Google Cloud

Next '24

Google Cloud Next '24

ラスベガス

サマリー セッション資料（日本語）

April 11, 2024

Google Cloud

Data Cloud

Modern Infra Cloud

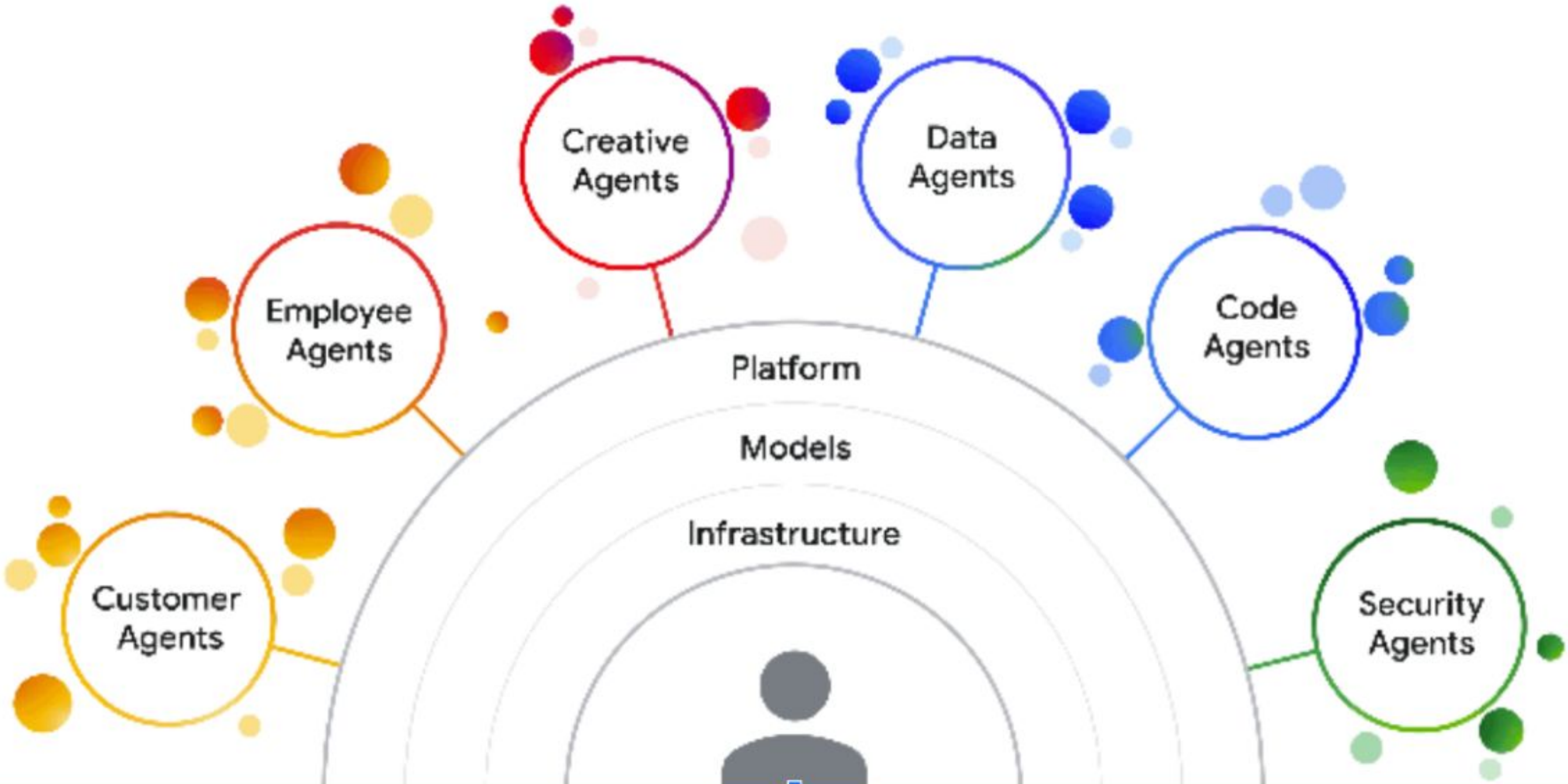
Collaboration Cloud

Security Cloud

Developer Cloud

AI Platforms & Models

Global Network of Cloud Regions



Creative Agents
Data Agents
Employee Agents
Code Agents
Customer Agents
Security Agents
Platform
Models
Infrastructure

Gemini
for Workspace & Google Cloud

お客様のエージェント型AI

Vertex AI

Gemini モデル

AI Hypercomputer

Infrastructure

Optimizing for AI workloads

Google's AI Hypercomputer Architecture

柔軟な利用形態

Dynamic Workload Scheduler | CUD | On demand | Spot

→ 動的なリソース確保とコスト最適化

オープンなソフトウェア

JAX　TensorFlow　PyTorch

マルチスライス学習、マルチホストでの推論、 XLA
Google Kubernetes Engine & Compute Engine

→ 高い柔軟性とユーザビリティ、多くの選択肢

パフォーマンスに最適化されたハードウェア

Compute (GPUs, TPUs) | Storage (Block, File, Object) | Networking (OCS, Jupiter)

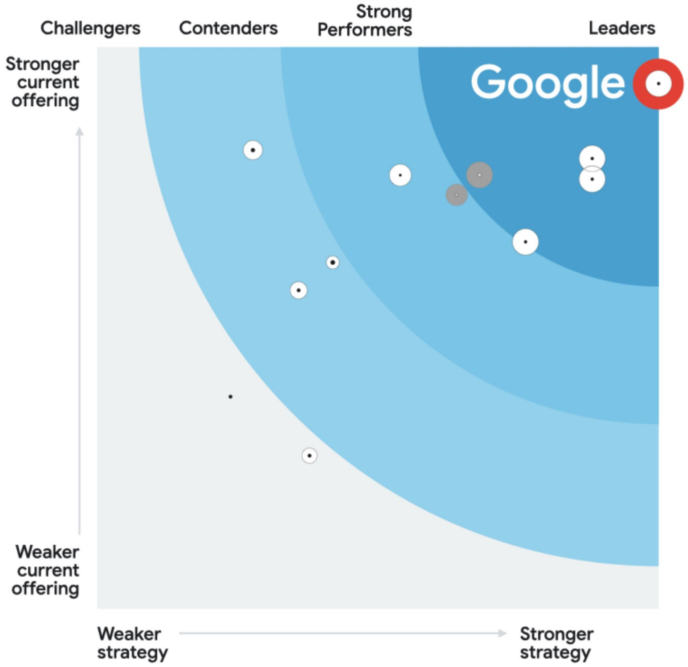
→ 卓越したパフォーマンスと性能

The Forrester Wave: AI Infrastructure Solutions

Google は 19 項目中 17 項目で
5 点満点の評価を得ました

- Solution Architecture
- Solution Ecosystem
- Data Workloads
- Training Workloads
- Inferencing Workloads
- Management Tools
- Developer Tools
- Fault Tolerance
- Efficiency

- Vision
- Roadmap
- Innovation
- Partner ecosystem
- Pricing Transparency
- Supporting services and offerings
- Number of customers
- Revenue



A3 Mega

一般提供

帯域幅を増強した NVIDIA H100 GPU 搭載の A3 マシン

2 倍の帯域幅

マルチノードの GPU 間通信で利用可能な帯域幅が A3 マシンの
2 倍である 1.6 Tbps に増強 (GPUDirect-TCPX の機能を利用した場合)

東京リージョンでも利用可能

2024 年 6 月から東京リージョンでも A3 Mega の提供を予定しており、国外へ持ち出せないデータを利用した学習が国内でも実行可能

GPUDirect-TCPX を使用して GPU ネットワークのパフォーマンスを最大化する
https://cloud.google.com/compute/docs/gpus/gpudirect?hl=ja

