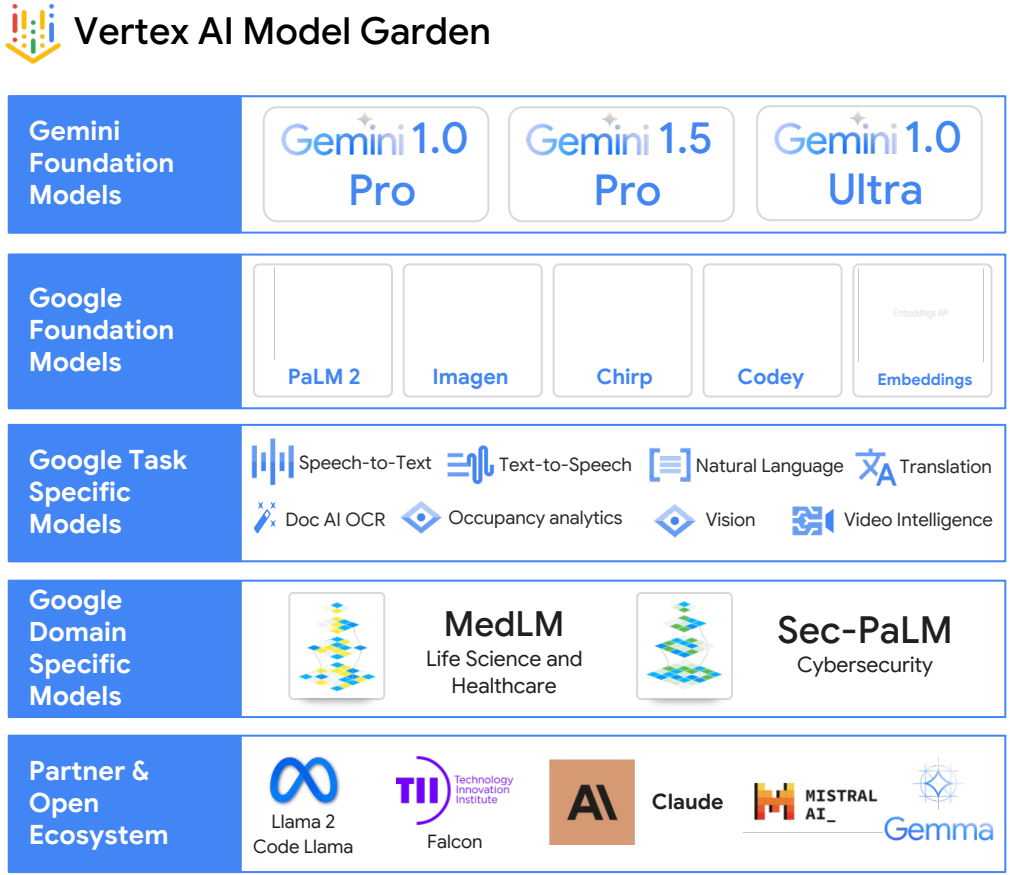
Feature Store

Model Registry

Model Evaluation

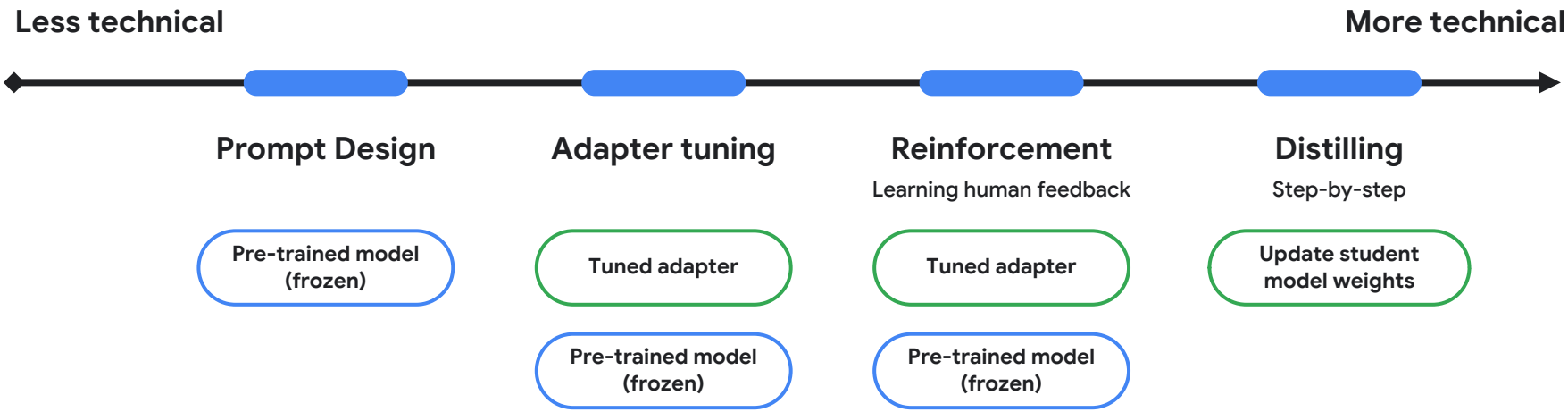
Pipelines

130+ enterprise-ready foundation models in Vertex AI Model Garden



- **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**

How to customize an LLM



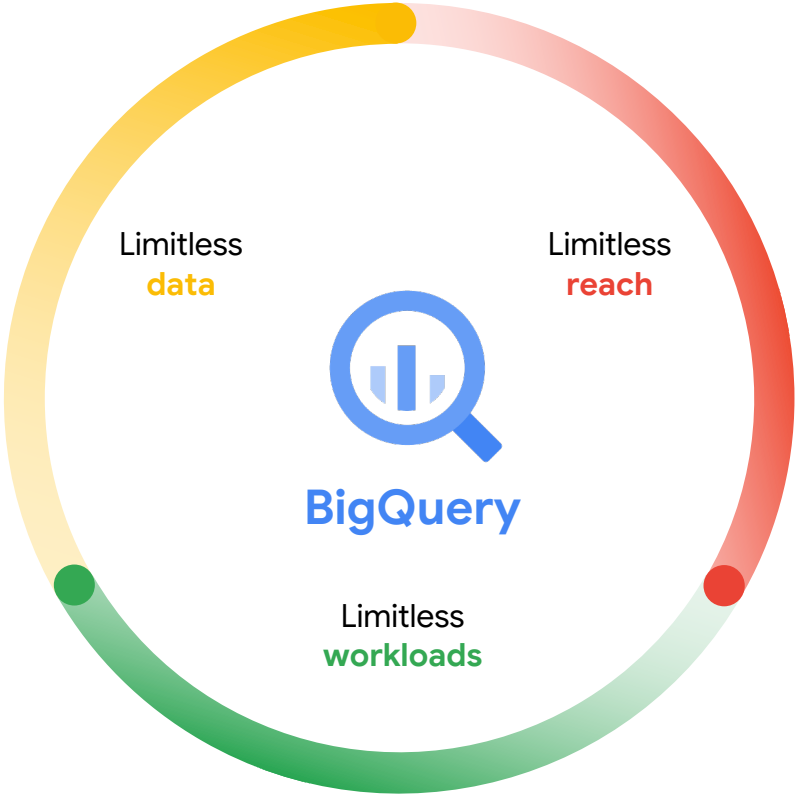
Vertex AI Agent Builder

Powering multiple use cases

- Food service ordering
- Market research support
- Customer service
- Travel assistance & booking
- Virtual banking assistants
- Personalized conversations
- Website search
- Document/media search, synthesis

BigQuery

The core of Google's Data Cloud to power your data-driven innovation.



100k+ data professionals have started their Data Cloud journey using BigQuery with trials growing nearly **150% YoY in 2021**.

Vertex AI Feature Store

Completely reimagined around **BigQuery** with advanced feature management and industry leading **low latency** online serving for predictive and **generative AI** workloads

Built on BigQuery

No more unnecessary data copying

Powerful Serving

Two options for real time serving optimized for latency or data size

Ready for Generative AI

Manage and serve embeddings with built-in high performance ANN retrieval

Document AI

JPEG, PNG, GIF, BMP, WEBP

Doc Prep

- Splitter
- Quality
- OCR
- Language
- Translation

Classification

- Specialized Classifier
- Custom Classifier

Entity Extraction

- Form Parser
- Specialized Schemas
- Custom Entity Extraction

Entity Enrichment

- Normalization
- Enterprise Knowledge Graph (EKG)

Training

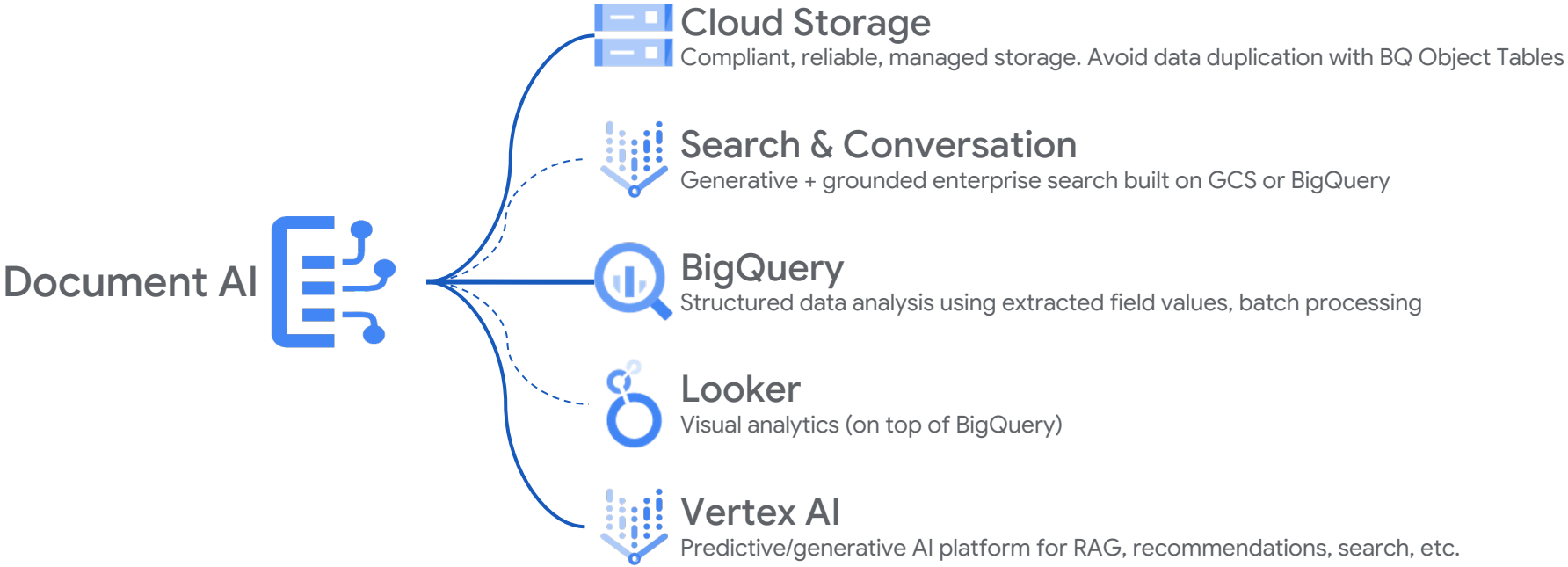
- Evaluation
- Datasets
- Processor Versions

Repository

- Warehouse

Human Review

Easily adopted with complementary products



Pega Infinity 8.8

Cloud Storage

Used to store case, artifact repositories in deployment pipelines, staging locations for data flows and storage for platform logs.

Offers support for Google Cloud Storage as a repository in platform

- **Ability to create a repository of type Google Cloud Storage**
- **Offers authentication via OAuth2.0**
- **Ability to provision OOTB repository instances for Pega cloud hosted on Google Cloud Platform**

PEGA

These Google Cloud services are configured in the Pega Platform using low-code configuration then easily referenced in decision strategies or case workflows. Pega can be offered aaS via Pega Cloud, hosted in a Google Cloud Project leveraging compute, storage and database services.

Pega Infinity 24.1

Pub/Sub

Used to publish and consume messages that are related to events within a cases. Also used for updating data cached in Pega for realtime decisioning.

Use Google Pub/Sub for streamlined and real-time data ingestion into Pega

Pega users can consume data from Google Pub/Sub, in addition to different real-time data sources such as Kafka, Kinesis, or Pega Stream

In this first release of Google Pub/Sub support, Pega supports ingesting from a publishing source via Google Pub/Sub for consumption in Pega, allowing users to ingest data such as customer data and customer-associated data like account data or product holdings

In essence, Google Pub/Sub is used for the real-time integration from any external source and serves as a middleware messaging layer from outside publishers into Pega

PEGA

These Google Cloud services are configured in the Pega Platform using low-code configuration then easily referenced in decision strategies or case workflows. Pega can be offered aaS via Pega Cloud, hosted in a Google Cloud Project leveraging compute, storage and database services.

Pega Infinity 24.1

Vertex AI

Used to supplement the adaptive AI models curated in Pega. AI predictions are used to proactively adapt the case process and aid decisions.

Leverage Predictive Models in Google Vertex AI

- **AutoML, XGBoost or scikit-learn models**
- **Replaces Google AI Platform support as it will be discontinued**
- **NLP topic detection models in Google Vertex AI**
- **Replaces the deprecated AutoML Natural Language**

Applies to Predictive AI; for generative AI support see Pega GenAI™ features

PEGA

These Google Cloud services are configured in the Pega Platform using low-code configuration then easily referenced in decision strategies or case workflows. Pega can be offered aaS via Pega Cloud, hosted in a Google Cloud Project leveraging compute, storage and database services.

Roadmap

BigQuery

To be used as a data warehouse to reference tables and views which can be referenced in data flows and elsewhere in the platform.

Bigtable

To be used for as a primary source of data in real time NBAs and other low-latency required data retrieval.

Gemini

Gemini

Alternative large lanbguage model used for Pega GenAI use cases (such as Knowledge Buddy and Coach).

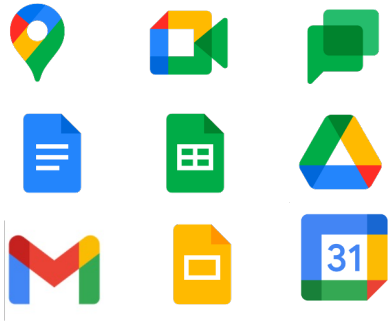
PEGA

These Google Cloud services are configured in the Pega Platform using low-code configuration then easily referenced in decision strategies or case workflows. Pega can be offered aaS via Pega Cloud, hosted in a Google Cloud Project leveraging compute, storage and database services.

Integrations

Google Data Sources

Google Workspace

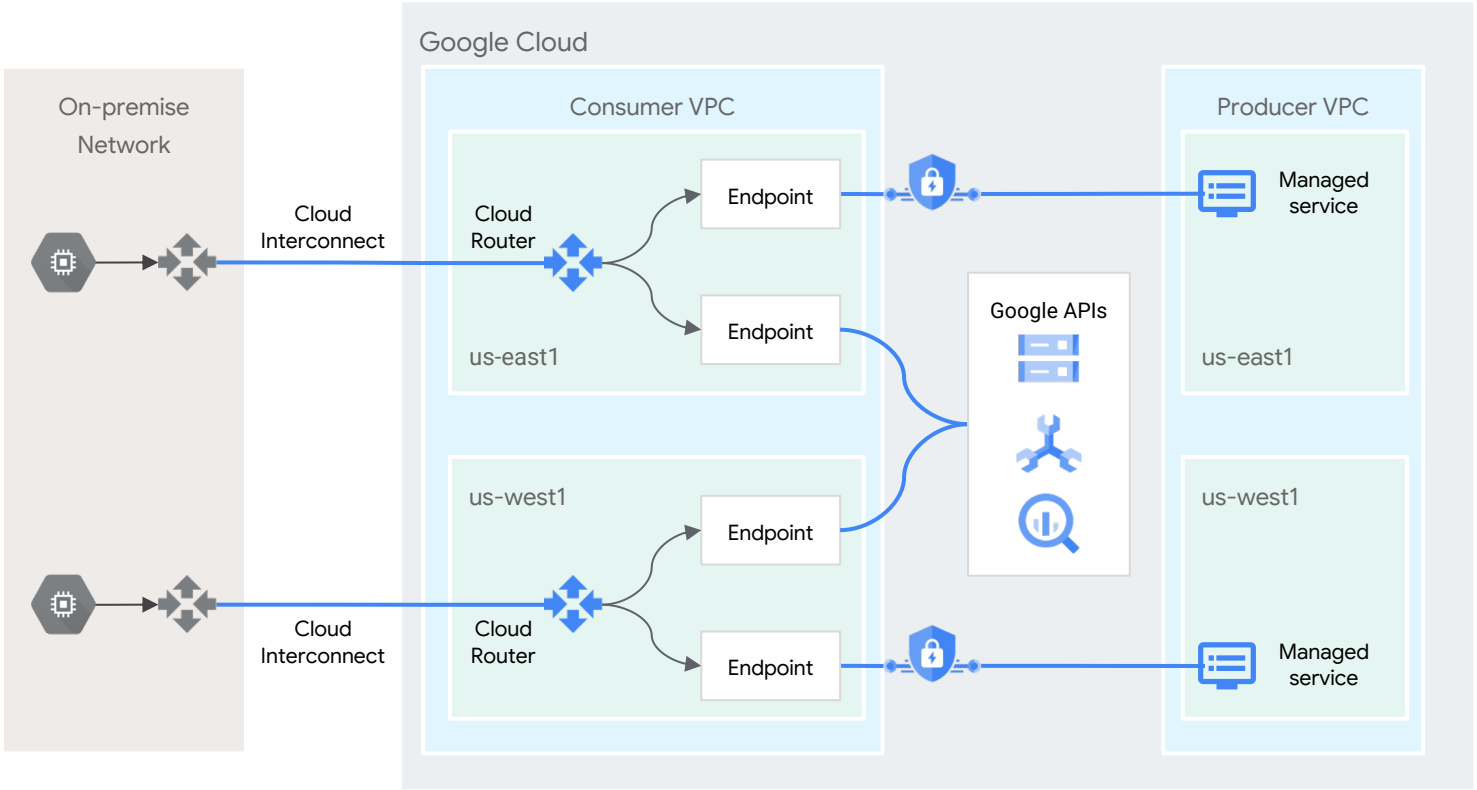


And many more...