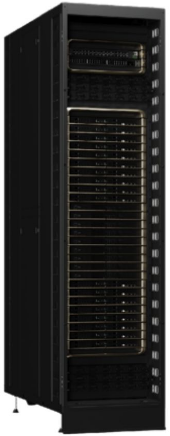
Google Cloud GPU のさらなる強化

ロードマップ

NVIDIA Blackwell が 2025 年上旬に登場予定

B200

メインストリームの学習やサービング



GB200 NVL72

フロンティア モデルの学習やサービング

一般提供

Google Cloud TPU v5p

最もパワフルでスケーラブルな TPU

パワフル

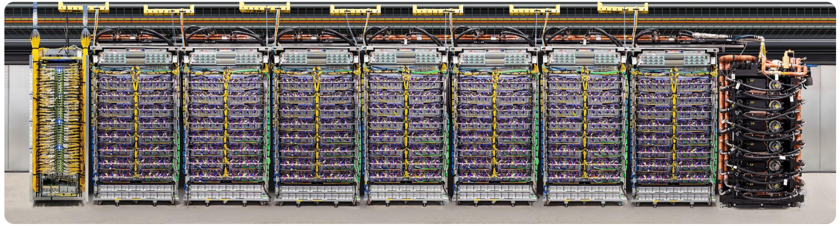
TPU v4 と比較して 2.8 倍の LLM（GPT3 175B）学習時間

スケーラブル

TPU v4 Pod と比較して 4 倍以上のスケーラビリティを持ち、簡単に 10,000 チップへのスケールが可能

柔軟

あらゆる種類（LLM、Recsys、マルチモーダル、MoE など）とサイズの AI モデルに対応



Dynamic Workload Scheduler

GPU リソースの柔軟な確保が可能に

一般提供

今までの課題

生成 AI の隆盛によって GPU の需要が高騰しているためにリソース確保のハードルが高い

解決策

リソースの連続使用日数の上限を設け、リソース デプロイの "順番待ち" を可能とすることで必ず GPU リソースのデプロイが可能

ユースケース

GPU の利用に時間的余裕のあるオフライン バッチ、実行時間の上限があるため短い時間で完了するジョブなどでの利用が最適

Calendar Mode

- 事前に利用期間や量を予約した上で利用
- 連続利用日数の上限は 14 日

Flex Start Mode

- オンデマンドでの利用
- 連続利用日数の上限は 7 日

Cloud Storage FUSE Read Cache

一般提供

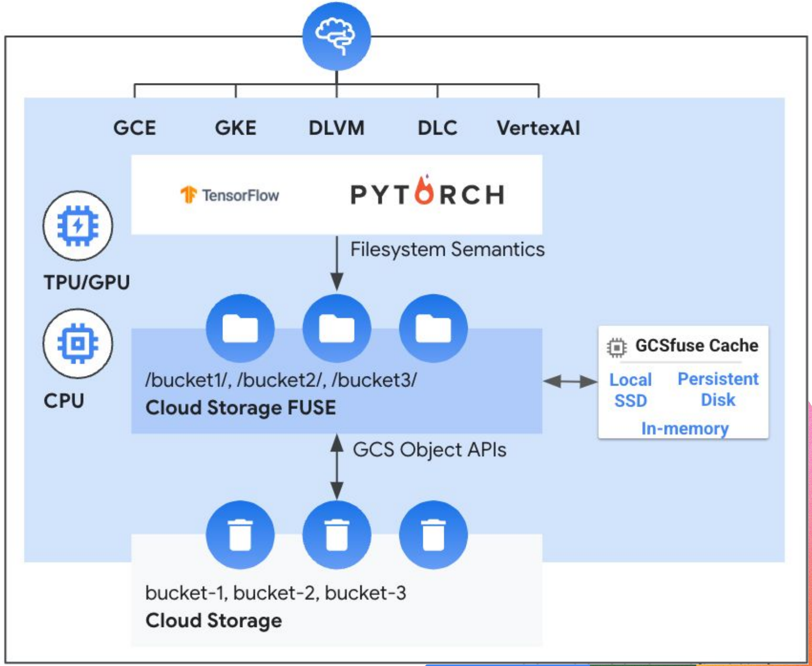
オブジェクト ストレージ内のデータへのアクセスを高速化

今までの課題

GCS FUSE によってデータ アクセスは簡易になるが、オブジェクト ストレージは小さいデータへの高頻度なアクセスやランダム I/O には最適化されておらず、レイテンシが高くなるために学習ワークロードでは学習時間が長くなる可能性がある

解決策

GCS FUSE Read Cache によって一度 GCS からダウンロードしたデータを Persistent Disk、Local SSD、メモリにキャッシュすることで同じデータへの繰り返しアクセスによって生じるオーバーヘッドが削減可能



Parallelstore

プライベート プレビュー

Intel DAOS ベースの低レイテンシの並列ファイル システム

高速でスケーラブル

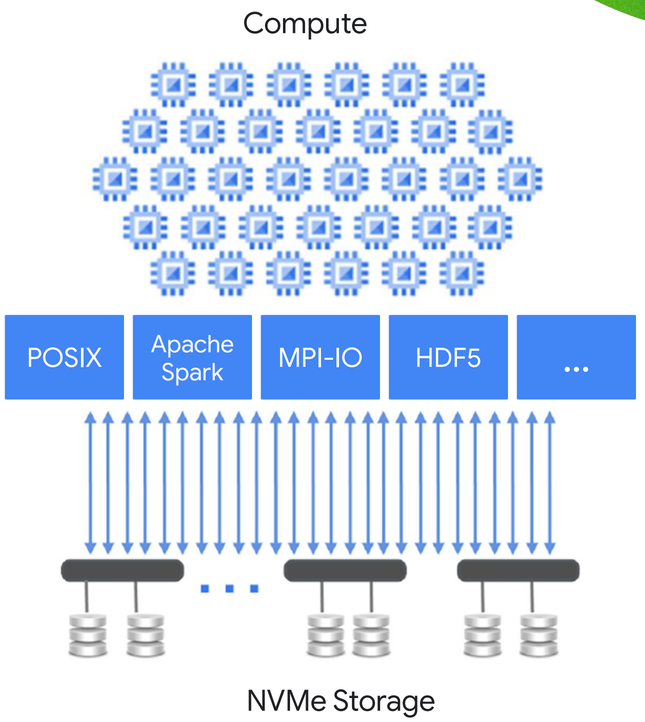
NVMe のストレージを利用した高帯域幅、高 IOPS、低レイテンシな並列ファイル システム

AI ワークロードや HPC に最適

複数のクライアントからの同時アクセスが発生するワークロードの場合は NFS やオブジェクト ストレージよりも並列ファイル システムが最適な選択肢

さまざまなクライアントに対応

Intel DAOS をベースとしているためにファイルへアクセスするための既存のエコシステム（POSIX、Apache Spark、HDF5 など）がそのまま利用可能



Hyperdisk ML

プライベート プレビュー

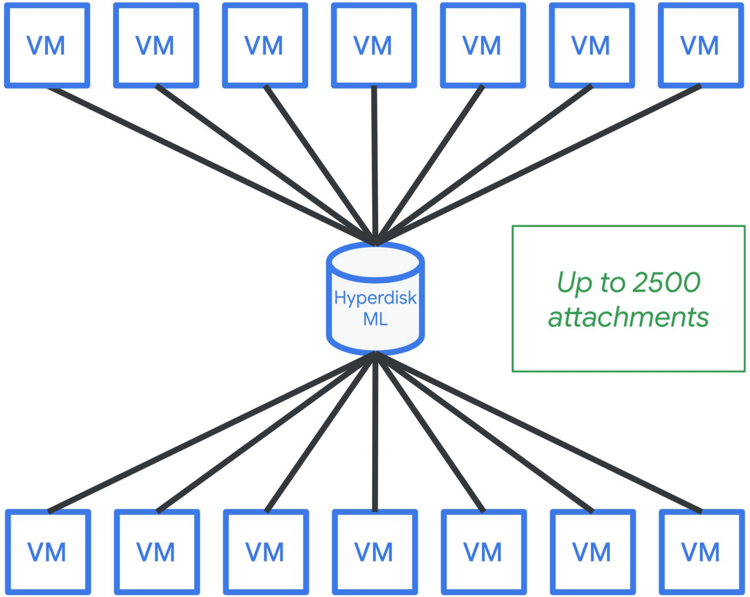
ML ワークロードに最適化されたブロック ストレージ

今までの課題

Hyperdisk を利用することでブロック ストレージのサイズに依存せず高いパフォーマンスを得られる一方で、ひとつのボリュームを複数のインスタンスへマウントができない（同じデータを参照する場合でも複数のボリュームが必要となるためにコストが増加）

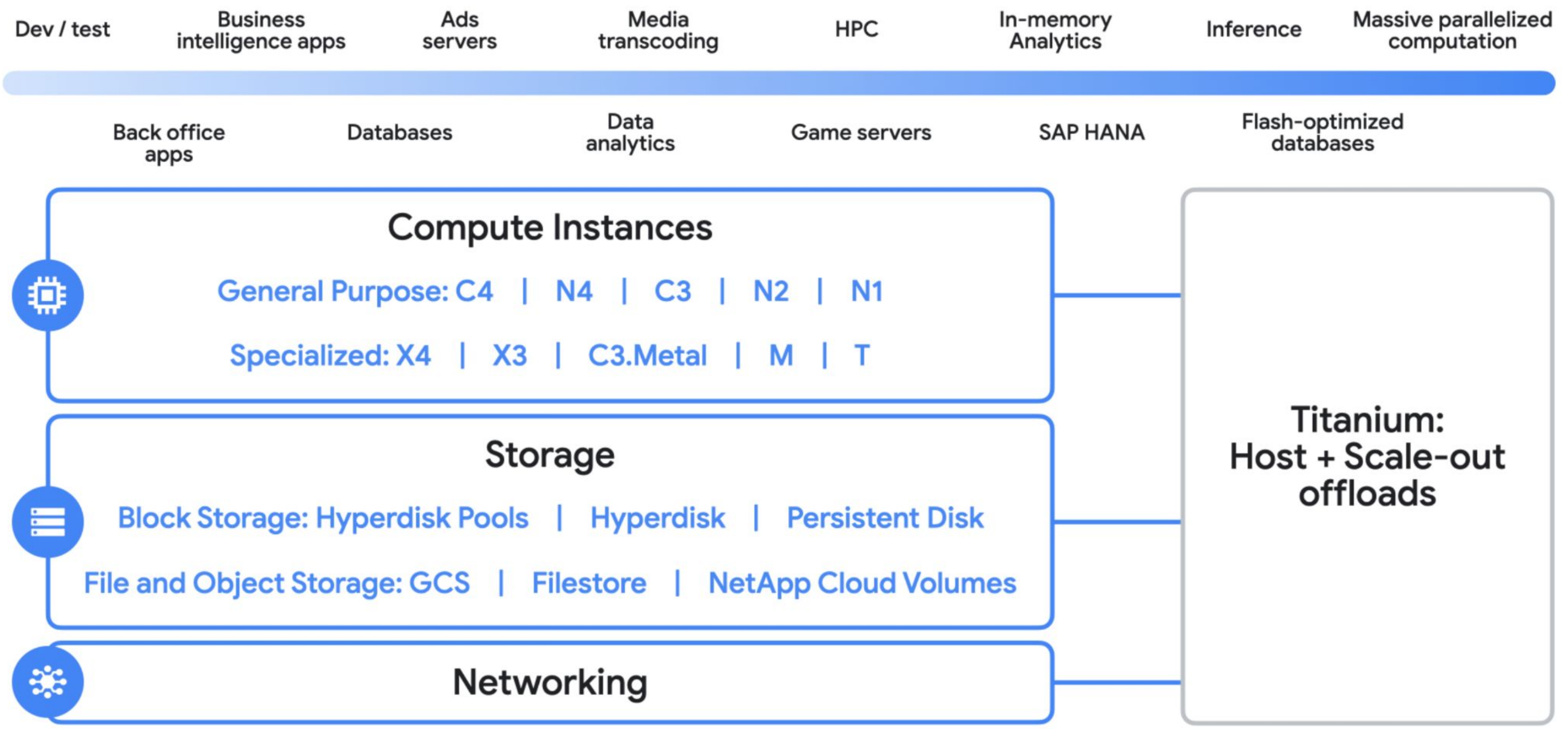
解決策

Hyperdisk ML はひとつのボリュームを最大 2,500 インスタンスから Read only で参照することが可能であるためにコスト削減につながり、最大 1.2 TiB/s のスループット（総合計）によってデータのロード時間を大幅に短縮可能



Optimizing for Cloud First and Enterprise workloads

ワークロードに最適化された Compute



N4 マシン ファミリー

一般提供

Titanium 上で利用可能な最初の N シリーズ

最新の CPU を搭載

クラウド マーケットで Intel 5th Gen Emerald Rapids CPU を初めて搭載した仮想マシン

コスト パフォーマンスの最適化

Web アプリケーションやバッチ、VDI などの汎用的なユースケースで利用可能なパフォーマンスを低コストで実現

既存マシン ファミリーからの性能向上

N シリーズの前バージョンである N2 と比較して 18% の性能向上

“N4 met my expectations in every way. It was roughly 66% faster than the N1.”

Michael Hurowitz, Software Architect

Bluecore

C4 マシン ファミリー

プライベート プレビュー

継続的に高いパフォーマンスが必要なワークロードに最適

最新の CPU を搭載

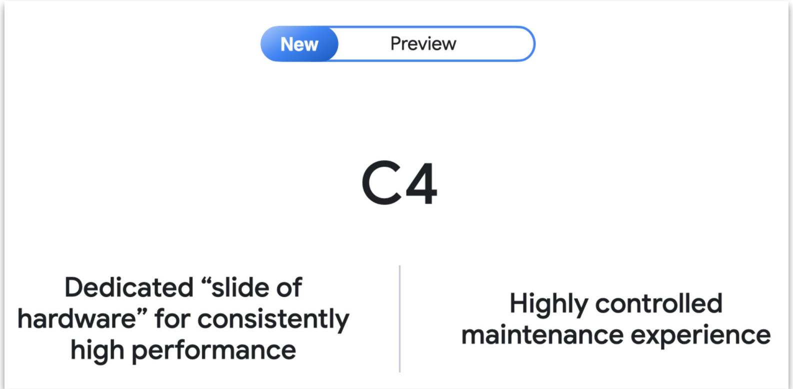
クラウド マーケットで Intel 5th Gen Emerald Rapids CPU を初めて搭載した仮想マシン

継続的に高いパフォーマンスを発揮

最新の CPU による高い処理能力のみならず、Titanium のオフロード技術に最適化することでネットワーク パフォーマンスとディスク アクセスのパフォーマンスが向上

メンテナンス コントロール

メンテナンス通知やメンテナンスの手動トリガーが可能



Z3 マシン ファミリー

初めてのストレージ最適化マシン ファミリー

一般提供

Local SSD によるパフォーマンス向上

最大 36 TB の Local SSD を持つためにディスク アクセスのパフォーマンスが高く、データベースやリアルタイム データ分析など高頻度にデータへアクセスするユースケースに最適

高いネットワーク パフォーマンス

大容量の Local SSD によるディスク アクセスのパフォーマンスのみならず、Tier_1 ネットワーキングにより最大 200 Gbps の下り帯域幅を実現

“Z3 instances match Aerospike's commitment to deliver superior cost performance for real-time database users. For large-scale workloads, customers can run clusters with 70% fewer nodes, significantly, reducing complexity and overall cost.”

Srini Srinivasan, Chief Technology Officer, Aerospace

C3 Metal Machines

ベアメタル上の C3 マシン ファミリー

プライベート プレビュー

ベアメタルだからこその パフォーマンス

仮想マシンではなくベアメタル サーバーであるためにオーバーヘッドを最小限に留めることができ、CPU やメモリを最大限に利用できるためにパフォーマンスの向上に寄与

ハイパーバイザーを自由に選択

ISV パートナーをはじめとしたお客様からのハイパーバイザーを選択したいというご要望に答え、ユーザーが自由にハイパー バイザーが選択可能

NUTANIX　Red Hat Virtualization　replit

+15%

performance over C3 on SIR17

"We're excited about Google's new C3-metal. Nutanix looks forward to working with Google to help organizations get the benefits of public cloud while achieving a universal hybrid operating model with simplified operations and improved TCO."

— Tarkan Maner,
Chief Commercial Officer at Nutanix

X4

メモリ最適化ベアメタル インスタンス

プライベート プレビュー

エンタープライズ アプリで
利用可能

32 TB のメモリとシングル インスタンスで 99.95% の SLA を提供するためにエンタープライズ アプリケーションで安心して利用可能

SAP ECS has been testing and integrating X4 systems into the SAP RISE offering, and we have been impressed with its reliability, performance and ease of management”

Lalit Patil, CTO, *SAP Enterprise Cloud Services*

Google Axion Processors

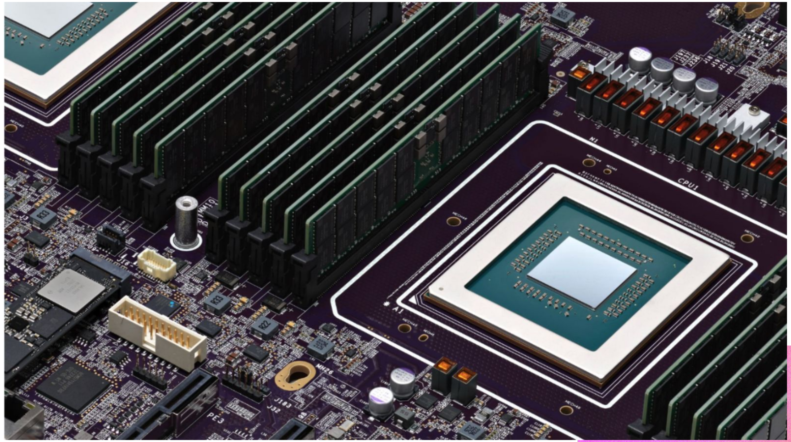
Google が作る初めての Arm ベースの CPU

近日リリース予定

既存の Arm ベース インスタンスよりも 30% 性能が向上

同世代の x86 ベース インスタンスよりも 50% 性能が向上

同世代の x86 ベース インスタンスよりも 60% エネルギー効率が向上



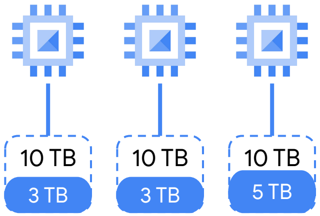
Hyperdisk Advanced Storage Pools

一般提供

TCO の削減を可能とするブロック ストレージ

今までの課題

ブロック ストレージの容量を最適化することは難しく、一般的には必要以上の容量をデプロイするためにコストが増大 (デプロイしている容量の 60% 以上が未使用のお客様も多い)

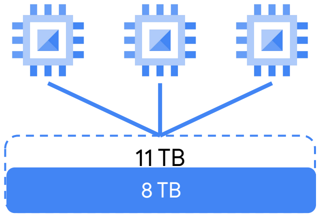


11 TB のみ利用しているにも関わらず 30 TB 分のコストが発生

一般的なブロック ストレージ

解決策

Thin provisioning (*) とデータ削減技術 (**) によってデプロイするブロック ストレージの容量を最適化することで TCO の削減が可能



データ圧縮によって利用容量とプールの容量が削減でき、**11 TB 分のみのコストが発生**

Hyperdisk Advanced Storage Pool

(*) Thin provisioning は Standard Storage Pool でこれまでも利用可能
(**) Advanced Storage Pool で新たに追加された機能