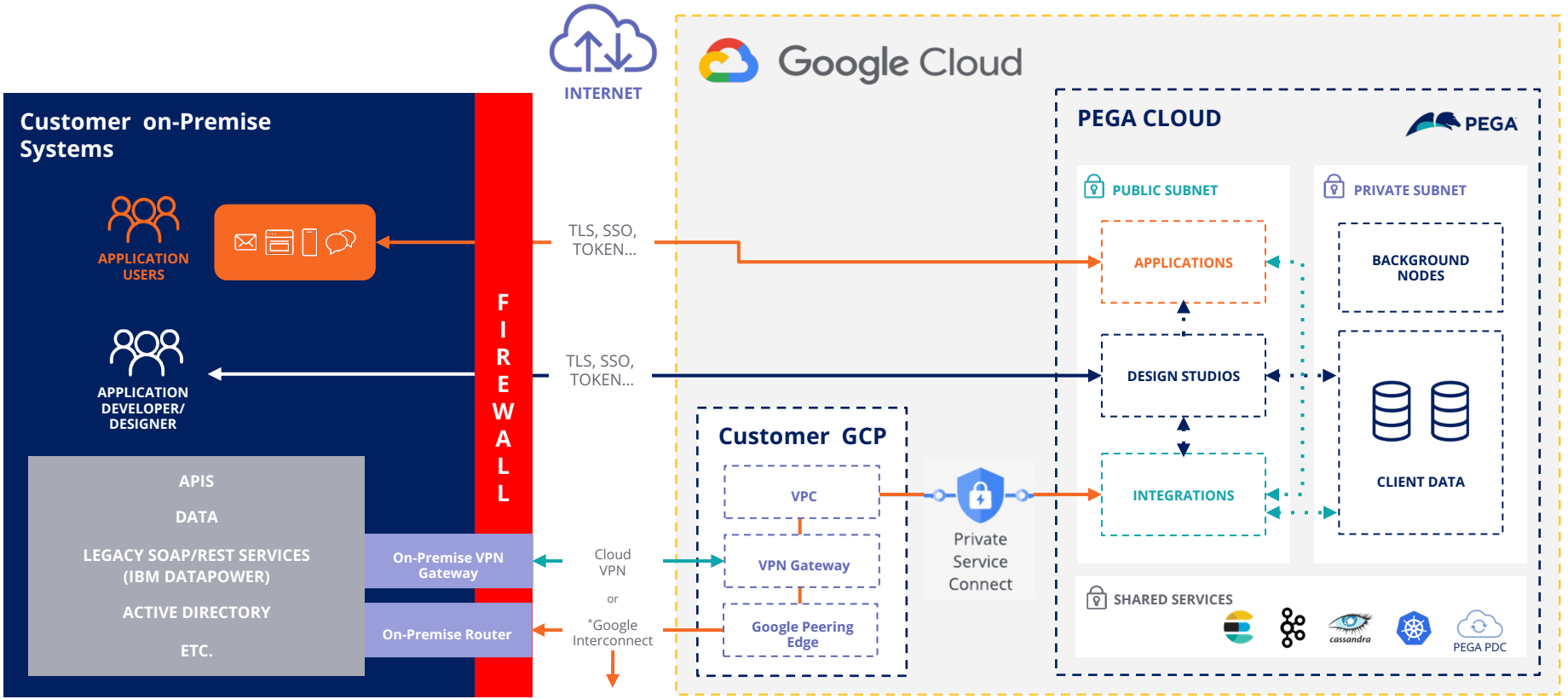
PEGA

Google Cloud

Pega Cloud Connectivity : Private Service Connect (PSC)



Pega Cloud Connectivity



Google Cloud & Pega
Innovating together

Top Generative AI powered use cases

Public Website Navigation

Effectively find information from a website via multi-modal inputs and conversational queries

Product / Content Catalog Discovery

Effectively find the most relevant Products / Content listings from a inventory catalogue

Intra-Knowledge Q & A

Conversationally query questions for answers from internal knowledge sources

Business Process Automation

Automating the information retrieval and recommendation step of a recurring business process

Regulatory Compliance

Interpret regulatory policy / documents to identify potential violations relative to operating procedures

Documentation Generation

Write new documentation based on summarization of other documents & software code

Customer Service Automation

Effectively service customers requests for information and service provisioning

Product / Content Recommendation

Recommend personalized Product / Content / Next Best Action from a catalogue

Document Search & Synthesis

Effectively find the most relevant documents and summarize their contents

Creative Assistance

Empower creative teams to create bespoke images and creative content for campaigns and editorial content

Research Acceleration

Find complex subject domain information across many disparate sources and synthesis the findings

Developer Efficiency

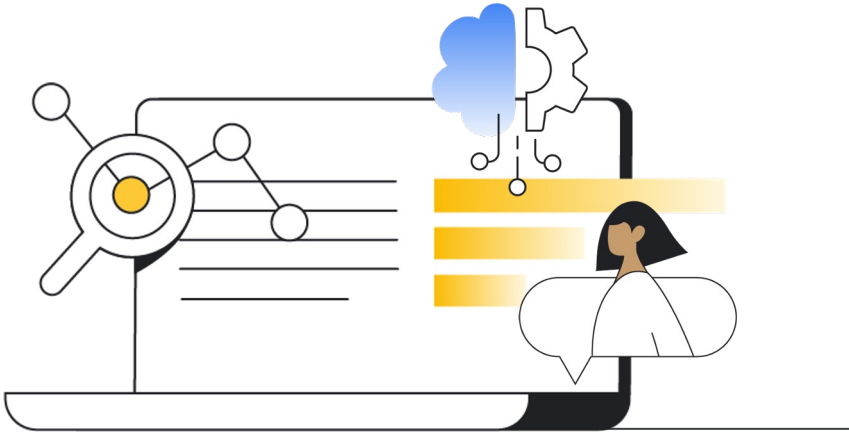
Complete and augment code to make your engineering team more efficient and effective

Complex data, intuitively accessible

Online interactions made conversational

Content generation at the click of a button

Realtime customer engagement



PEGA

Google Cloud

Use case: Generative AI - Hyper-personalized customer experience

SUMMARY

CSPs are facing immense downward pressure on ARPU, this is further leading to flat revenue growth & EBDITA margins. Return on capital invested is declining consistently. CSPs are spending to acquire and retain customers, marketing & sales activities, meeting the demands for high-speed connectivity and spending to upgrade Network Capacity and IT Infrastructure. Create a Gen AI-based solution engineered to provide customers a personalized real-time omnichannel experience driven by a Next-Best-Action central brain

VALUE

- Increase lifetime value by up to 35%
- Increase revenue by up to 24%
- Reduce churn by 20%

PEGA

Google Cloud

Delivering a personalized, real-time end-to-end customer experience

Illustrative customer journey with Pega Customer Decision Hub (CDH) and Google Cloud data and AI/ML capabilities

Meet Matt! Matt has been a loyal customer for the past 6 years

Recently, Matt's internet service quality has dropped significantly. Disappointed with the service quality, he has started to look for alternative service providers, as well as browse the Terms & Conditions section of the service providers website.

1. Google Cloud collects streaming network events and applies ML algorithms to identify that Matt has a high risk of Churning to an alternative provider.
2. Website clickstream events are sent automatically to Google Cloud, and the browsing of the T&Cs is detected.
3. Pega is informed by the events via Pub/Sub, and starts to automatically compile the Next Best Action to avoid customer churn, using the churn prediction from Google. Using customer profile data Pega identifies that Matt is a movie lover.
4. Pega calculates the Next Best Action in real time, scoring & ranking all the options. The offer is for a three-month free trail of a video streaming service. GenAI is used to create a curated message that will resonate with Matt and is empathetic to his situation

Matt receives a notification about the free video streaming offer together with an apology for the service interruptions, and an assuring message that the problems are being addressed. The offer contains products and services that are targeted specifically for Matt, so he feels valued and understood by his provider, which leads him to hold off on looking for a new supplier.

5. Matt is sent a follow up message with tips to ensure that his service maintains its maximum speed – GenAI is used to create a curated message in a tone that suits Matt's Tech Savvy profile. Matt's service setting are adjusted to get the best performance and he is sent a message thanking him for his loyalty
6. AI models are re-scored, and re-ranked by Pega, using data in Big Query and models in VertexAi.

   As Matt still has a high churn risk, he is sent a second offer of a proactive speed monitoring service.

Matt notices the improved service stability and speed and accepts the proactive monitoring service quickly & easily with a response to the text.

Matt receives proactive personalised, empathetic service and is secured as a happy customer, which reduces his churn score.

Matt contemplated leaving the provider, but having had his problem identified and fixed in a timely manner, and provided with an offer that made him feel valued as a loyal customer, he decides to stay with his provider, securing ARPU and creating opportunities for up-sell.

Proactive retention trigger/alert → Customer sentiment analysis → Next-best offer/action compilation → Customer retention and conversion

Data collected and transformed with Google Cloud detecting relevant actions/alerts and informing Pega

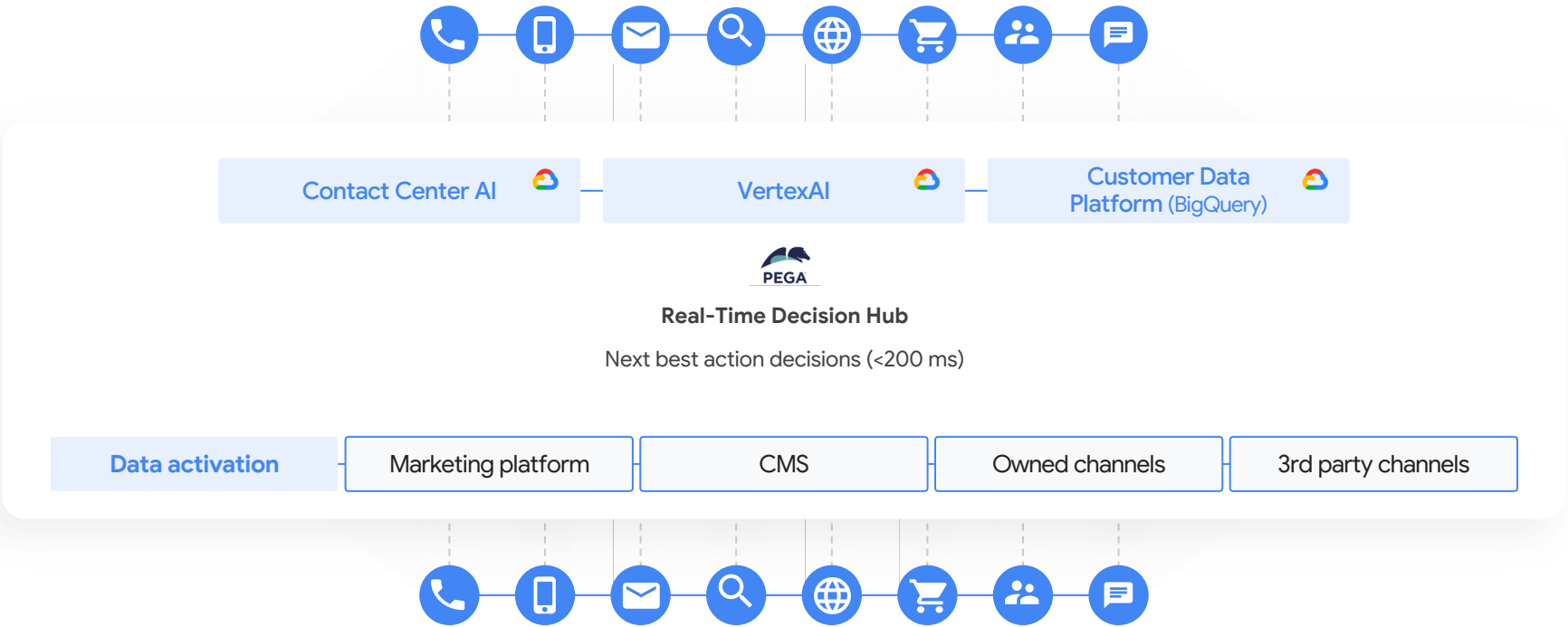
Leverage AI and ML-enriched insights to enhance Pega's data-driven decisioning

Engage customers with curated GenAI powered actions to deliver targeted next best action/offer

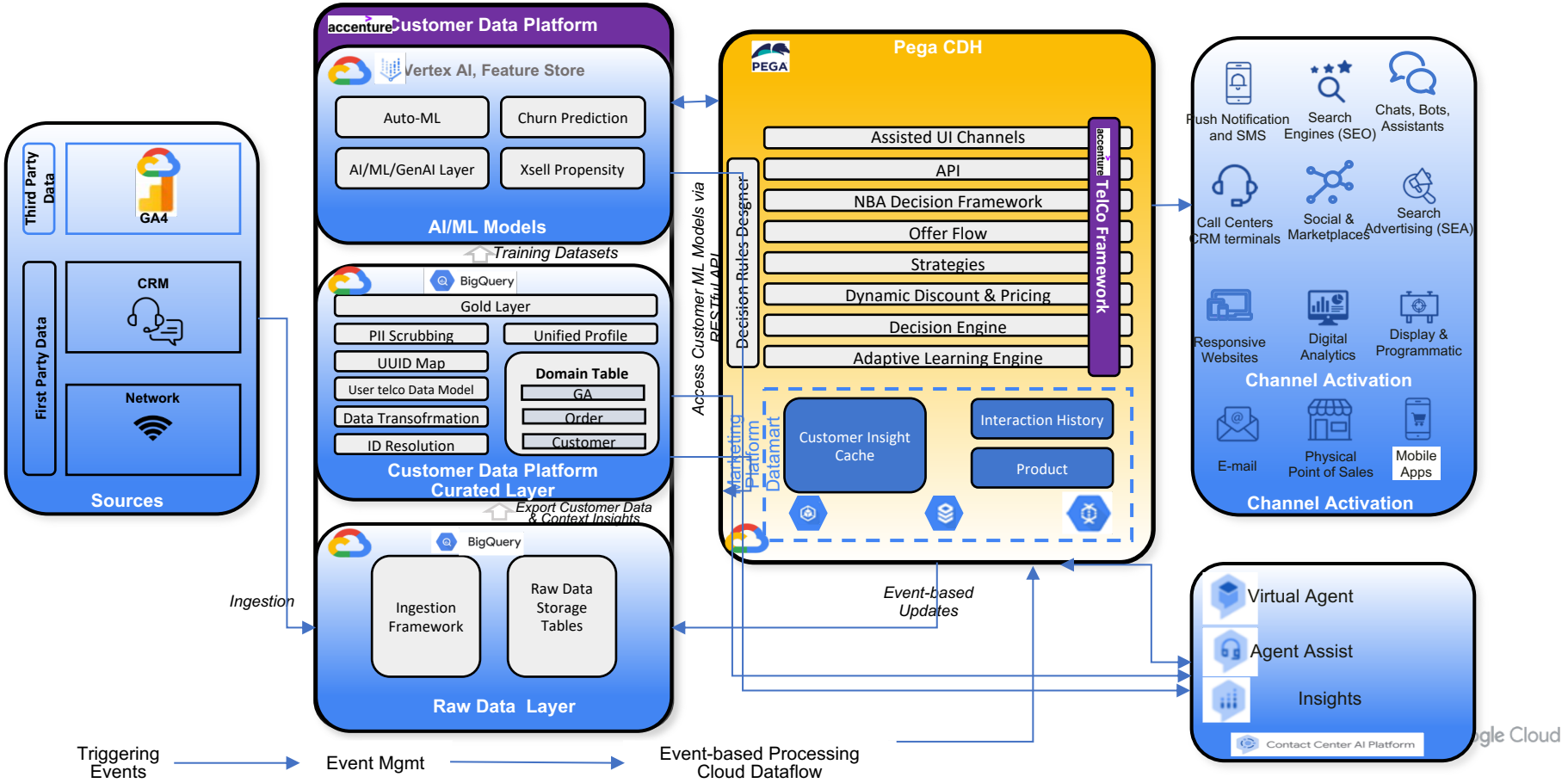
PEGA

Google Cloud

Pega & Google Cloud unlock a modern, fully integrated customer engagement stack



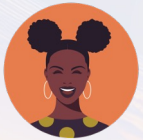
Reference architecture



Act I - From Modem Freeze to Fiber Upgrade: AI-Powered Troubleshooting and Self-Service Installation

Start

Ann's WIFI experienced degradation and Routing freeze



Ann

Proactive automated action based on customer profile , modem rebooted. Ann notified

After querying **CDH**, the workflow confirmed Ann's eligibility for a Fibre upgrade with unlimited mobile hotspotting until installation, which she accepted

The issue persisted, Ann engaged in a multi-modal troubleshooting session with **Gemini**, which ultimately resulted in a workflow to order a new modem

Mei

The self-service workflow confirmed no permits were needed, the nearest fibre node had capacity and equipment was available. Ann then booked an installation appointment within the next two days

Khalil

Developer

Automated subflows sent a pre-installation reminder with a rescheduling option. On installation day, the installer received a list of required materials and devices, including those for Ann's home. The fibre connection and modems were then exchanged. Upon updating their mobile device, the installer automatically updated Ann's account to reflect the new modem and billing plan

Success

Customer

CSP Customer Service Representative

CSP SWE

Powered by

Gemini

Vertex AI Search & Conversation

CDH

CCAI

Pub/Sub

CCAI

BigQuery

Gemini

Automation

Cloud Build

Gemini

Automation

IDEs

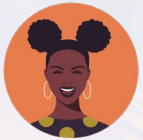
PEGA

Google Cloud

Act II - Automated Identification and Resolution of Network-Wide Modem Failure Incident

Start

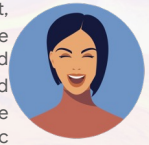
An increasing number of modems are failing across the CSP's network



Ann

The increasing volume of automated modem replacement cases has triggered a separate Network Operations workflow due to a statistically significant deviation from baseline modem failure levels

The network operations workflow assigned the case to the on-duty analyst, who initiated a root cause analysis. The analysis quickly identified that all failed modems were older models that had recently received the same firmware upgrade, despite no geographic commonalities



Mei

Using version control and Gemini Code Assist, the analysis identified a thread synchronization error in the firmware code, which would only manifest under specific circumstances and cause severe overheating. After correcting and testing the code, an emergency patch was issued to all impacted modems using an automated process.

The automated workflow coordinated the patch distribution to the affected modems. For modems that could not be successfully patched remotely after three attempts, clients were notified through the CSP mobile app to initiate an explicit firmware update

Success

Modems failing to receive the critical patch will automatically initiate a prioritized replacement workflow, ensuring swift resolution based on impact severity and device availability. This seamless process is entirely managed through self-service and automation, minimizing disruption for customers

Khalil

Developer

Customer

CSP Customer Service Representative

CSP SWE

Powered by

Gemini

Vertex AI Search & Conversation

CDH

CCAI

Pub/Sub

CCAI

BigQuery

Gemini

Automation

Cloud Build

Gemini

Automation

IDEs

PEGA

Google Cloud

Act III - GenAI-Enhanced Customer Service: Resolving Issues and Improving Experience

Start

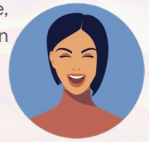
Ann, taken aback by an unexpectedly higher bill due to an unannounced increase in modem rental fees (now including pods), decides to contact her CSP's customer service for clarification

Ann

Customer

The CSR engages in a chat with Ann, reviewing the recent history of interactions, including the modem failure and upgrade. Simultaneously, Pega VoiceAI analyzes the conversation in real-time, identifying keywords and context relevant to Ann's reason for calling

The CSR chats with Ann, reviewing her recent modem failure and upgrade, while GenAI analyzes the conversation in real-time to provide support to the CSR



Mei

CSP Customer Service Representative

When asked if she needed further assistance, Ann mentioned connectivity drops when moving around the house, suspecting it was due to switching between pod WiFi signals. Pega Knowledge Buddy provided the CSR with instructions to enable Fast Roaming, which they shared with Ann. The call then ended

Pega CDH, based on the context, suggests waiving the fee increase for 24 months, which Ann accepts. An automated workflow then updates her billing and sends her a confirmation email

Khalil

Developer

CSP SWE

Success

The customer service app harnesses the power of GenAI to instantly generate a call summary, which the CSR can easily tailor. Leveraging this advanced technology, the CSR effortlessly creates a personalized email for Ann, complete with a direct link to the WiFi Fast Roaming article specifically tailored for her modem/router combination

Powered by

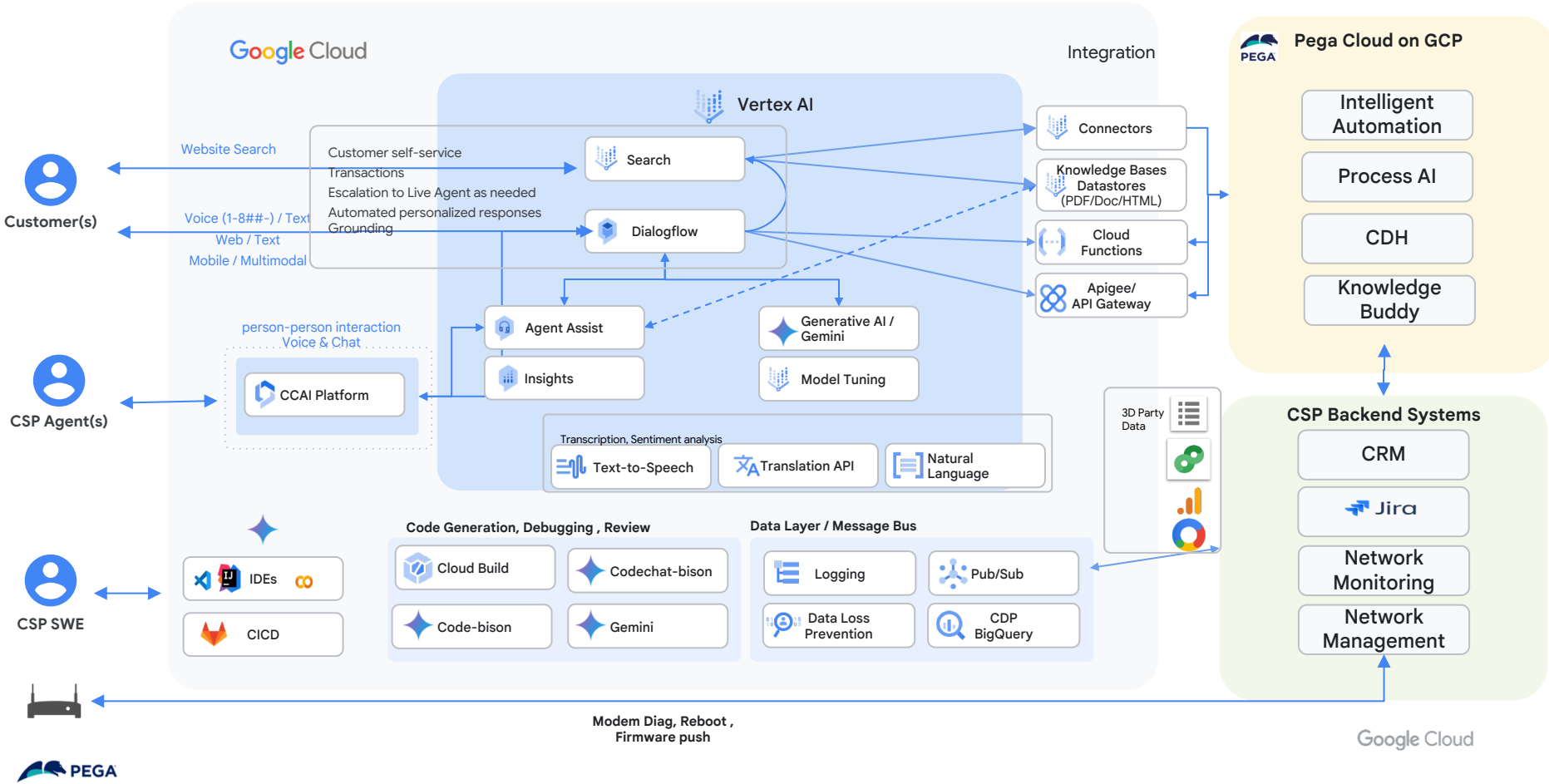
Gemini · Vertex AI Search & Conversation · CDH · CCAI · Pub/Sub

CCAI · BigQuery · Gemini · Automation

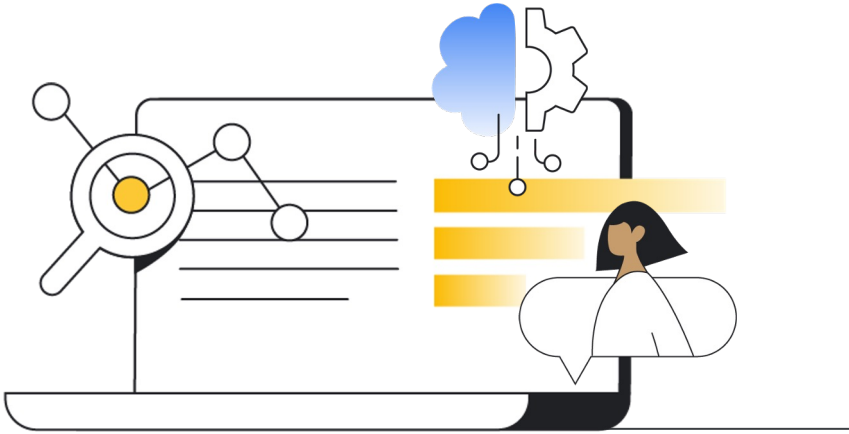
Cloud Build · Gemini · Automation · IDEs

PEGA

Google Cloud



Claims Review



PEGA

Google Cloud

Claims Review

Business Problem

Analysing coverage clauses, inclusion and fine prints, all submitted documents to quickly identify fraud, filter-out 'not covered' claims for rejection remains a time consuming process

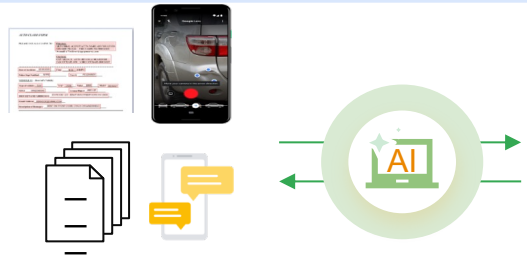
Description

Gen AI can help:

- Review policy documents and submissions in real-time and filter out covered vs non-covered portions of the claim
- Study past fraudulent claims data to identify potential cases of fraud or other irregularities.
- Data driven approach to adjudicate claims, provide reasonable, quick and accurate claim settlement.

Google/ Pega Value Proposition

GenAI App builder, Doc AI, Vertex AI, Pega Process AI



Revenue

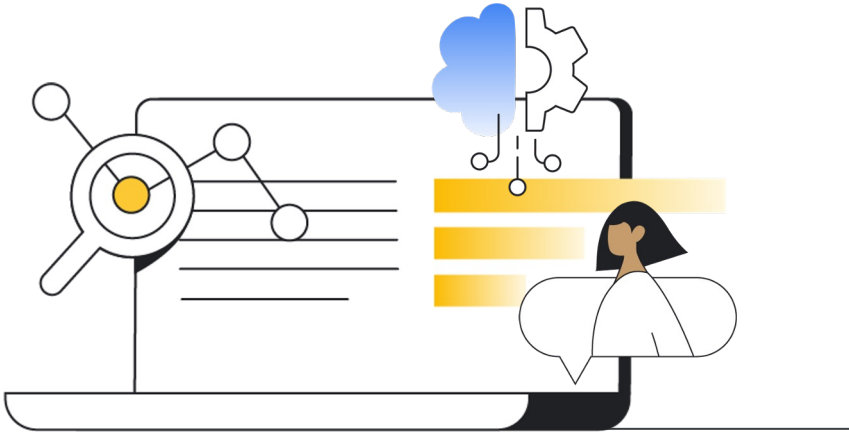
Time to Market

CSAT

PEGA

Google Cloud

Content & Ads Generation



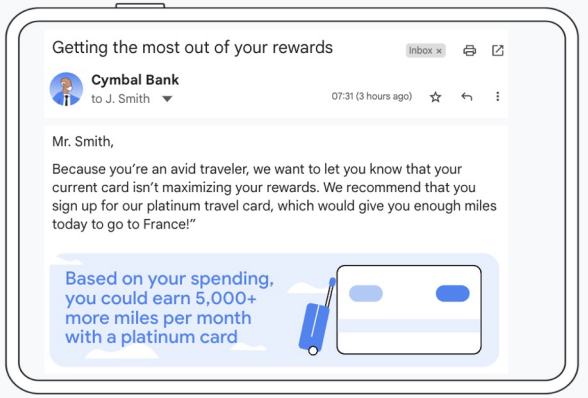
PEGA

Google Cloud

Use case: Creative Generation

SUMMARY

Create copy and images for marketing campaigns. Build multiple assets for personalization for consumers or cohorts leveraging the brand's own data-driven segments.



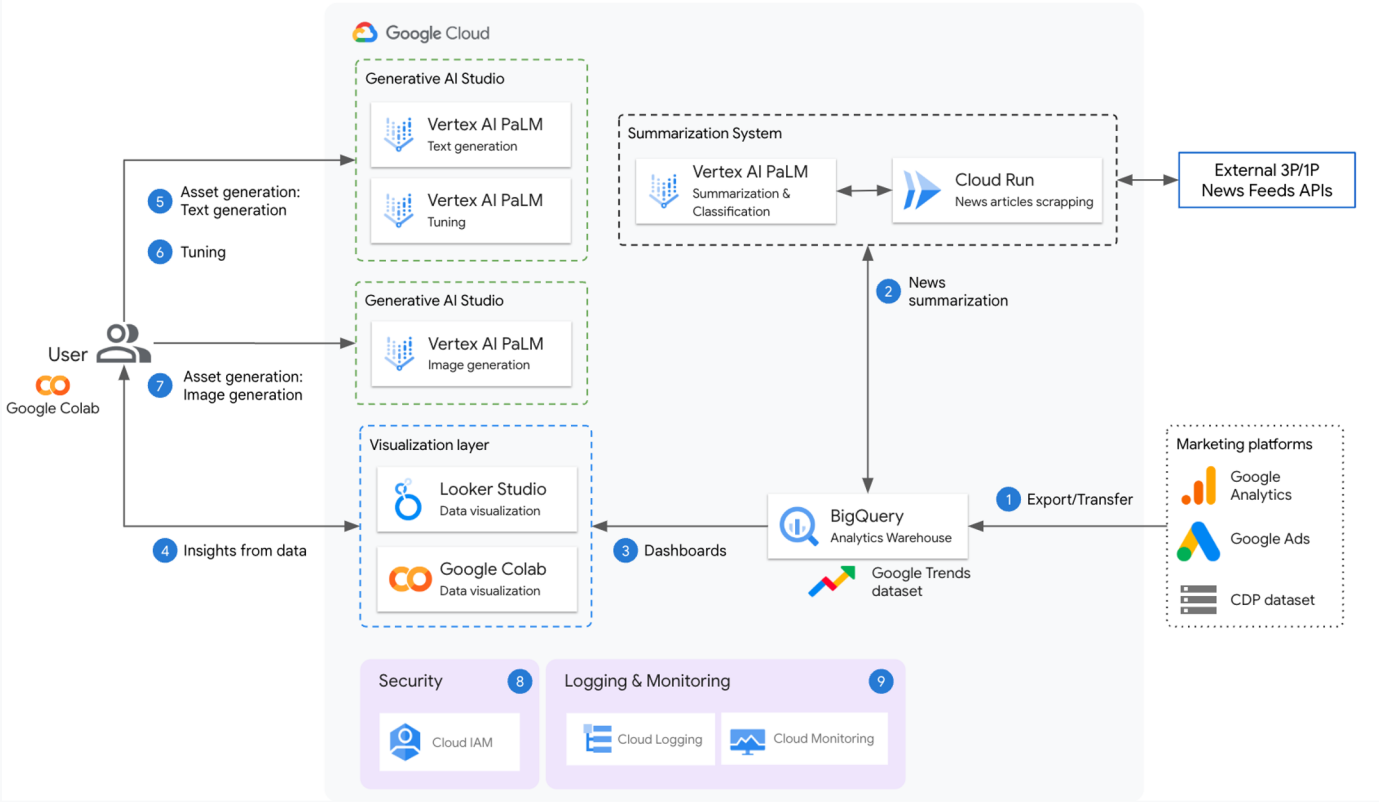
VALUE

- More creative options
- Increased agility and productivity
- Improved engagement & conversion

PEGA

Google Cloud

Use case: Creative Generation



Components

1. Export data from marketing platforms: A systems administrator sets up marketing platforms (Analytics, Ads and CDP) to automatically export their data to BigQuery, a serverless data warehouse platform.
2. News summarization: News articles are retrieved from a third-party API based on insights from Google Analytics, Ads and CDP. The text is pre-processed and the Vertex AI PaLM API provides a summary and a classification of the article. The data is then written to BigQuery.
3. Dashboards: A consolidated visualization of the collected data is provided on a Looker Studio or a Colab notebook.
4. Insights from data: To understand which ads are performing the best, one can explore data from Google Ads to analyze their performance metrics, such as number of clicks and conversion rate. Data from Google Analytics and CDP can be used to determine the audience that is consuming these ads, and data from Google Trends can provide additional insights on search terms.