Networking

モデルをサービスエンドポイントとして提供

1 **App Hub**

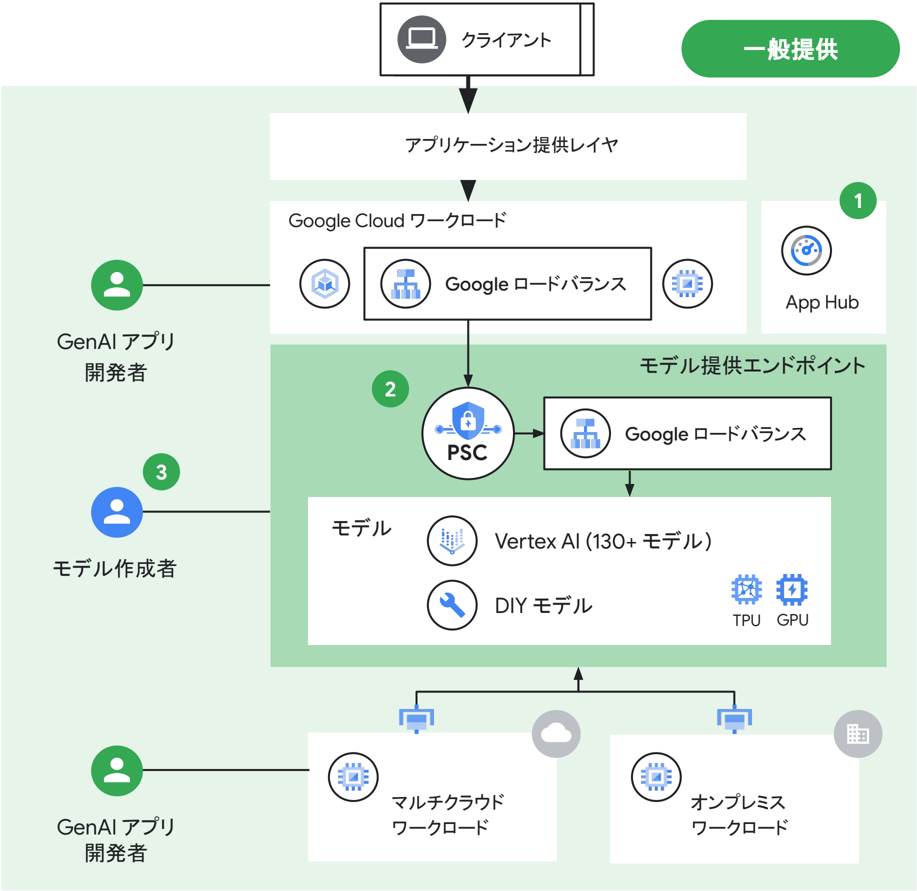
アプリケーションサービスのリストを AI アプリケーションや AI モデルも含め、簡単に使えるように管理

2 **Private Service Connect**

Private Service Connect を通して利用者-提供者形式でモデルにアクセスすることで、クリーンなアーキテクチャを実現

3 **モデル作成者とアプリ開発者の責任分界**

モデル作成者はモデル提供エンドポイントを管理し、アプリ開発者はそのエンドポイントに接続する



推論に最適化されたロードバランス

近日リリース予定

LLM に最適なロードバランス

モデル単位のルーティング

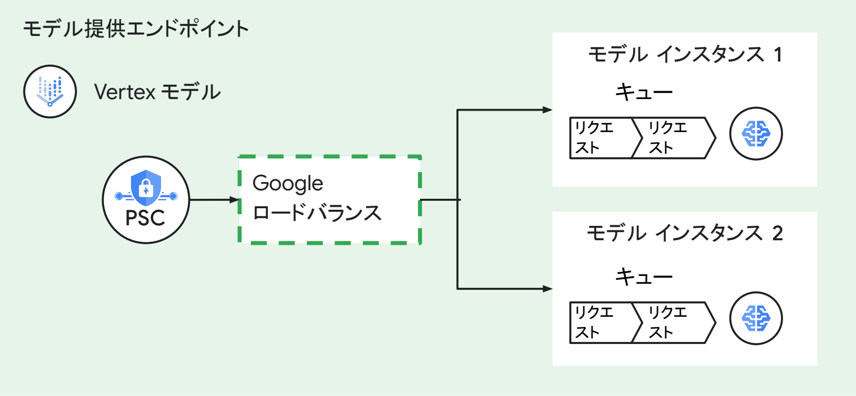
サービスエクステンションによる拡張性

カスタム メトリクスによる負荷分散

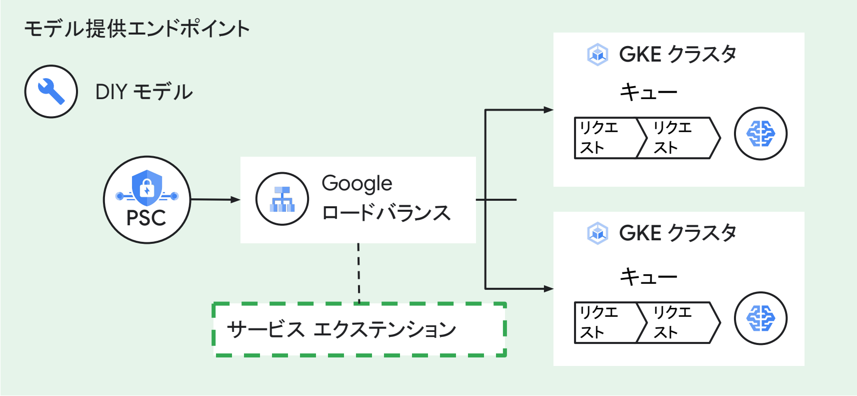
1. アプリケーションがレスポンスヘッダーにORCA (Open Request Cost Aggregation)の値をセット
2. セットされた値に基づきルーティングとスケーリングを実施