   Asset generation - Text: A prompt that incorporates instructions and examples derived from the insights generated by previous steps is used to generate an asset.
5. Tuning: The model can be further adapted to generate better assets by tuning on a high quality datasets build based on top performing ad copies.
6. Asset generation - Image: After creating some ad copies, you can use them, along with a detailed description of how to generate an image, to create images for your ad copies. Use Vertex AI Generative AI Studio for that.
7. Security: Leverage all the security and data residency features of Vertex AI and BigQuery platforms.
8. Logging/Monitoring: Use logging and monitoring features from BigQuery and Vertex AI to understand data usage and resources consumption over time. This information can help you identify potential problems with your data pipelines or infrastructure, and make necessary adjustments to improve performance.
9.

Thank you.

Google Cloud

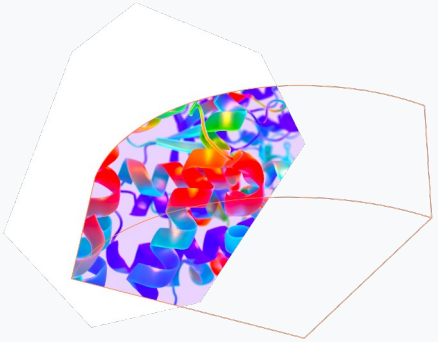
Appendix

Healthcare

Use case: Generative AI for drug discovery

SUMMARY

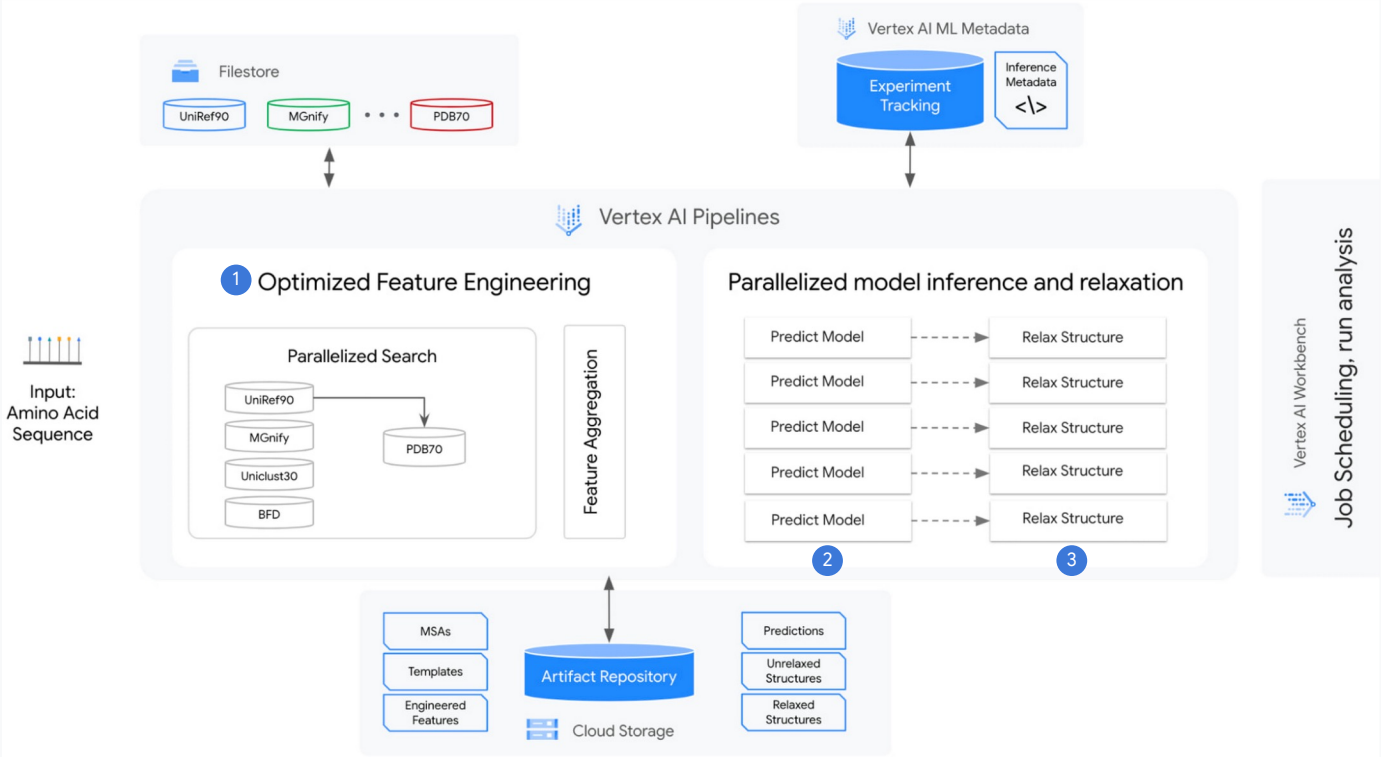
Accelerating the development of drugs through faster protein prediction using an operationally ready solution for running DeepMind's Alphafold on Google Cloud



VALUE

- Faster time to market
- Run experiments at scale with tracking
- Cost optimized and operationally ready

Use case: Protein Folding



Components

1. Feature preprocessing. Search through genetic sequences with common open source tools such as JackHMMER, HHBlits, and HHSearch. Returned multiple sequence alignments (MSAs) and structural templates are processed as inputs to inference.

2. Model inference. AlphaFold pretrained models, including predicting monomer and multimer structures. By default, one prediction is generated per model when folding monomer models, and five predictions are generated per model when folding multimers.

3. (Optional) Structure relaxation. In the AlphaFold system, you use the OpenMM molecular mechanics simulation package to perform a restrained energy minimization procedure. This resolves structural violations from inference.

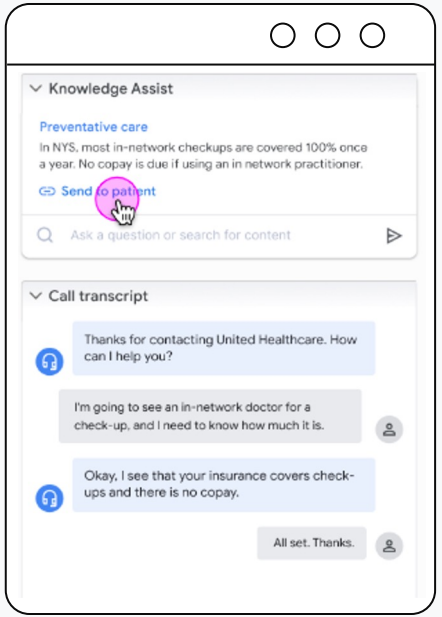
Use case: Digital Member Concierge

SUMMARY

Locate, summarize, and generate health plan customer service responses to improve operational efficiencies while also improving the customer experience.

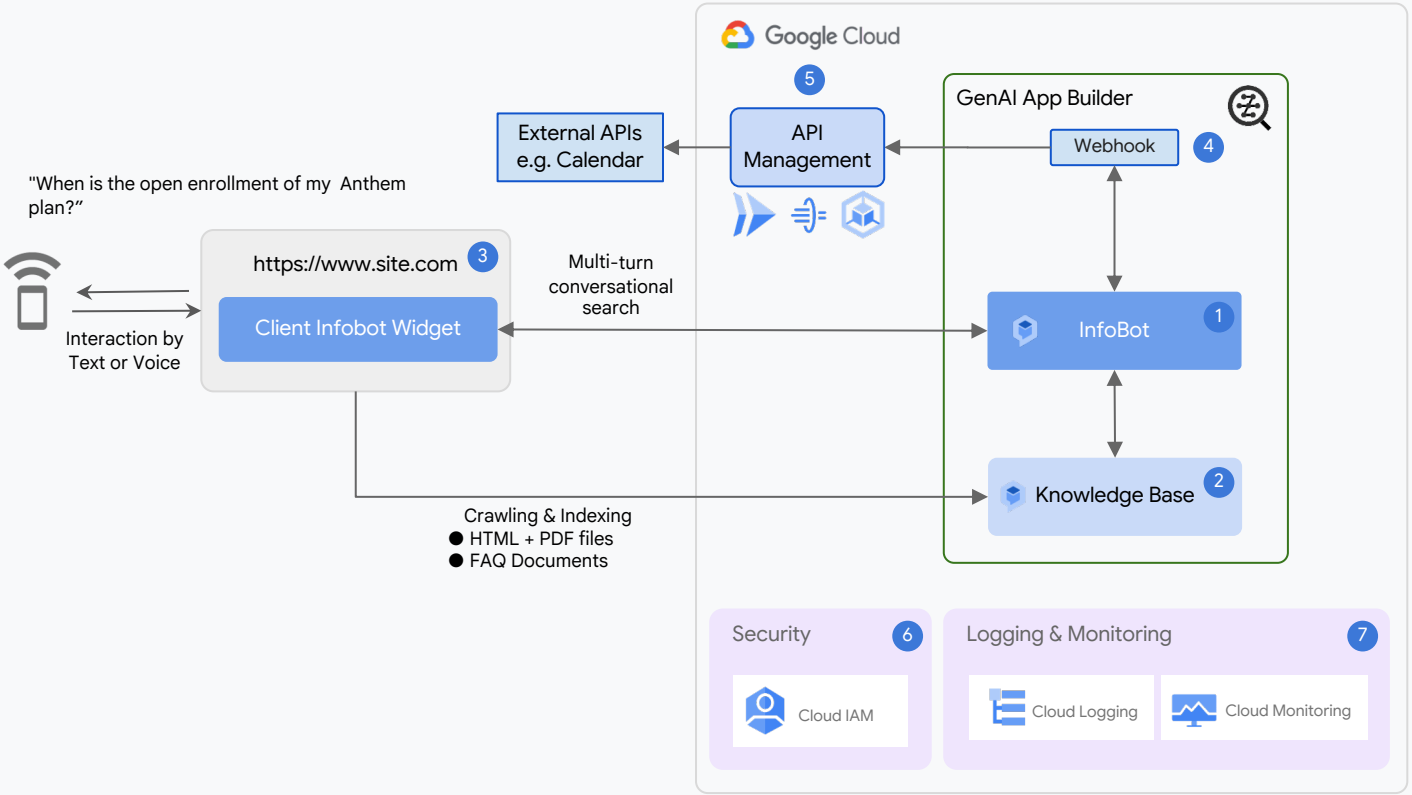
VALUE

- Improved patient satisfaction
- Reduction in call center calls and support
- Lower costs for consumer and patient care support



Patient can ask a chatbot specific coverage questions

Use case: Digital Member Concierge



Components

1. **Configure Infobot:** An Admin User configures Infobot, a virtual agent powered by large language models, to create experiences such as answering questions based on the organization's own contents. Infobot is a Dialogflow CX feature that is part of Generative AI App Builder.

2. **Configure Knowledge Base (KB):** Infobot uses Dialogflow CX KB to find answers for User's questions. The KB can be composed of your website domain, private documents, or FAQ pairs. After configuring Infobot, the virtual agent is published for users to interact with, using built-in integrations such as Dialogflow Messenger or custom widgets hosted on the organization's website.

3. **User interaction with Infobot widget:** User interacts with GenAI App Builder Infobot widget on the organization's website to get answers to their questions. The user interact with the Infobot using text .

4. **Call external APIs via Webhook:** Use Infobot webhooks to call APIs for external sources, such as calendar API

5. **API Management:** Build, manage, and secure API endpoints with tools such as Cloud Functions, Cloud Run, API Gateway, Apigee, etc.

6. **Security:** Infobot leverages all the security and data residency features from Dialogflow, like: Access control, Security Settings, VPC Service Controls, mutual TLS authentication, Regionalization, Custom CA certificates and Access Transparency.

7. **Logging/Monitoring:** Use Dialogflow CX monitoring and logging capabilities to monitor Infobot's conversation history and the analytics tool for Infobot's statistics.

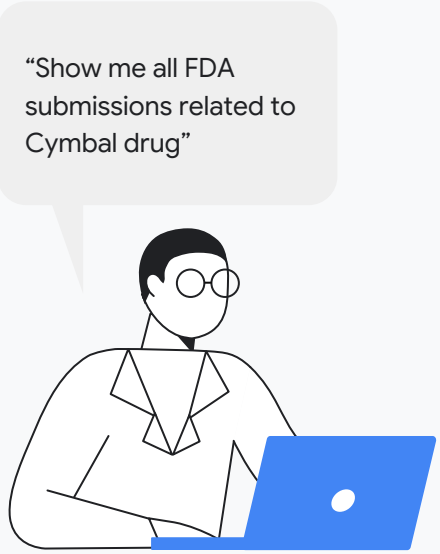
Use case: Public & Private Contextual Search

SUMMARY

Query and extract insights from from public datasets (medical literature) and private datasets (ELN). Quickly and accurately source insights from public and private datasets. Summarize research into plain language

VALUE

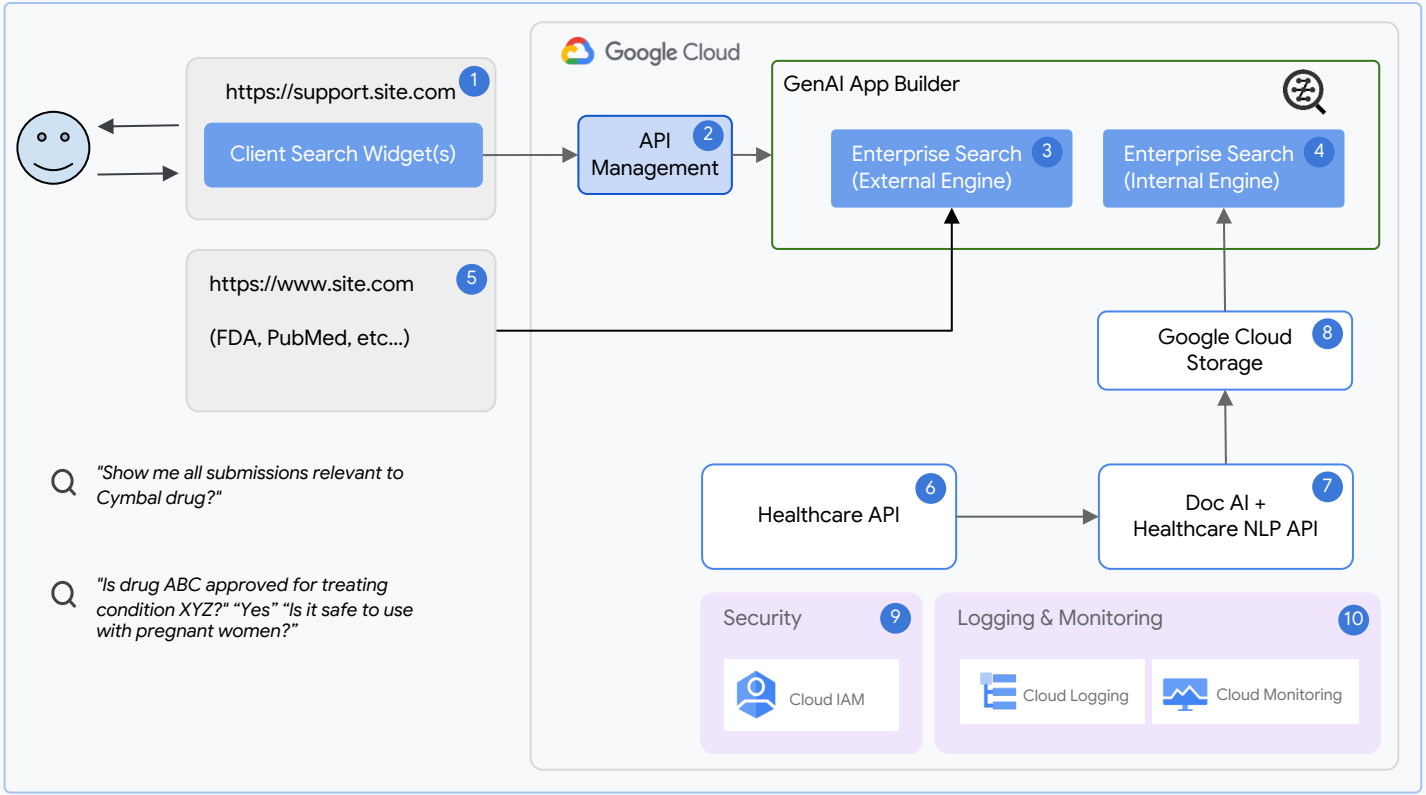
- Improved collaboration and communication between different personas in the healthcare industry
- Reduced burnout among field practitioners
- Improved patient satisfaction



PEGA

Google Cloud

Use case: Public & Private Contextual Search



Components Description

1. **Enterprise Search** provides HTML code for Widgets to embed on a page.
2. **API Management** to build, manage, and secure API endpoints with tools such as Cloud Functions, Cloud Run, API Gateway, Apigee, etc.
3. **Enterprise Search** External Engine, indexes and searches an external knowledge base.
4. **Enterprise Search** Internal Engine, indexes and searches internal documents and files.
5. **External knowledge bases:** External sources such as the FDA website, PubMed, etc...
6. **Healthcare API** Managing and storing internal documents.
7. **Doc AI + Healthcare NLP API** Ingests copies of internal technical documents and ELNs.
8. **Google Cloud Storage** Used to store structured or unstructured copies of the internal documents ingested.
9. **Security:** Enterprise Search leverages Identity and Access Management of Google Cloud.
10. **Logging/Monitoring:** Enterprise Search provides metrics such as click through rate, devices, etc.

PEGA

Use case: Clinician Notes Summarization

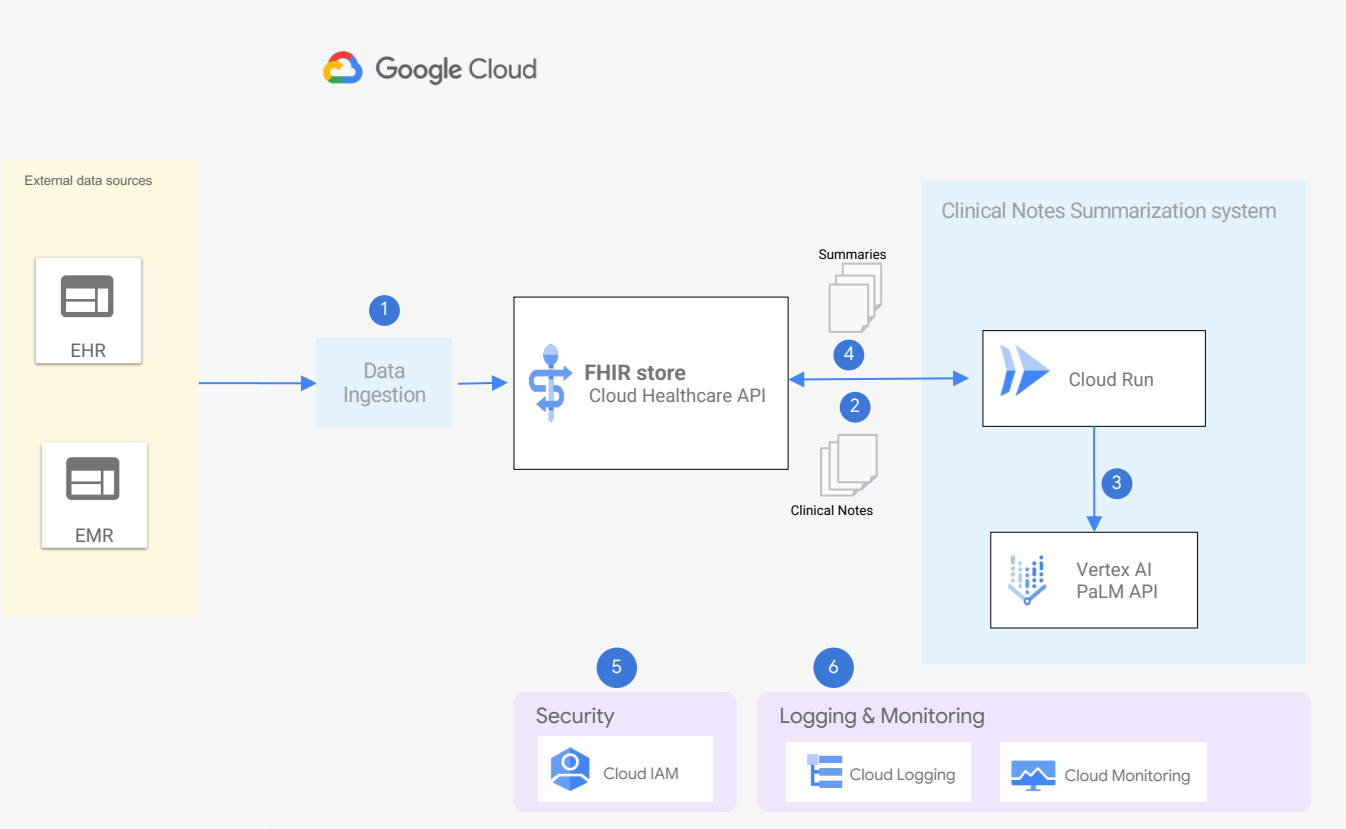
SUMMARY

Summarization of patient encounter notes to extract information (e.g. disease severity), which can be used in downstream processes such as population health management

VALUE

- Provides nurse practitioners and care providers timely information
- Ability to quickly get insights from data from integrated systems
- Reducing overheard, greater efficiency & cost reduction

Use case: Summarization of Clinical Notes

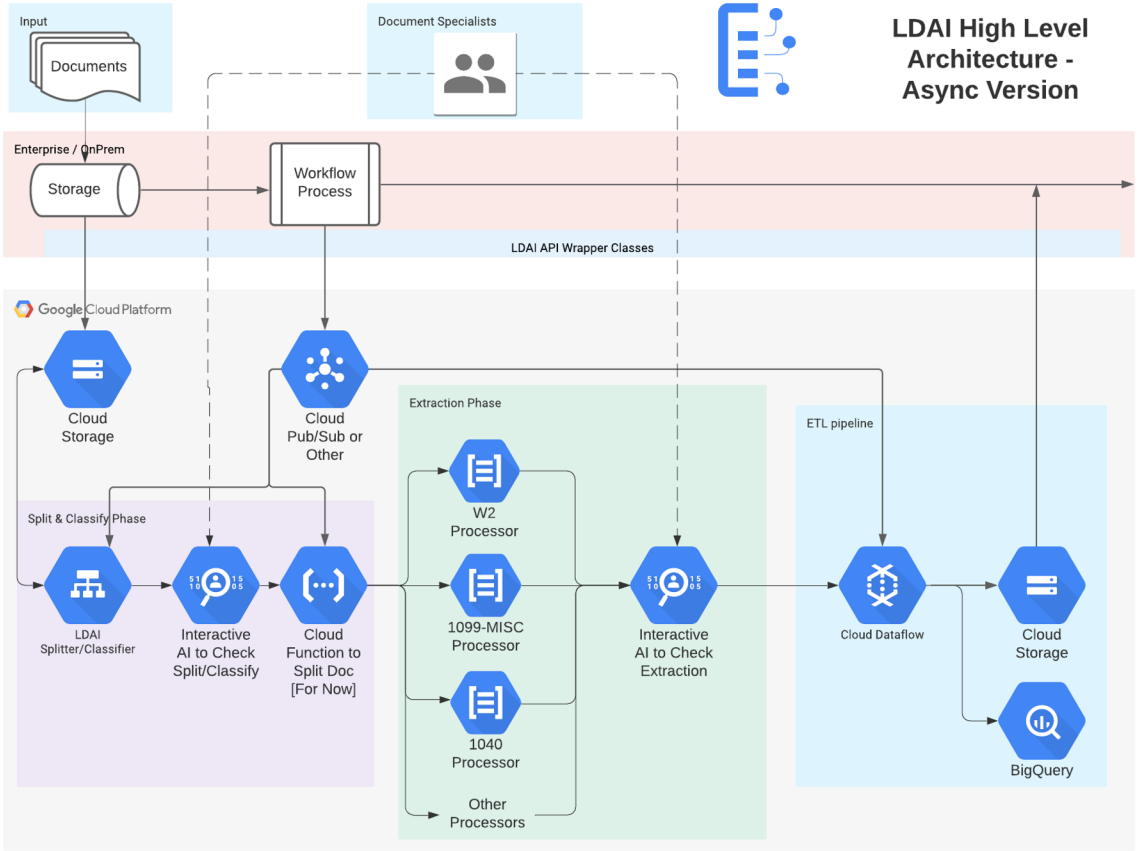


Components

1. Ingest data to Cloud Healthcare API: Data from external clinical systems is ingested into a Cloud Healthcare API FHIR Store..
2. Clinical Notes Extraction: Clinical notes for a patient or cohort of patients are extracted from the FHIR Store.
3. Summarization: Vertex AI PaLM API is invoked to summarize clinical notes. The notes are passed as context of the summarization prompt.
4. Integrate summaries to a clinical record. The generated summaries of clinical notes are integrated back into the clinical record in FHIR Store.
5. Security: Leverages Google Cloud IAM for unified authorization management.
6. Logging/Monitoring: Leverages Cloud Logging and Cloud Monitoring for unified and integrated management.

Financial Services

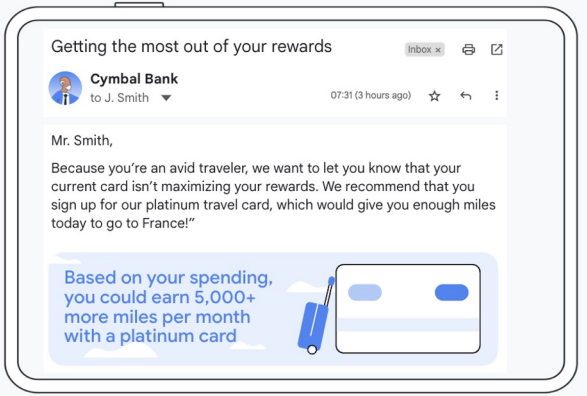
End-to-end GCP Reference Architecture - Async



Use case: Creative Generation

SUMMARY

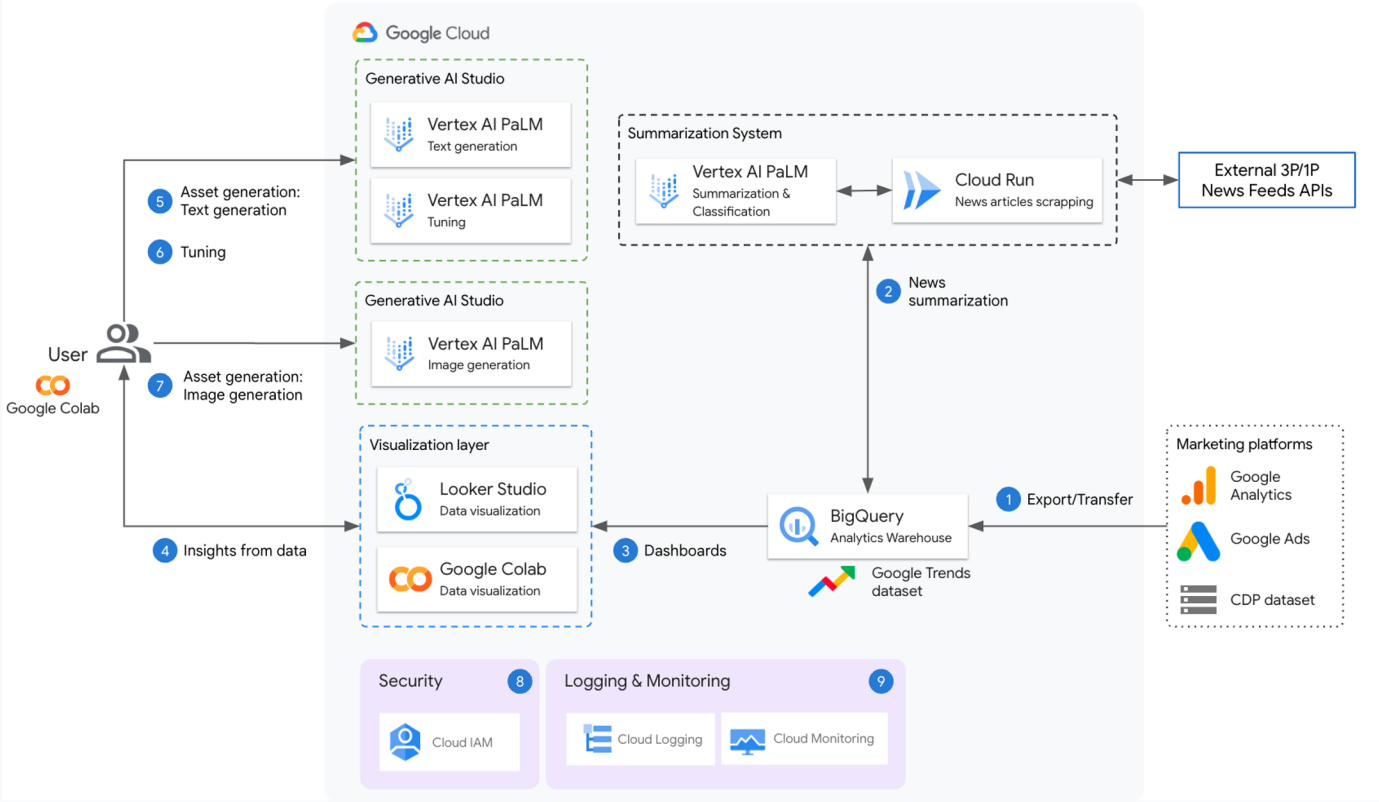
Create copy and images for marketing campaigns. Build multiple assets for personalization for consumers or cohorts leveraging the brand's own data-driven segments.



VALUE

- More creative options
- Increased agility and productivity
- Improved engagement & conversion

Use case: Creative Generation



Components

1. Export data from marketing platforms: A systems administrator sets up marketing platforms (Analytics, Ads and CDP) to automatically export their data to BigQuery, a serverless data warehouse platform.
2. News summarization: News articles are retrieved from a third-party API based on insights from Google Analytics, Ads and CDP. The text is pre-processed and the Vertex AI PaLM API provides a summary and a classification of the article. The data is then written to BigQuery.
3. Dashboards: A consolidated visualization of the collected data is provided on a Looker Studio or a Colab notebook.
4. Insights from data: To understand which ads are performing the best, one can explore data from Google Ads to analyze their performance metrics, such as number of clicks and conversion rate. Data from Google Analytics and CDP can be used to determine the audience that is consuming these ads, and data from Google Trends can provide additional insights on search terms.

   Asset generation - Text: A prompt that incorporates instructions and examples derived from the insights generated by previous steps is used to generate an asset.
5. Tuning: The model can be further adapted to generate better assets by tuning on a high quality datasets build based on top performing ad copies.
6. Asset generation - Image: After creating some ad copies, you can use them, along with a detailed description of how to generate an image, to create images for your ad copies. Use Vertex AI Generative AI Studio for that.
7. Security: Leverage all the security and data residency features of Vertex AI and BigQuery platforms.
8. Logging/Monitoring: Use logging and monitoring features from BigQuery and Vertex AI to understand data usage and resources consumption over time. This information can help you identify potential problems with your data pipelines or infrastructure, and make necessary adjustments to improve performance.
9.

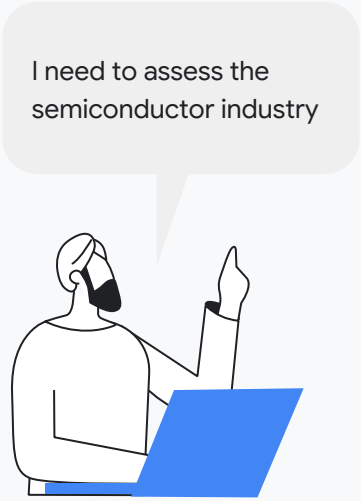
Use case: Capital Markets Research

SUMMARY

With the power of Google's Enterprise Search, Investment Advisors can generate meaningful, digestible and actionable insights from publicly available information for selected investment products from vast amounts of data including news, analyst reports and annual reports.