例：LLM における**キューの深さ**によってリクエストを分散させる

Vertex に最適化されたトラフィック管理

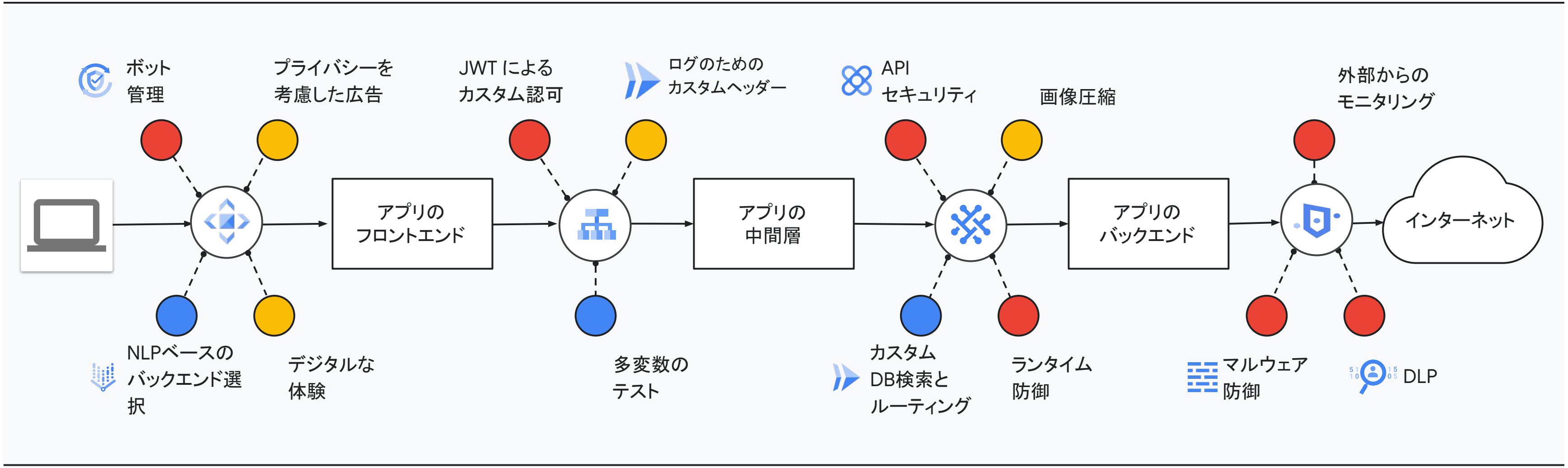


サービスエクステンションを活用した先進的な負荷分散



サービス エクステンションによる オープンでプログラマブルなデータプレーン

データプレーンにサービス(SaaS)や、お客さまによるカスタマイズを導入

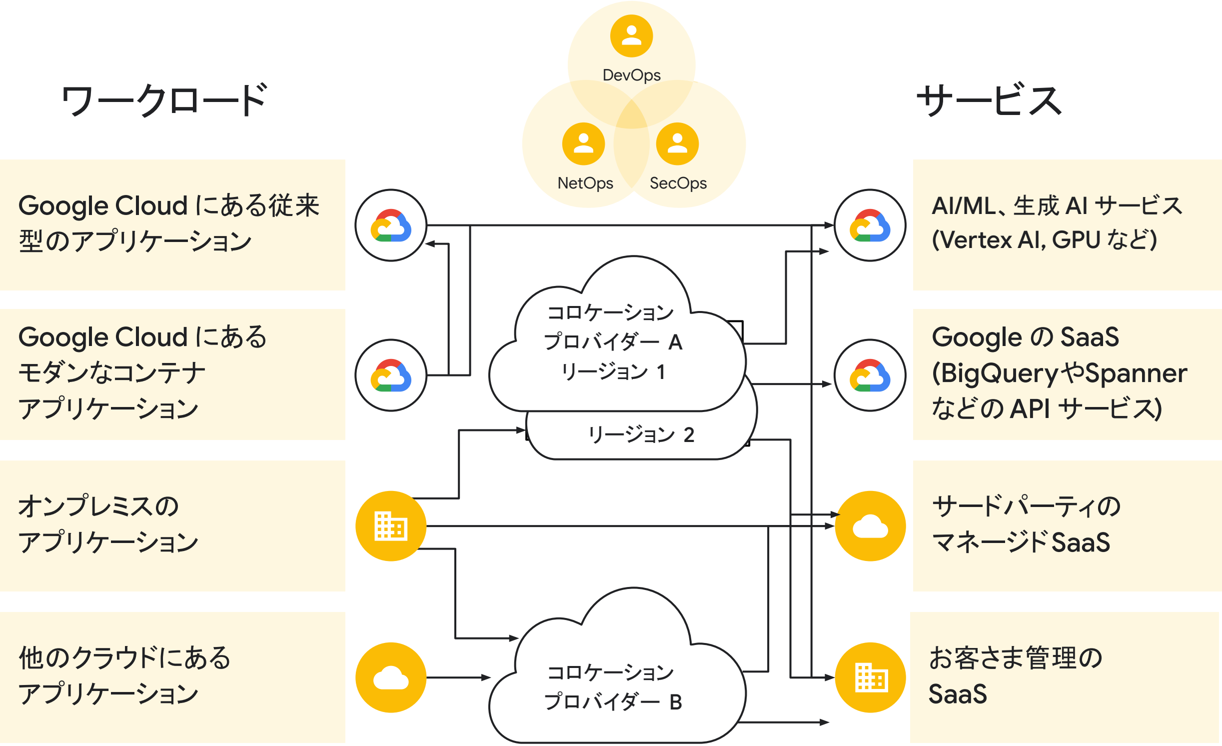


ルーティング サービス
→ バックエンドの選択をカスタマイズ

セキュリティ サービス
→ レイヤ 7 の分析とポリシー管理

トラフィック サービス
→ リクエスト/レスポンスヘッダの変更、ロギングのカスタマイズなど

個別のネットワークによる複雑化



個別に構築されたネットワーク

分断されており、ばらばらのセキュリティ

複雑な運用
NetOps、SecOps そして DevOps チームをまたがる

振り返り: Cross-Cloud ネットワークの 3 つの主なユースケース

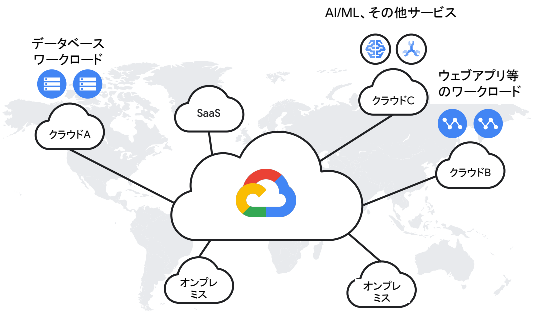
5000 社以上のお客さまにご利用いただいています

オープンな
プラットホーム

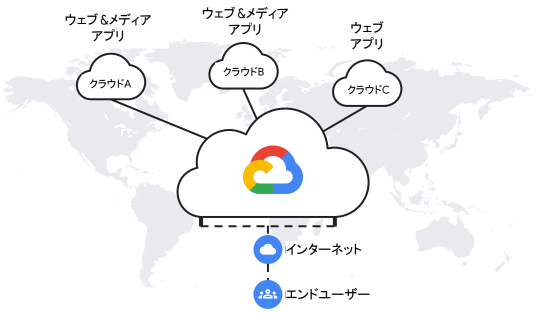
機械学習ベースの
セキュリティ

パフォーマンスに
最適化

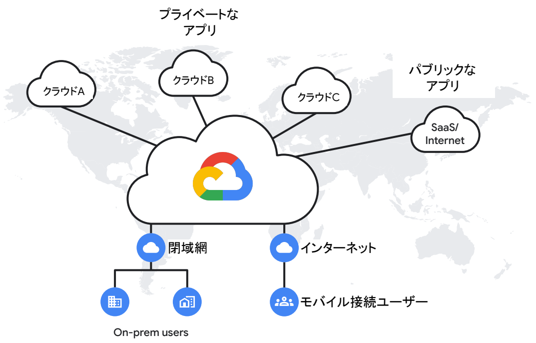
構築: 分散アプリケーション



配信: インターネット向けアプリ・コンテンツ



セキュア: ハイブリッドな利用環境



Google Cloud の地球規模のバックボーン上に構築

Cross-Cloud ネットワークでどんなアプリも安全に

オンプレミスのユーザーも簡単に接続:

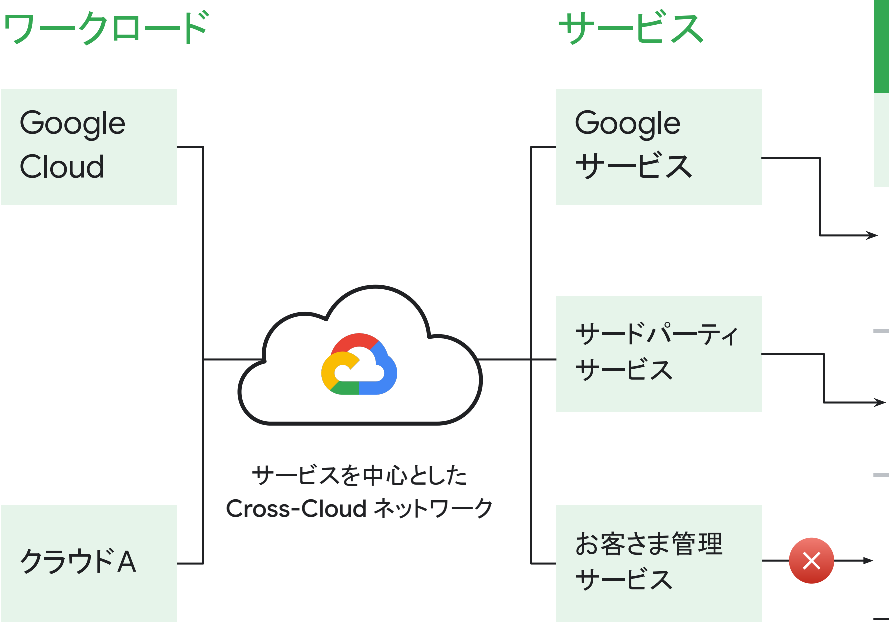
リモート接続のユーザーや、支社からも
Cross-Cloud ネットワークへ簡単に接続可能

**統合的なセキュリティ対策をすべての
ユーザー、デバイスに対して提供:**

サードパーティエコシステムから最適な
ソリューションをお客さまが選択可能

IAM によるゼロトラストアクセス

パブリック プレビュー



アクセスポリシー		
接続元	接続先	アクセス
Google Cloud ワークロード A	Google サービス BigQuery	許可
Google Cloud ワークロードB	サードパーティ MongoDB	許可
クラウド A ワークロード X	お客さまサービス SQL データベース	拒否

サービスを中心とした Cross-Cloud ネットワーク

AI/ML、生成 AI サービス
- Vertex AI
- GPU/TPU

Google Cloud サービス
- BigQuery
- Spanner

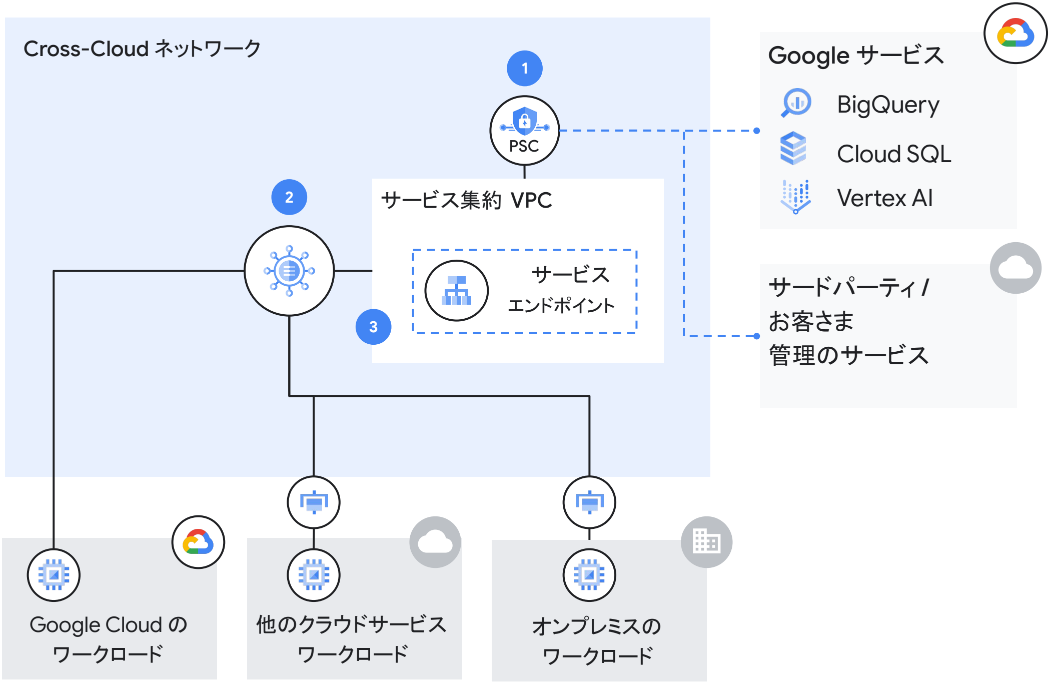
Partner サービス
- MongoDB
- Confluent など

お客さま管理サービス
- SQL DB

ワークロード
- VM ベース
- コンテナベース

- Cross-Cloud Interconnect
- Private service connect
- Cloud NGFW
- Network Connectivity Center
- Cloud Secure Web Proxy
- ロードバランサー

PSC を活用したネットワーク集約



① DevOps

PSC で集約ポイントにサービスを提供

- カナリアリリースのためのトラフィック管理
- サービスのレートリミット プレビュー
- ヘルスチェックによるリージョン間
  フェイルオーバー プレビュー

② NetOps

ワークロードと集約ポイントをシンプルな
ハブ&スポークモデルで接続

- Google Cloud にあるサービスと他クラウドにある
  サービスを抽象化し統一的に扱う プレビュー

③ SecOps

集約ポイントで、ワークロードとサービス間に
統一的なセキュリティ機能を提供

- 機械学習ベースの脅威排除
- サービスアカウント /IDベースの認可 (mTLS) プレビュー
- 通信中データへの Data Loss Prevention (DLP) プレビュー

Application

Gemini Code Assist accelerates application development

Gemini Code Assist の新しい機能

IDE に統合された開発アシスト機能の進化

Repository-wide Context

コード リポジトリ全体のコンテキストを考慮した上で実装をアシスト（Private P

プライベートプレビュー

Code Transformation

コードを分析し、リファクタリング、最適化して変換

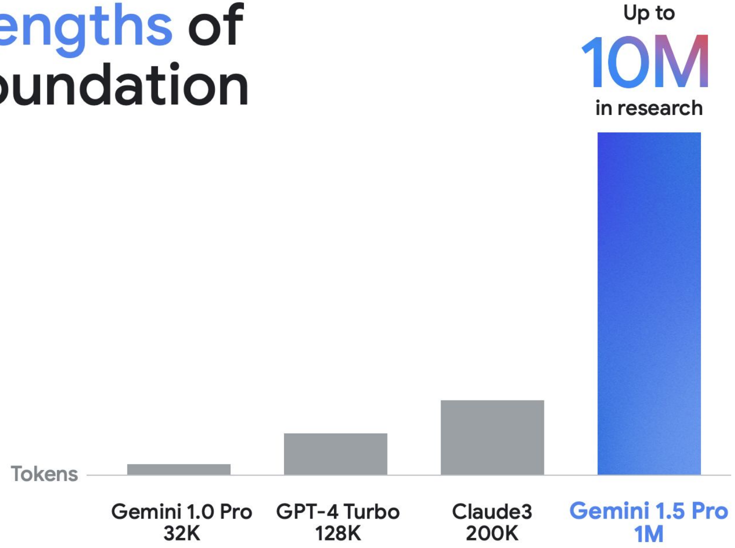
パブリックプレビュー

Code Customization

プライベートのコードベースに基づくコンテキストを考慮したレスポンスに対応

プライベートプレビュー

Context lengths of leading foundation models

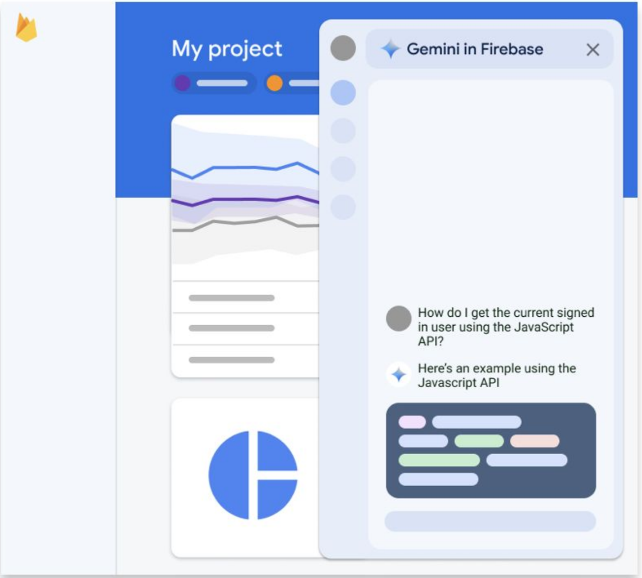


Gemini in Firebase

パブリック プレビュー

Firebase 開発者にも Gemini のサポートを提供

- Firebase の製品と機能に関する質問に回答
- Firebase のコードを生成して説明し、
  開発を高速化
- アプリの品質を向上させるための
  トラブル シューティング サポートを提供
- Firebase で必要な学習コストを緩和



Build GenAI applications to innovate faster

AI アプリケーションを容易にデプロイ

一般提供

LangChain / LangServe でアプリ開発、1 コマンドで Cloud Run にデプロイ

LangChain Support for Cloud Run

LangChain や LangServe を利用して、容易に生成 AI アプリケーションを実装可能

コンテナの知識（Dockerfile）不要

OSS の Buildpacks を利用して、1 コマンドで AI アプリケーションを Cloud Run にデプロイ可能

gcloud run deploy --source=.

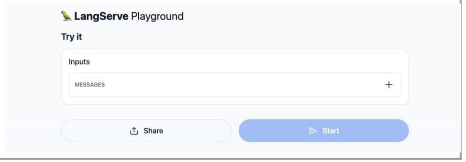
LangServe を利用した例

```
[main.py]
from fastapi import FastAPI
from langchain_google_vertexai import ChatVertexAI
from langserve import add_routes

app = FastAPI()
add_routes(app, ChatVertexAI(model_name="gemini-pro"), path="/chat")

[Procfile]
web: uvicorn main:app --host 0.0.0.0 --port "$PORT"
```

```
$ gcloud run deploy [SERVOCE_NAME] --source=. --region=asia-northeast1
```