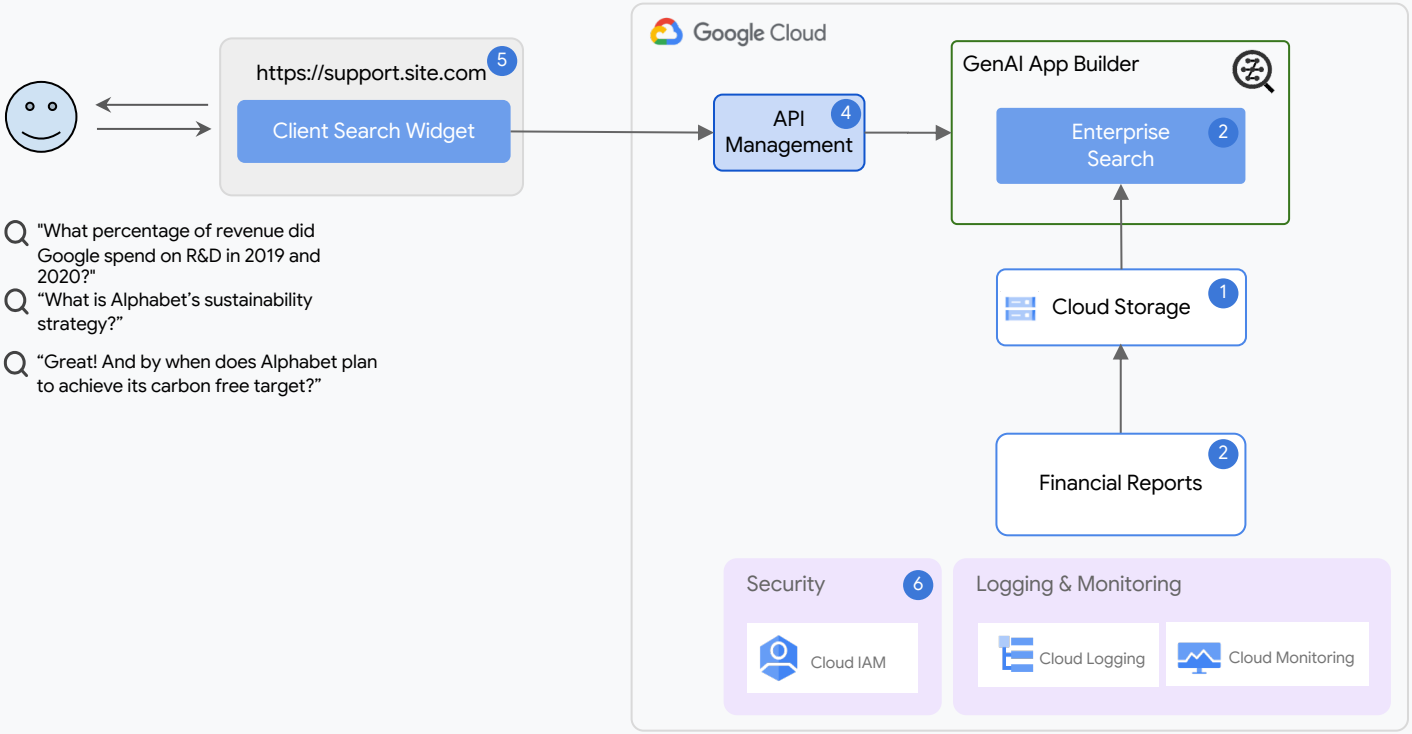
VALUE

- Increased employee productivity and efficiency with shorter time to research and accelerate time to insights
- Higher return on portfolio by getting consumable insights from massive amount of data
- Reduced risks for errors and inaccurate summarization



Investment manager

Use case: Capital Markets Research



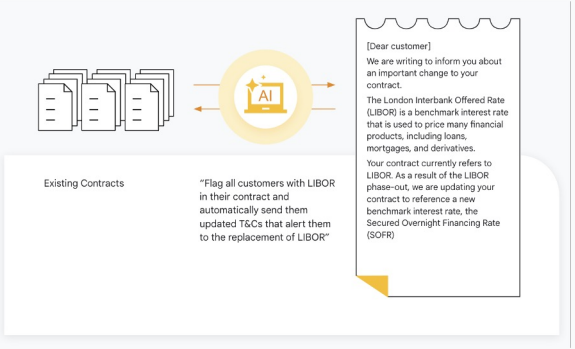
Components

1. Google Cloud Storage is used to store financial documents to be searched and analyzed.
2. Document loader financial documents are loaded and indexed.
3. Enterprise Search is a search platform for developers to build AI enabled, LLM-enriched, embedded, search capabilities and vertical solutions.
4. API Management to build, manage, and secure API endpoints with tools such as Cloud Functions, Cloud Run, API Gateway, Apigee, etc.
5. Enterprise Search provides HTML code for a Widget to embed on a page.
6. Security: Enterprise Search leverages Identity and Access Management of Google Cloud.
7. Logging/Monitoring: Enterprise Search provides metrics such as click through rate, devices, etc.

Use case: Financial Document Search & Synthesis

SUMMARY

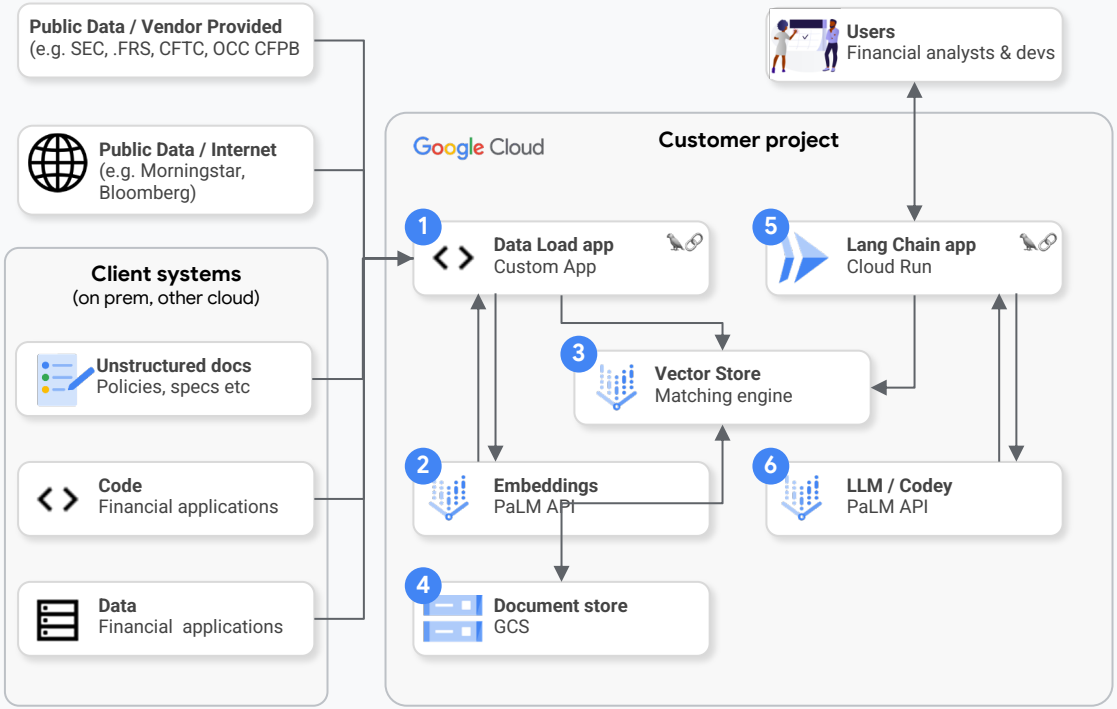
Improve the ability to understand large document sets to help identify, summarize, and explain relationships to other documents.



VALUE

- Faster document processing times
- Lower operational costs
- Improved employee productivity

Financial Document Search & Synthesis



Components

1. A **data load application** reads the data from a number of sources, which include:
   - Public data (e.g. SEC, .FRS, CFTC, OCC CFPB)
   - Vendor provided data
   - Proprietary (1st party) data, including policies, data and code
2. During data load, the **PaLM embeddings API** is called to create a vectorial representation of the content
3. The embeddings are stored in **matching engine**, that provides a managed service to store and query the embeddings. Alternative embeddings databases can be explored by customer if they need to implement retrieval mechanisms other than Approximate Nearest Neighbor
4. GCS is used by Matching Engine as a document store
5. Users interact with a conversational LLM implemented in Langchain, served by Cloud Run. The App retrieves the relevant documents from matching engine, and prepares the prompt for the LLM
6. The PaLM LLM API (text-bison and codey) are called to generate answers to the user queries

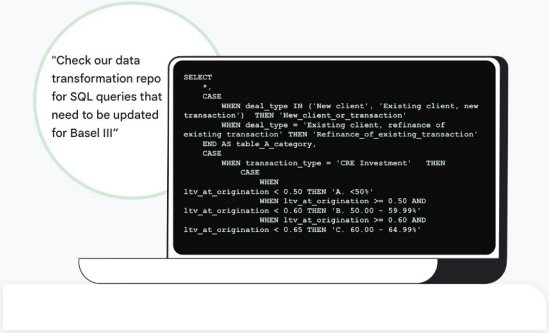
Use case: Regulatory Code Change Consultant

SUMMARY

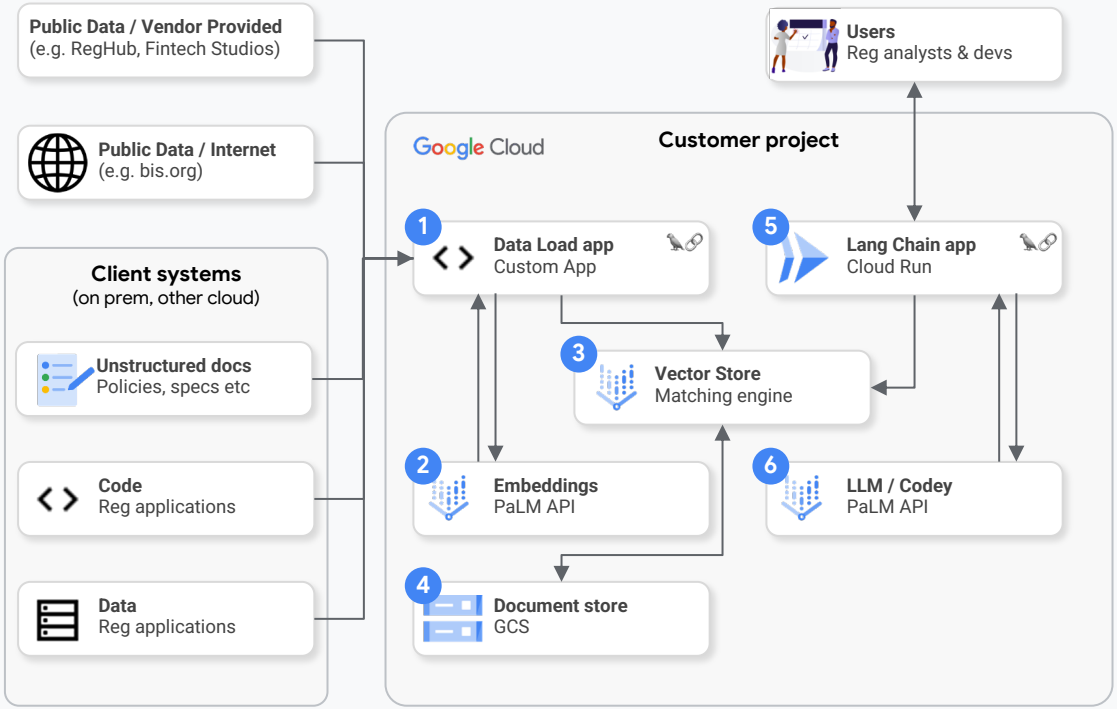
Help developers understand the underlying regulatory or business changes that will require them to change code, and assist in automating coding changes

VALUE

- Increased code robustness & scalability
- Accelerated time to launch
- Greater developer productivity



Regulatory Code Change Consultant



Components

1. A **data load application** reads the data from a number of sources, which include:
   - Public data (e.g. Basel 3 website)
   - Vendor provided data
   - Proprietary (1st party) data, including policies, data and code
2. During data load, the **PaLM embeddings API** is called to create a vectorial representation of the content
3. The embeddings are stored in **matching engine**, that provides a managed service to store and query the embeddings. Alternative embeddings databases can be explored by customer if they need to implement retrieval mechanisms other than Approximate Nearest Neighbor
4. GCS is used by Matching Engine as a document store
5. Users interact with a conversational LLM implemented in Langchain, served by Cloud Run. The App retrieves the relevant documents from matching engine, and prepares the prompt for the LLM
6. The PaLM LLM API (text-bison and codey) are called to generate answers to the user queries

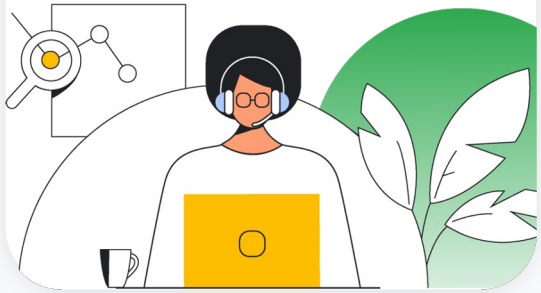
Retail

Generative AI for Retail

Generating and optimizing **content**

Creating **conversational commerce** experiences

Automating and improving **operations**

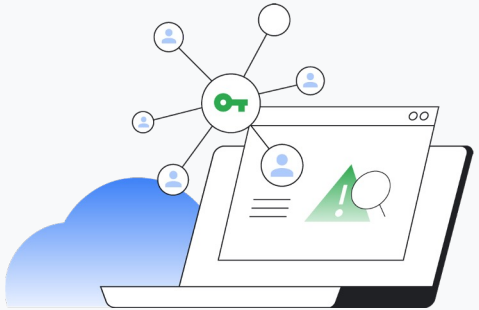


Use case: Products on digital shelf

SUMMARY

Use LLMs to automate and streamline the creation of rich product attributes, and detailed product descriptions from basic image or text data.

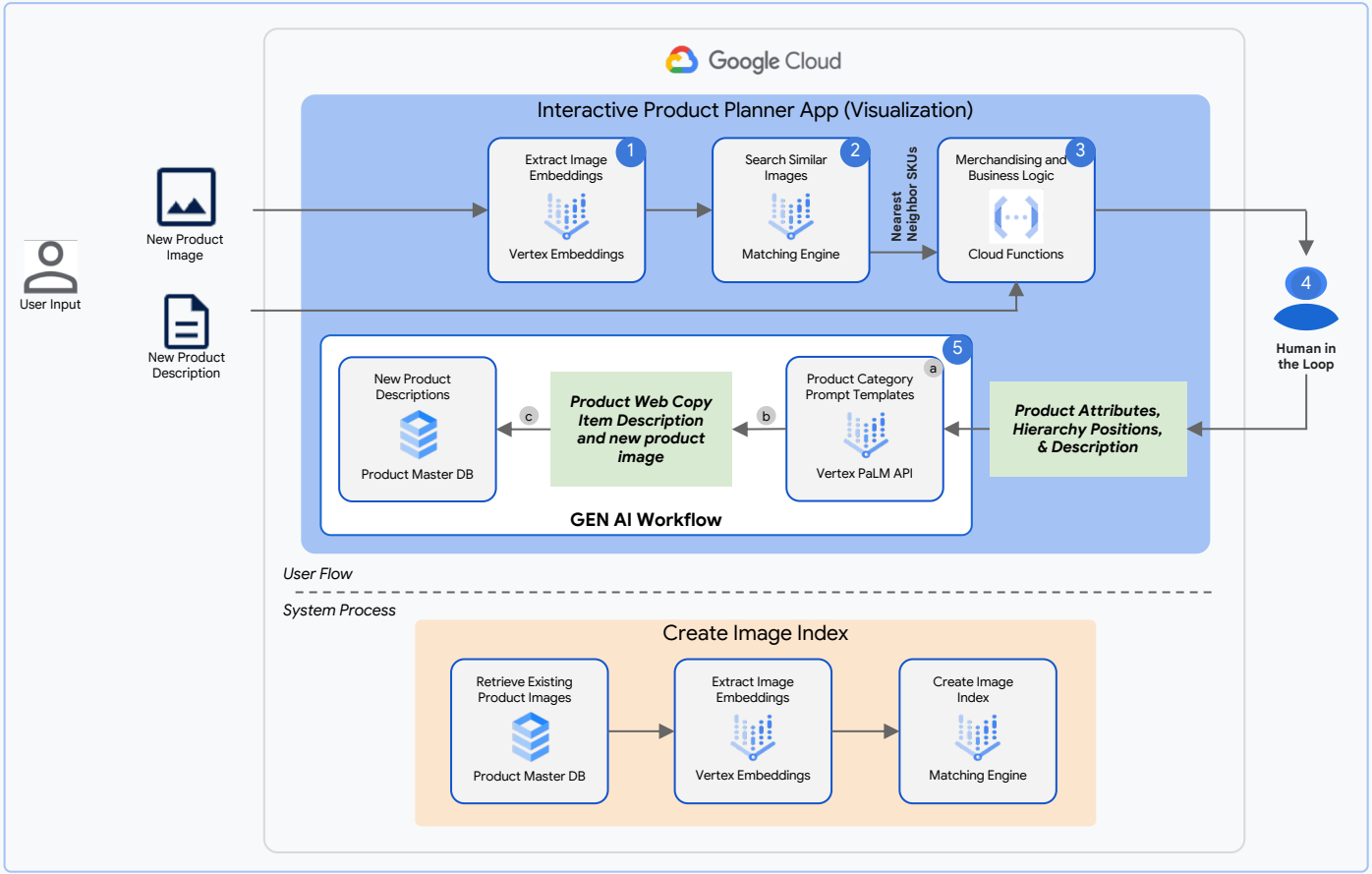
Power more consistent product ontologies and categorization with generative AI that drive better marketing, search, and conversion outcomes for your products.