Cloud Run エンドポイントにアクセス

指定モデルを利用したチャット インターフェースが存在



Cloud Run application canvas

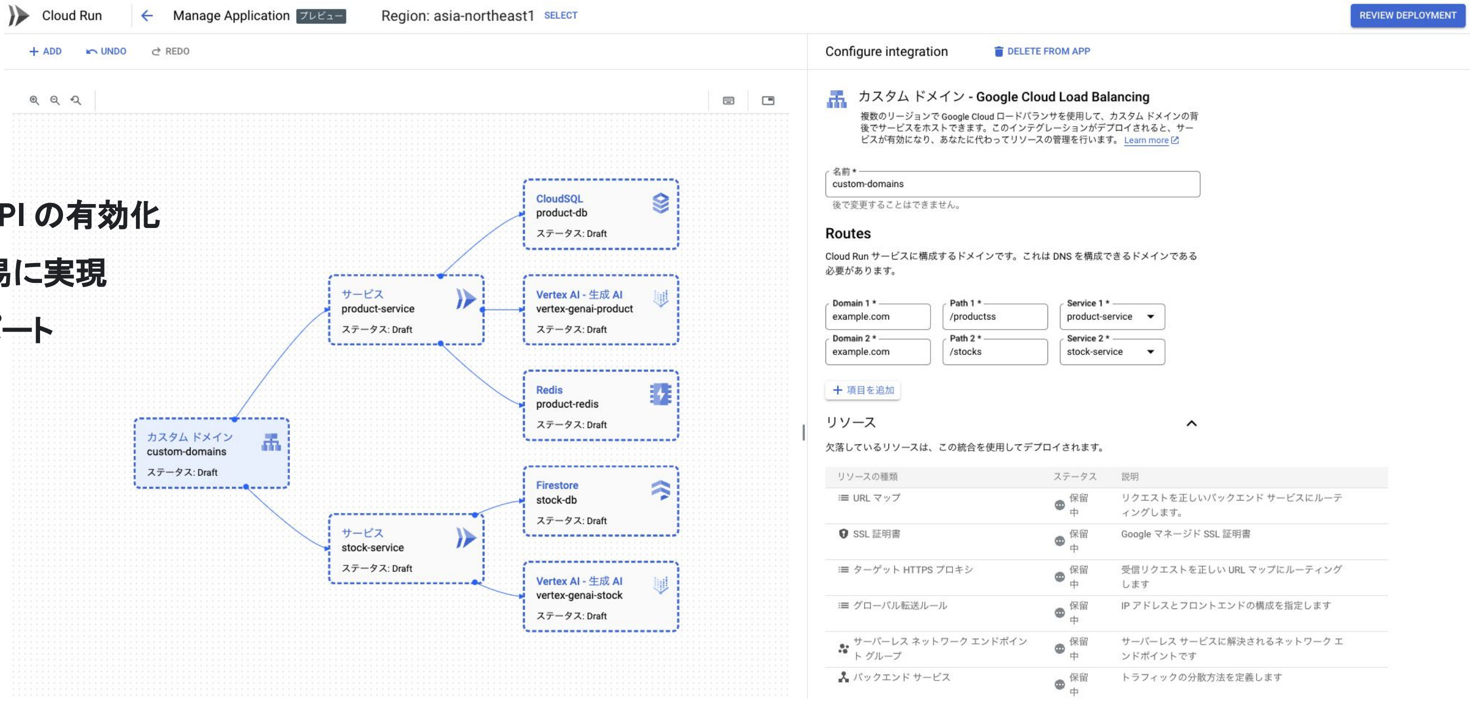
プライベート プレビュー

自然言語や GUI 操作によるプロダクトの統合サポート

リソースの作成、削除、API の有効化

プロダクト間の統合を容易に実現

自然言語による構築サポート

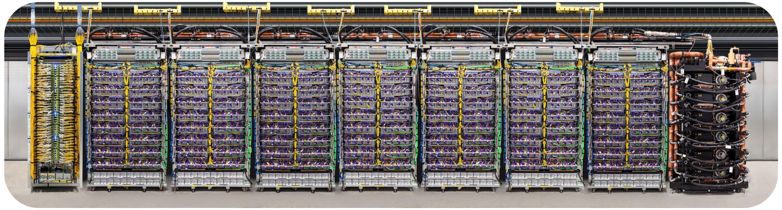


Unlocking AI with a Platform Strategy for the next decade

GKE で TPU v5p が利用可能に

一般提供

高スケールな学習・推論を実現



概要

- GKE Standard / Autopilot で TPU v5p が利用可能に

利点

- 最先端の AI モデル開発期間を数ヶ月から数週間へ短縮
- あらゆるレンジ（サイズ・タイプ）の AI モデルを高速化

2.8X Faster LLM Training
(vs TPU v4)

1.9X Faster Embeddings Training
(vs TPU v4)

Full Range of AI Models
(LLMs, Recsys, Multimodal, MoE)

GKE で A3 Mega が利用可能に

一般提供

帯域幅を増強した NVIDIA H100 GPU 搭載の A3 マシン

概要

- GKE で A3 Mega が利用可能に
- A3 マシンと比べて 2x のネットワーク帯域幅を提供

利点

- 高パフォーマンス、高スループット、低レイテンシーが求められるワークロードに適したマシンを利用可能に

NVIDIA Multi-Process Service (MPS)

パブリック プレビュー

概要

- GKE ノードに接続された単一の物理 GPU を、複数のコンテナが共有できるように
- MPS (Multi-Process Service) は、複数の CUDA アプリケーションがコンテキスト切り替えのオーバーヘッドなしに単一の GPU デバイス上で同時に実行できるように設計されている

利点

- より多くの GPU を使用することで、複数ワークロードのスループットを向上させる
- GPUのアイドル状態による無駄を削減できる
- 必要に応じて、異なるプロセスにより細かい単位でGPUリソースを割り当てることができる

Deploy workloads using MPS

```yaml
apiVersion: batch/v1
kind: Job
metadata:
  name: nbody-sample
spec:
  completions: 3
  parallelism: 3
  template:
    spec:
      hostIPC: true
      nodeSelector:
        cloud.google.com/gke-gpu-sharing-strategy: mps
      containers:
        - name: nbody-sample
          image: nvidia/samples:nbody
          command: ["/tmp/nbody"]
          args: ["-benchmark", "-i=5000"]
          resources:
            limits:
              nvidia.com/gpu: 1
      restartPolicy: "Never"
  backoffLimit: 1
```

GCS FUSE Read Cache サポート

一般提供

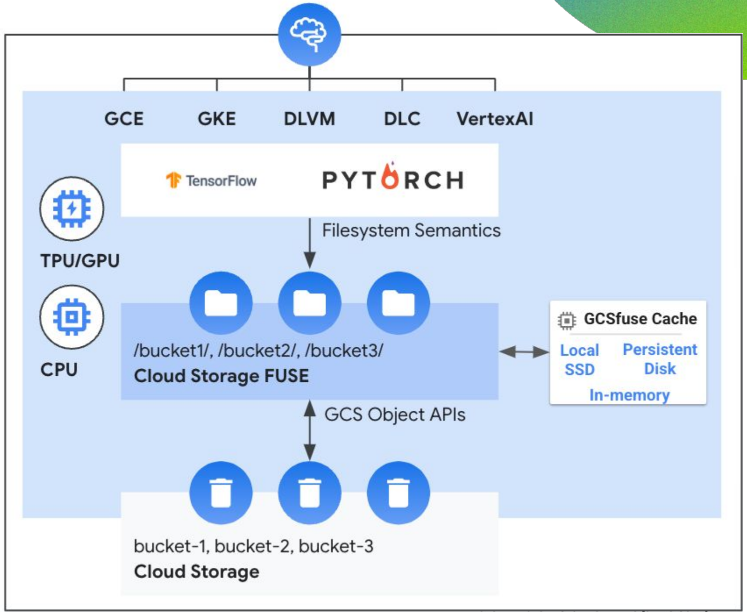
オブジェクト ストレージ内のデータへのアクセスを高速化

概要

- GKE で GCS Fuse CSI ドライバを使って、Read Cache 機能を利用することが可能に

利点

- 一度 GCS からダウンロードしたデータを Persistent Disk、Local SSD、メモリにキャッシュすることで同じデータへの繰り返しアクセスによって生じるオーバーヘッドが削減可能



Kueue サポート

パブリック プレビュー

概要

- Kubernetes ネイティブなジョブ キューイングの GKE アドオン
- Dynamic Workload Scheduler (DWS) とネイティブに統合
- マルチクラスタでのジョブキューイングをサポート

利点

- キューの優先順位付けと公平なリソース共有による、汎用性の高いジョブキューイングおよびスケジューリング
- クラスタ使用率の最適化によるコスト効率の向上

Parallelstore サポート

プライベート プレビュー

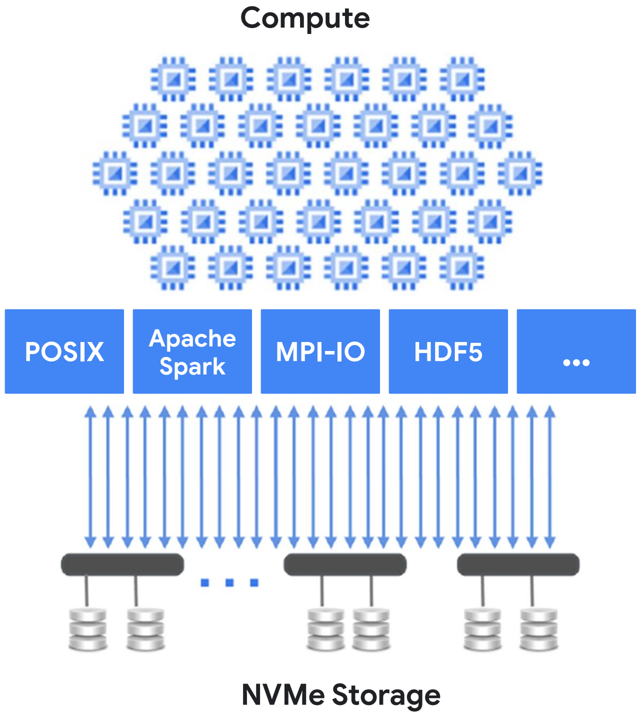
AI/ML に最適化されたストレージ ソリューション

概要

- AI/ML に最適化され、完全にマネージドされたストレージソリューションを CSI ドライバ経由で利用可能に
- 高いスケーラビリティと低遅延のデータアクセスを必要とするハイパフォーマンスなアプリケーションの処理速度を向上

利点

- 超低遅延、高パフォーマンス
- 分散メタデータ管理、極めて高い IOPS、Key Valueアーキテクチャを備えた AI に最適化された構造
- GKE マネージドの CSI ドライバを提供し、コンテナワークロードから簡単に利用可能