VALUE

- Robust and detailed product hierarchies, rich and accurate product attributes.
- Improved product categorization, and more accurate demand planning and assortment optimization.
- Improved product descriptions and Ecommerce content.

Products on the Digital Shelf [Retail]



Components Description

1. Extract embeddings from uploaded image: We extract the image embeddings with Vertex AI Embedding Extractor. We will use these embeddings to search for similar images.

2. Search for similar images: We call the Vertex Matching Engine match function, and pass it the image embeddings. Vertex Matching Engine will return top ~n nearest neighbor image matches. Matching Engine image index created by Create Image Index process.

3. Business logic check: Ensure that the outputs of the automated ingestion process are consistent with the department and category of the new SKU, and with merchandising and risk and compliance rules. If there are any data quality issues, semantic anomalies, or business rule violations, an alert is generated. New product description is added to flow. Human-in-the-loop completes and audit. Implementation will vary by customer.

4. Human process: Review description from matched product retrieval, add product hierarchy information.

4. Send product attributes, hierarchy position, and description to GenAI workflow

5. A: few-shot PaLM API with category-specific templates.
B: generate web copy and (optionally) image(s) from PaLM API
C: insert new record into (customer) Product Master Database

Use case: Creative Generation

SUMMARY

Create copy and images for marketing campaigns. Build multiple assets for personalization for consumers or cohorts leveraging the brand's own data-driven segments.

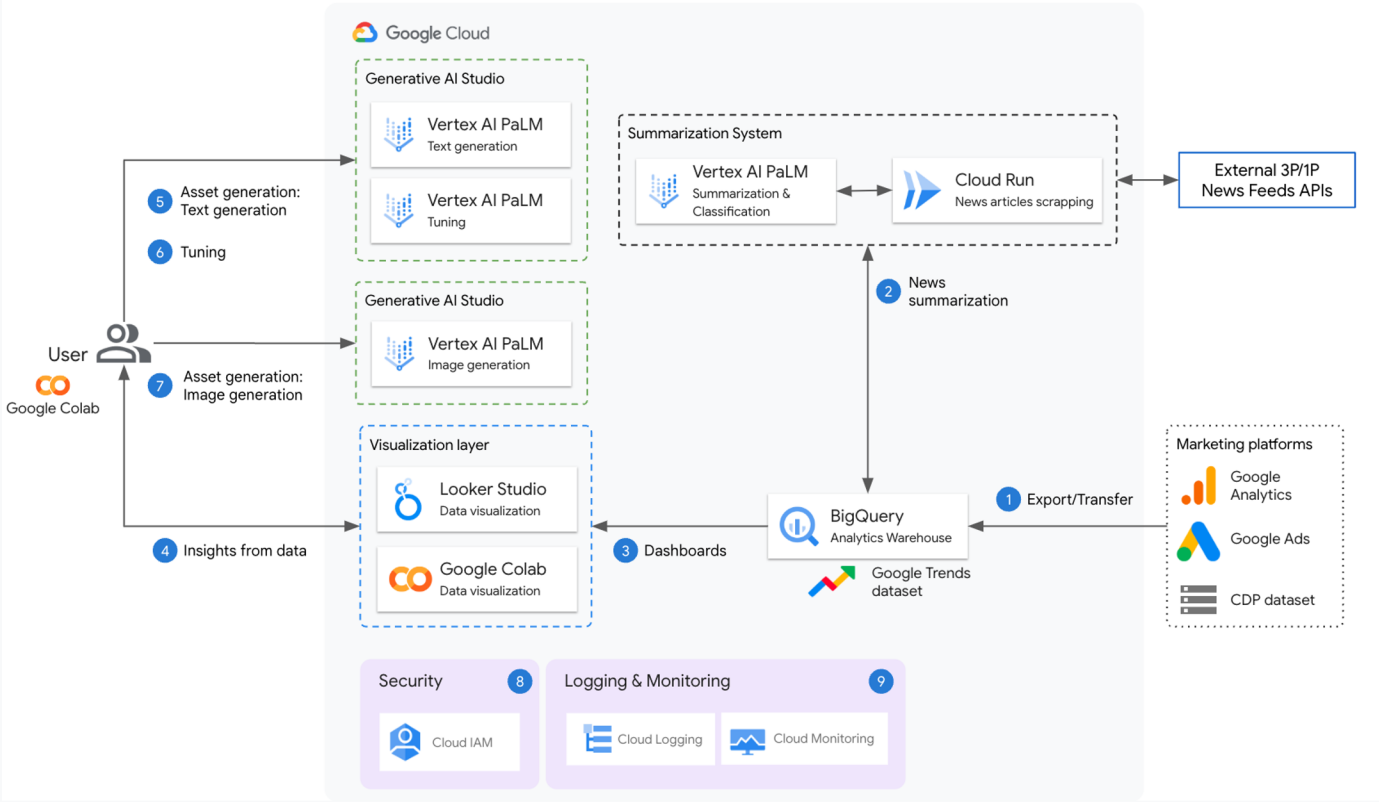
VALUE

- More creative options
- Increased agility and productivity
- Improved engagement & conversion

"Create several different versions of social media ads for our new campaign"

Advertising or Marketing Professional

Use case: Creative Generation



Components

1. Export data from marketing platforms: A systems administrator sets up marketing platforms (Analytics, Ads and CDP) to automatically export their data to BigQuery, a serverless data warehouse platform.
2. News summarization: News articles are retrieved from a third-party API based on insights from Google Analytics, Ads and CDP. The text is pre-processed and the Vertex AI PaLM API provides a summary and a classification of the article. The data is then written to BigQuery.
3. Dashboards: A consolidated visualization of the collected data is provided on a Looker Studio or a Colab notebook.
4. Insights from data: To understand which ads are performing the best, one can explore data from Google Ads to analyze their performance metrics, such as number of clicks and conversion rate. Data from Google Analytics and CDP can be used to determine the audience that is consuming these ads, and data from Google Trends can provide additional insights on search terms.

   Asset generation - Text: A prompt that incorporates instructions and examples derived from the insights generated by previous steps is used to generate an asset.
5. Tuning: The model can be further adapted to generate better assets by tuning on a high quality datasets build based on top performing ad copies.
6. Asset generation - Image: After creating some ad copies, you can use them, along with a detailed description of how to generate an image, to create images for your ad copies. Use Vertex AI Generative AI Studio for that.
7. Security: Leverage all the security and data residency features of Vertex AI and BigQuery platforms.
8. Logging/Monitoring: Use logging and monitoring features from BigQuery and Vertex AI to understand data usage and resources consumption over time. This information can help you identify potential problems with your data pipelines or infrastructure, and make necessary adjustments to improve performance.
9.

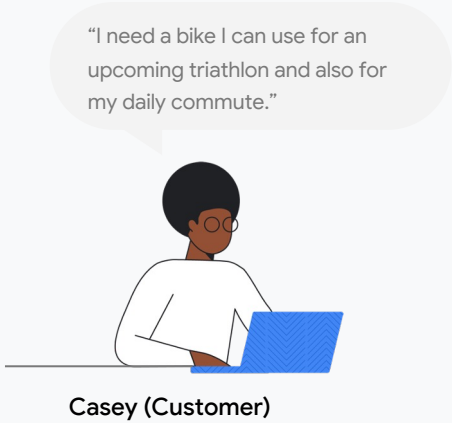
Use case: Search and Recommendations

SUMMARY

Build personalized search and contextual conversational experiences to help your customers have a shopping experience that's customized to the viewer in real-time.

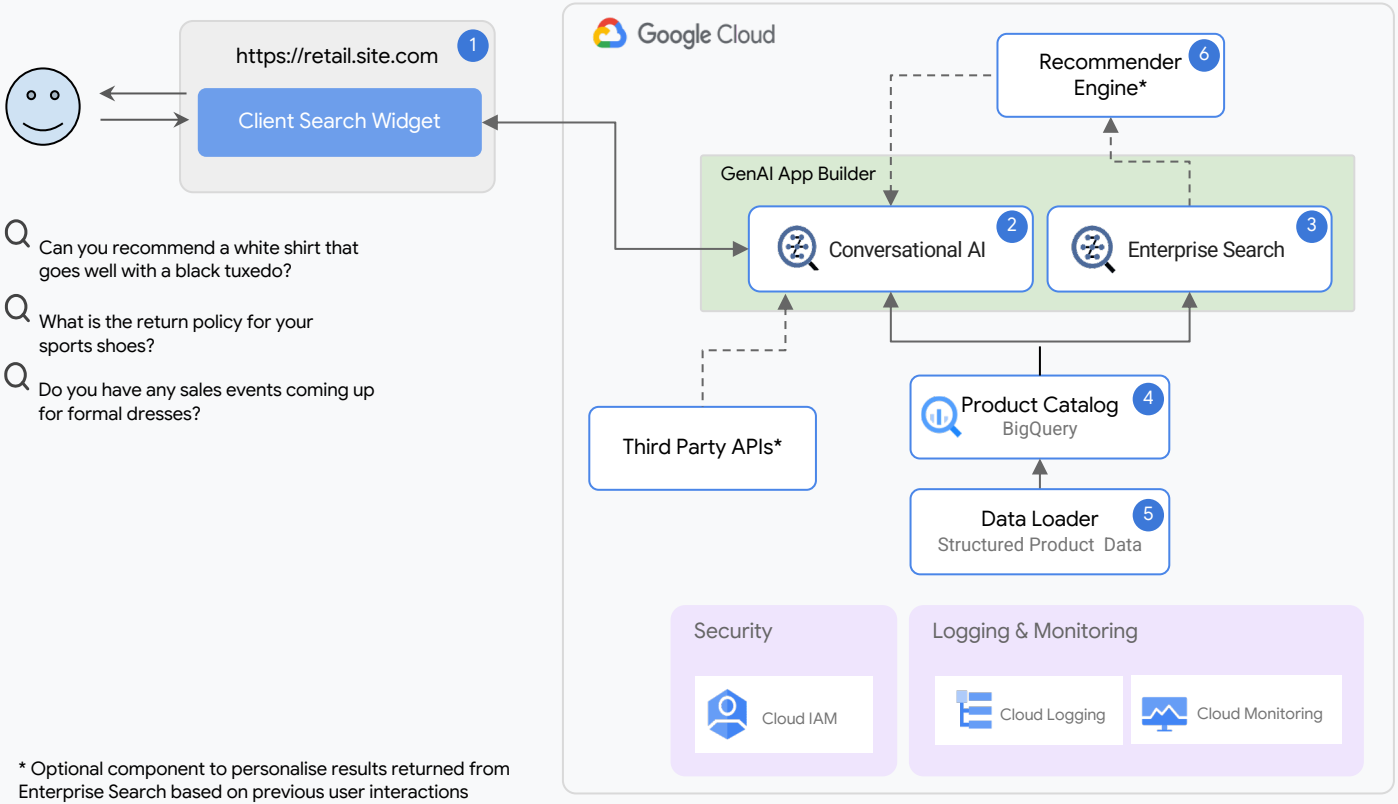
VALUE

- Drive sales
- Improve conversion
- Increase average order value
- Increase customer retention



Casey (Customer)

Use case: Search and Recommendations



* Optional component to personalise results returned from Enterprise Search based on previous user interactions

01 Search Widget is using HTML code embed on a page that is provided by the Enterprise Search service.

02 Enterprise Search is a search platform for developers to build AI enabled, LLM-enriched, embedded, search capabilities and vertical solutions.

03 Conversational IA allows chat-like interactions with users. Responses are formed from a predefined data sources (websites/gcs buckets/BigQuery).

04 BigQuery is used as a product catalog repository that is ingested by Enterprise Search and Conversational AI.

05 Data loader loads detailed product data into BigQuery. Product data should have enriched information so that chat and search requests can be matched using semantic information retrieval

06 Recommender Engine can be used in the conversation flow to filter and reorder Enterprise Search results based on users previous views and preferences.

Use case: Creative Generation

SUMMARY

Create copy and images for marketing campaigns. Build multiple assets for personalization for consumers or cohorts leveraging the brand's own data-driven segments.

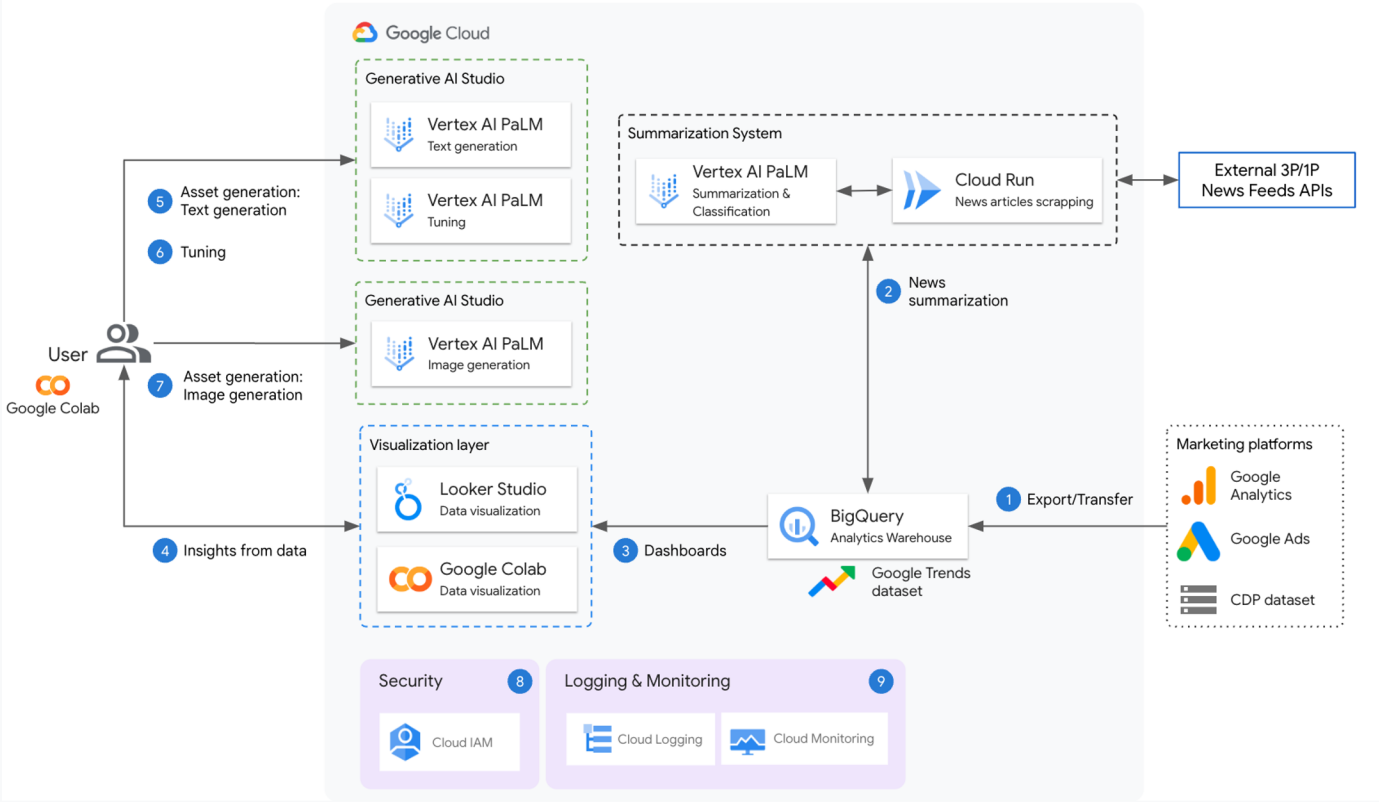
VALUE

- More creative options
- Increased agility and productivity
- Improved engagement & conversion

"Create several different versions of social media ads for our new campaign"

Advertising or Marketing Professional

Use case: Creative Generation



Components

1. Export data from marketing platforms: A systems administrator sets up marketing platforms (Analytics, Ads and CDP) to automatically export their data to BigQuery, a serverless data warehouse platform.
2. News summarization: News articles are retrieved from a third-party API based on insights from Google Analytics, Ads and CDP. The text is pre-processed and the Vertex AI PaLM API provides a summary and a classification of the article. The data is then written to BigQuery.
3. Dashboards: A consolidated visualization of the collected data is provided on a Looker Studio or a Colab notebook.
4. Insights from data: To understand which ads are performing the best, one can explore data from Google Ads to analyze their performance metrics, such as number of clicks and conversion rate. Data from Google Analytics and CDP can be used to determine the audience that is consuming these ads, and data from Google Trends can provide additional insights on search terms.

   Asset generation - Text: A prompt that incorporates instructions and examples derived from the insights generated by previous steps is used to generate an asset.
5. Tuning: The model can be further adapted to generate better assets by tuning on a high quality datasets build based on top performing ad copies.
6. Asset generation - Image: After creating some ad copies, you can use them, along with a detailed description of how to generate an image, to create images for your ad copies. Use Vertex AI Generative AI Studio for that.
7. Security: Leverage all the security and data residency features of Vertex AI and BigQuery platforms.
8. Logging/Monitoring: Use logging and monitoring features from BigQuery and Vertex AI to understand data usage and resources consumption over time. This information can help you identify potential problems with your data pipelines or infrastructure, and make necessary adjustments to improve performance.
9.

Use case: New Product Concept Development

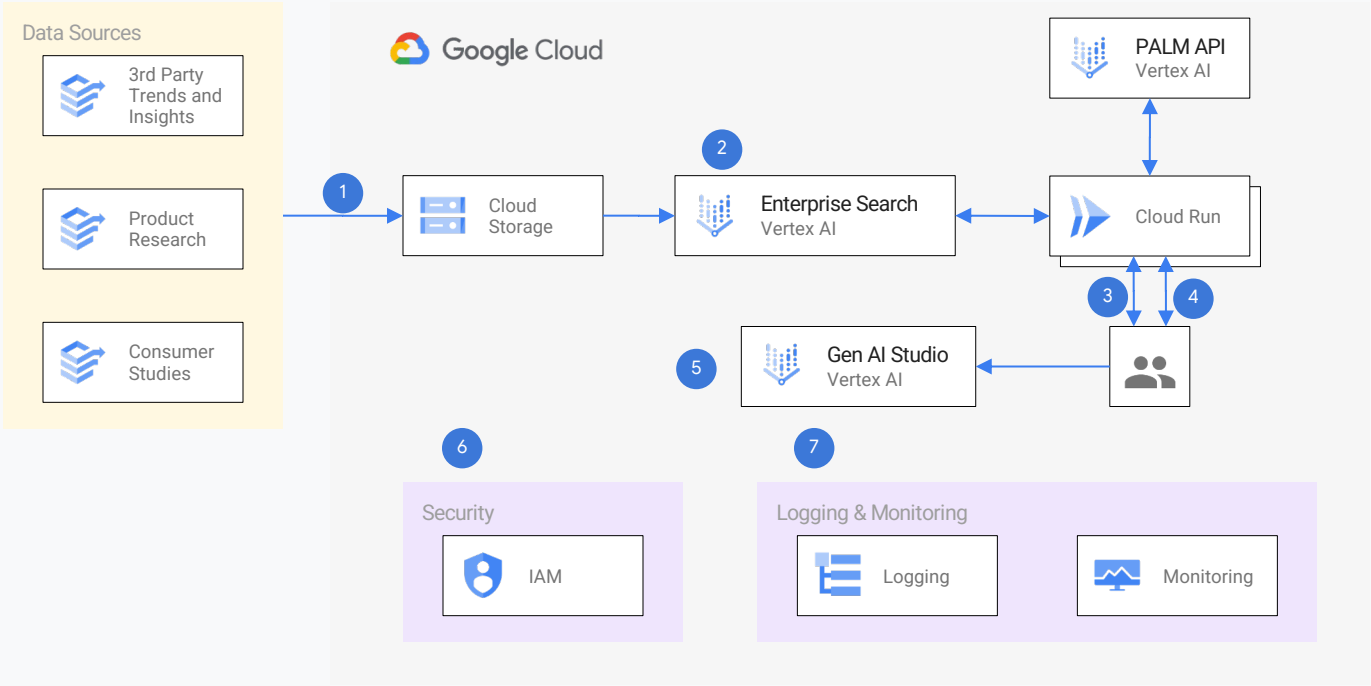
SUMMARY

Query consumer studies, product research and relevant external data. Summarize the key insights. Generate concepts, claims and images for testing.

VALUE

- Faster, more successful innovation
- Improved team productivity

Use case: New Product Concept Development



Components

1. Compile Source Materials: Compile source materials that would be useful in deriving new product insights
2. Index using Enterprise Search (ES): ES creates semantic embeddings and indexes the source materials
3. Insight Generation: User submits queries (e.g. 'What are the biggest unmet needs in the natural deodorant market?"). The query is used to fetch relevant documents from the ES index, which are in turn used as context to prompt the PALM LLM.
4. Concept Generation: Building on generated insights, user prompts for concepts (e.g. 'Write a story about natural deodorant that lasts longer'). These prompts are fed directly to PALM, bypassing ES.
5. Image Generation: From the leading concepts, use Gen AI Studio to develop imagery (e.g. 'Develop an image of a consumer using deodorant in the desert)
6. Security: Leverages Google Cloud IAM for unified authorization management.
7. Logging/Monitoring: Leverages Cloud Logging and Cloud Monitoring for unified and integrated management.

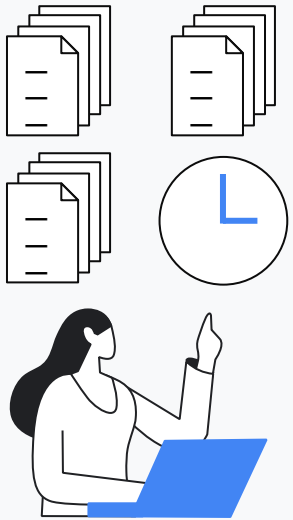
Use case: Employee Knowledge Search

SUMMARY

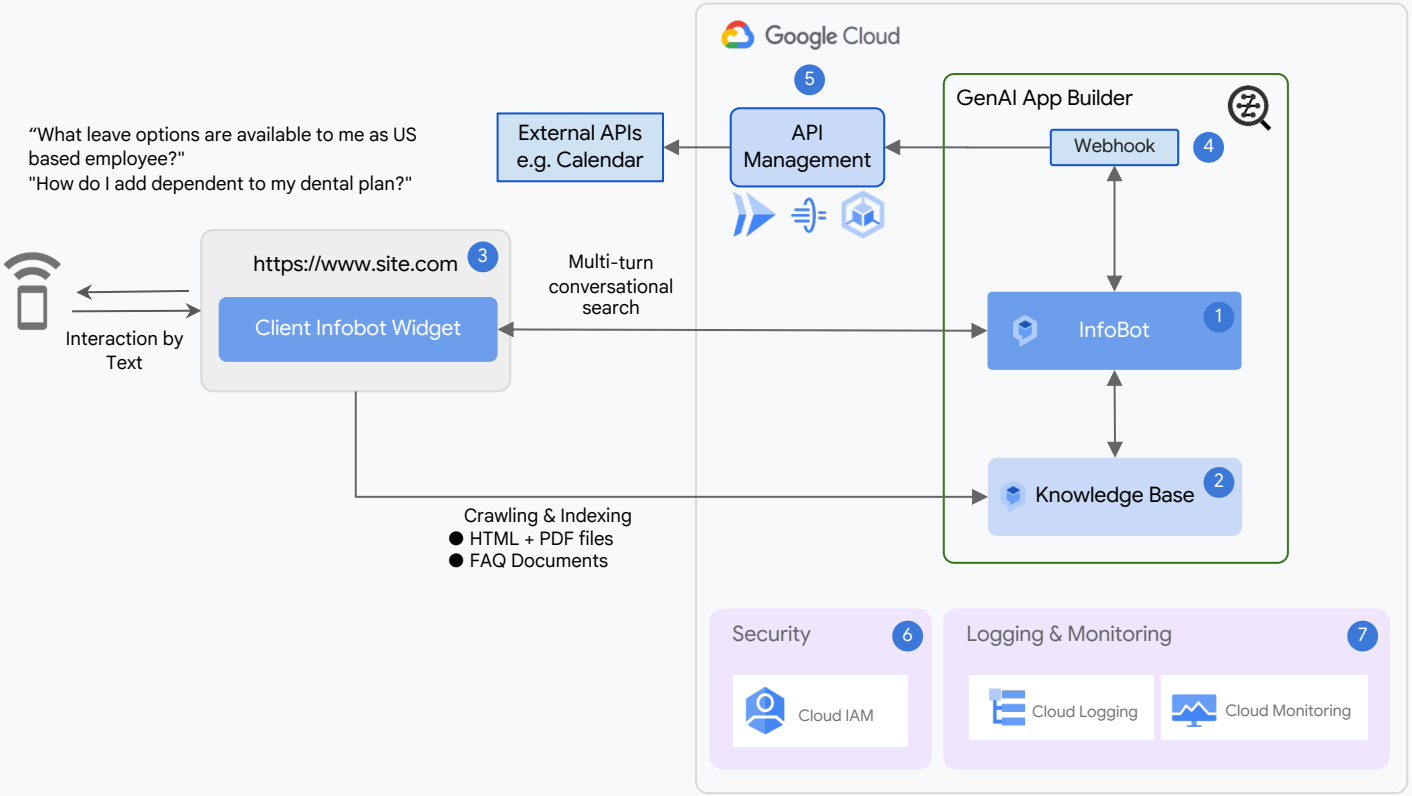
Quickly find the most relevant content via natural language search to reduce hours of manual work

VALUE

- Reduced toil
- Increased employee productivity



Use case: Better Employee Search



Components

1. **Configure Infobot:** An Admin User configures Infobot, a virtual agent powered by large language models, to create experiences such as answering questions based on the organization's internal HR contents. Infobot is a Dialogflow CX feature that is part of Generative AI App Builder.

2. **Configure Knowledge Base (KB):** Infobot uses Dialogflow CX KB to find answers for User's questions. The KB can be composed of your website domain, private documents, or FAQ pairs. After configuring Infobot, the virtual agent is published for users to interact with, using built-in integrations such as Dialogflow Messenger or custom widgets hosted on the organization's internal HR website.

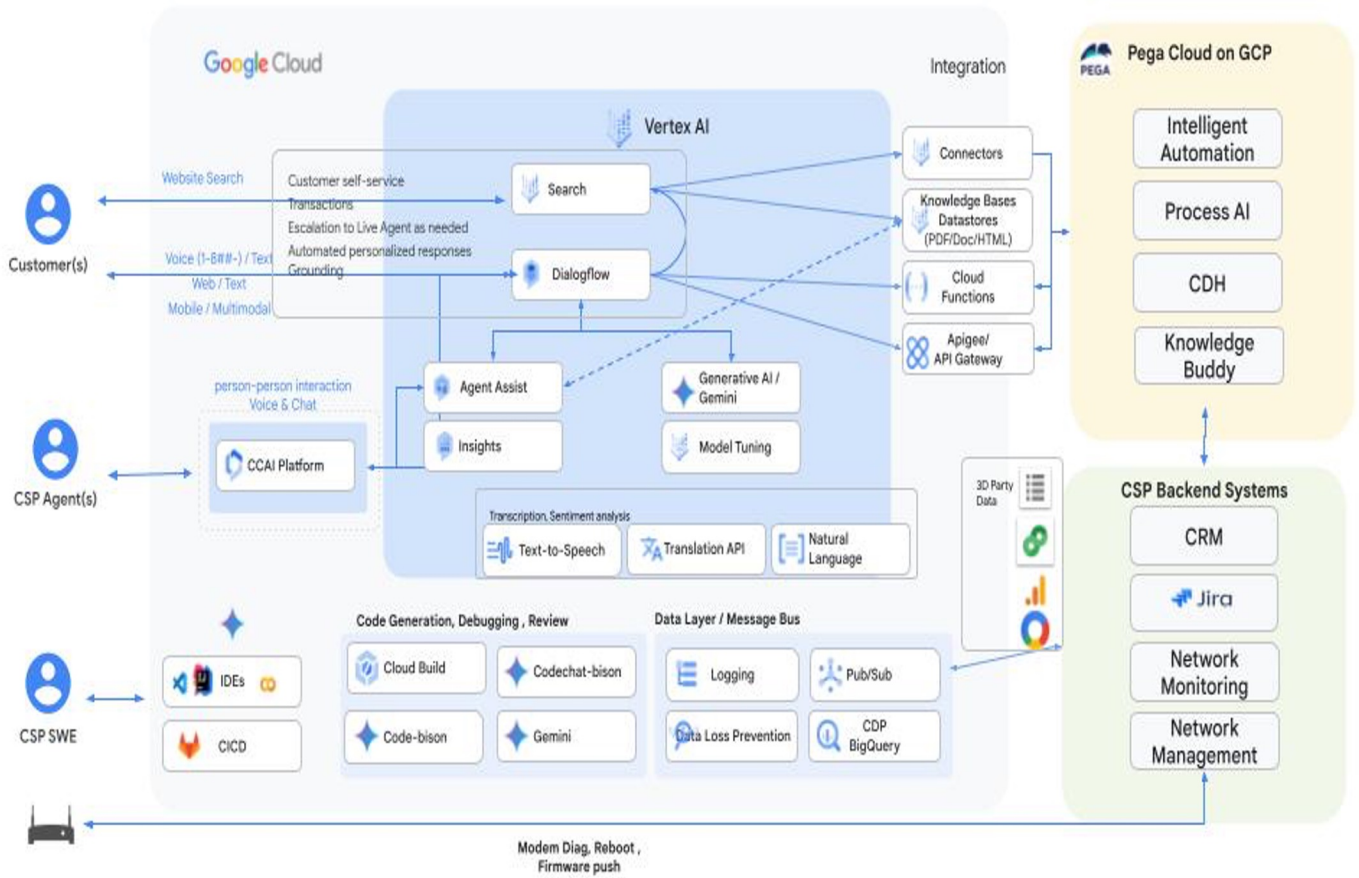
3. **User interaction with Infobot widget:** User interacts with GenAI App Builder Infobot widget on the organization's internal HR website to get answers to their questions. The user interact with the Infobot using text .

4. **Call external APIs via Webhook:** Use Infobot webhooks to call APIs for external sources, such as calendar API for easy scheduling

5. **API Management:** Build, manage, and secure API endpoints with tools such as Cloud Functions, Cloud Run, API Gateway, Apigee, etc.

6. **Security:** Infobot leverages all the security and data residency features from Dialogflow, like: Access Control, Security Settings, VPC Service Controls, mutual TLS authentication, Regionalization, Custom CA certificates and Access Transparency.

7. **Logging/Monitoring:** Use Dialogflow CX monitoring and logging capabilities to monitor Infobot's conversation history and the analytics tool for Infobot's statistics.



1

2 Ann has a combination modem + router that has had its WIFI routing freeze.

3 EVENT: Freeze conditions met

Automated action – Based upon customer preferences and the time-of-day , the modem is rebooted, remotely, and Ann is notified via push message to her CSP mobile app

4

5 The modem fails outright. Ann opens up her CSP mobile app (on 5G)

She initiates a voice-chat (in her native tongue) with the Gemeni Chat and explains the problem. The CSP AI asks her to take videos of the front and back of the modem. The AI her to turn the modem off and on again and, after waiting the necessary time, asks her to take pictures of the front and back again.

6

By comparing the state of the modem and looking at the WAN health in the area, the AI suggests that a new modem is necessary. It opens up a workflow for a self-service modem exchange within the app.

7

The workflow queries CDH and determines that Ann is eligible for an upgrade to Fibre in her area and that, if she is willing to wait until the installer's appointment, she will be given unlimited mobile hotspotting up to seven days. She accepts the offer.

8

The workflow enters self-service mode and determines that: 1) no permits are needed; 2) her nearest fibre node still has capacity and is within range of her home; and, 3) The equipment is available. The workflow then presents available time slots for installation within the next week. Ann books a time 2 days into the future.

9

Subflows automation send out a reminder prior to the installation with an option to reschedule. The day of the installation, the installer is presented with a list of all the materials and devices needed for that days' work . The fibre connection is installed, and the modems are exchanged. When the installer updates information on their mobile device, the workflow automatically updates