Container Image Preloading

AI ワークロードのコールドスタートを 30% 短縮

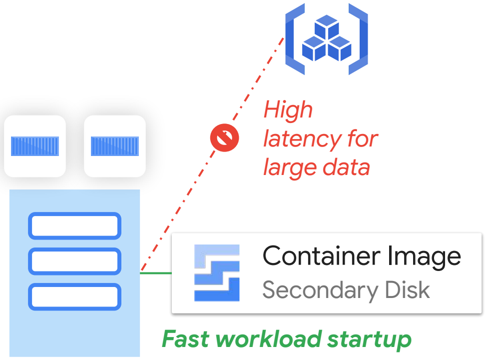
パブリック プレビュー

概要

- GKE Node Pool にセカンダリディスクを追加することが可能に
- コンテナイメージやデータを Node 上に Preload し、AI/ML ワークロードなどサイズの大きいコンテナの立ち上げ速度を向上

利点

- コンテナイメージをセカンダリディスクとしてマウントすることにより、Node の動的プロビジョニング時も Image Pull が不要となりコンテナの立ち上げ速度が向上
- コンテナに必要なライブラリ等をセカンダリディスクに格納することにより、コンテナイメージ自体のサイズを削減することも可能



GKE Threat Detection

脅威検知の継続的なモニタリング

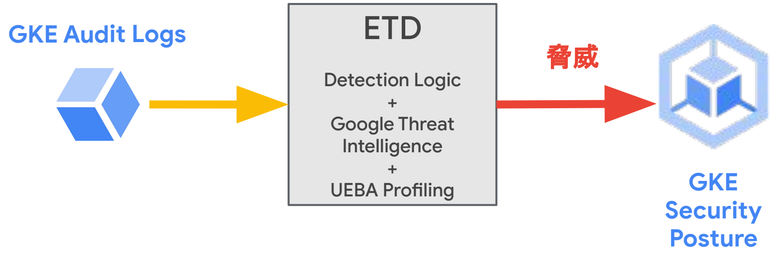
パブリック プレビュー

概要

- GKE Audit Log を基に、GKE クラスタ上の脅威を検出
- 検出された脅威は Security Posture ダッシュボード上で確認可能

利点

- 複数クラスタ環境に継続的にスキャンを実行しアクティブな脅威を迅速に検出
- 検出された脅威は MITRE ATT&CK フレームワークにより分類される
- Security Command Center の Premium Tier を利用している場合は、SCC でも脅威情報を確認可能



GKE Compliance

パブリック プレビュー

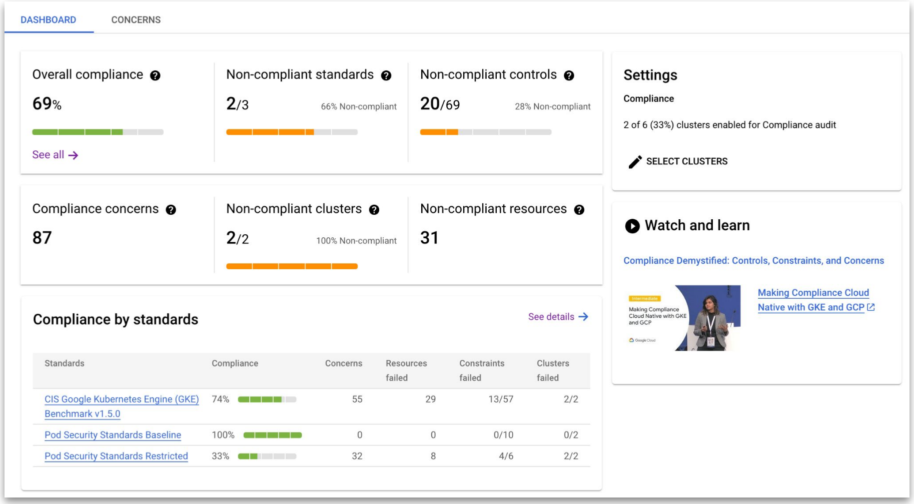
継続的なコンプライアンス遵守の可視化とレポーティング

概要

- GKE クラスタの業界ベンチマークやセキュリティ スタンダードの遵守状況を継続的にモニタリングするダッシュボードを提供

利点

- クラスタ構成からコンテナ ワークロード構成まで包括的なコンプライアンス評価を実施
- 複数クラスタのコンプライアンス状況をダッシュボードで一元的に管理
- 対象クラスタのコンプライアンスレポートを自動的に生成



AI

Google Cloud AI

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

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

Vertex AI 開発者向け エンタープライズレディな生成 AI

Google および 業界内の
最高のモデルを提供

レベルを問わない
エンドツーエンドのモデル開発プ
ラットフォーム

より早くエージェントを開発、デプ
ロイし、
エンタープライズの信頼性ある回
答をグラウンディングで提供

エンタープライズのレディネスの
基礎を構成

Vertex AI 開発者向け エンタープライズレディな生成 AI

Google および 業界内の最高のモデルを提供	レベルを問わないエンドツーエンドのモデル開発プラットフォーム	より早くエージェントを開発、デプロイし、エンタープライズの信頼性ある回答をグラウンディングで提供	エンタープライズのレディネスの基礎を構成
Gemini 1.0 Pro 一般提供 Gemini 1.5 Pro パブリックプレビュー Imagen 2.0 編集 | Live Images | デジタル透かし Gemma: CodeGemma と RecurrentGemma Hugging Face 統合	プロンプト マネジメント Gen AI Evaluation	Grounding on Google Search パブリックプレビュー Vertex AI Agent Builder Retrieval Augmented Generation (RAG): DIY	デジタル透かし データ レジデンシーと拡大されたリージョン対応 +ML処理
Vertex Embeddings モデルのアップデート Claude 3: Haiku | Sonnet Claude 3: Opus (近日リリース予定) Chirp 2	Supervised Tuning for Gemini Pro Feature Store 2.0 Dataplex Lineage による Pipelines Integration	LangChain on Vertex AI Extensions Function Calling Search Primitives Connectors	Responsible AI (RAI) Tooling モデルアクセスのための Org policy (IAM)

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 | Training | Feature Store | Pipelines | Monitoring

Vertex AI Model Garden

Google | Open | Partner

Model Garden : Gemini モデルのアップデート

最大 100 万トークンの入力が可能なマルチモーダル言語モデルにより
あらたな可能性を開く

Gemini 1.5 Pro パブリック プレビュー

最大 100万トークンの入力が可能、様々なベンチマークで 1.0 Pro を上回り、1.0 Ultra に遜色ないパフォーマンスを発揮。ビデオ、画像、テキストに加えて、ビデオ内オーディオ、PDF が input として利用可能に。

Gemini 1.0 Pro 一般提供

Gemini 1.0 Pro (Gemini-1.0-Pro-002) が 一般提供を開始。より品質が向上。バージョニングポリシーが適用され、安心して商用環境で利用いただけるようになったほか、生成物の著作権侵害に対する補償にも対応。

Gemini

Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context

Gemini 1.5 Pro	Relative to 1.0 Pro	Relative to 1.0 Ultra
Long-Context Text, Video & Audio	from 32k up to 10M tokens	from 32k up to 10M tokens
Core Capabilities	Win-rate: 87.1% (27/31 benchmarks)	Win-rate: 54.8% (17/31 benchmarks)
Text	Win-rate: 100% (13/13 benchmarks)	Win-rate: 77% (10/13 benchmarks)
Vision	Win-rate: 77% (10/13 benchmarks)	Win-rate: 46% (6/13 benchmarks)
Audio	Win-rate: 60% (3/5 benchmarks)	Win-rate: 20% (1/5 benchmarks)

Table 1 | Gemini 1.5 Pro compared to Gemini 1.0 family. Gemini 1.5 Pro maintains high levels of performance even as its context window increases. Detailed results are presented in Table 7.

Model Garden : Google モデルのアップデート

Imagen 2 (画像生成、編集)、Embeddings API をアップデート
より高機能・品質を向上

Imagen 2 一般提供

テキストからの画像生成

テキストからの Live Image の生成

編集機能の向上 (Inpainting / Outpainting / Product Image Editing)

デジタル透かし (powered by Google Deepmind SynthID)

ユーザー定義の Safety Filter (子供、人間の顔等)

Embeddings API 一般提供

`text-embedding-preview-0409`
`text-multilingual-embedding-preview-0409` を追加し
Retrieval および 分類タスクの性能を向上。

DeepMind の論文:
https://deepmind.google/research/publications/85521/

Frog croaking joyfully in a puddle during a rain shower.

Display bamboo steamer baskets opening to reveal perfectly steamed dumplings, with steam rising and showcasing the delicate pleats of the dumplings.

Model Garden : オープン / パートナーモデル

様々なモデルが日々開発・公開される今日、最先端のモデルを **選択** できるプラットフォーム

オープン モデル

- Google DeepMind のオープンモデル Gemma が一般提供
- Code Gemma と RecurrentGemma を追加
- Mistral 7B, Mixtral 8x7B
- Grok-1
- Hugging Face から 1 クリックでのデプロイ

パートナーモデル

- Anthropic の Claude 3.0 を追加
    - Haiku **一般提供**
    - Sonnet **一般提供**
    - Opus **近日リリース予定**

オープン モデル / パートナー モデル

Gemma

Google の新たにリリースした軽量オープンモデル
7B / 2B パラメータサイズにおいて最高の性能を発揮

Hugging Face 連携

数クリックで Vertex AI でオープンモデルをすぐ利用可能

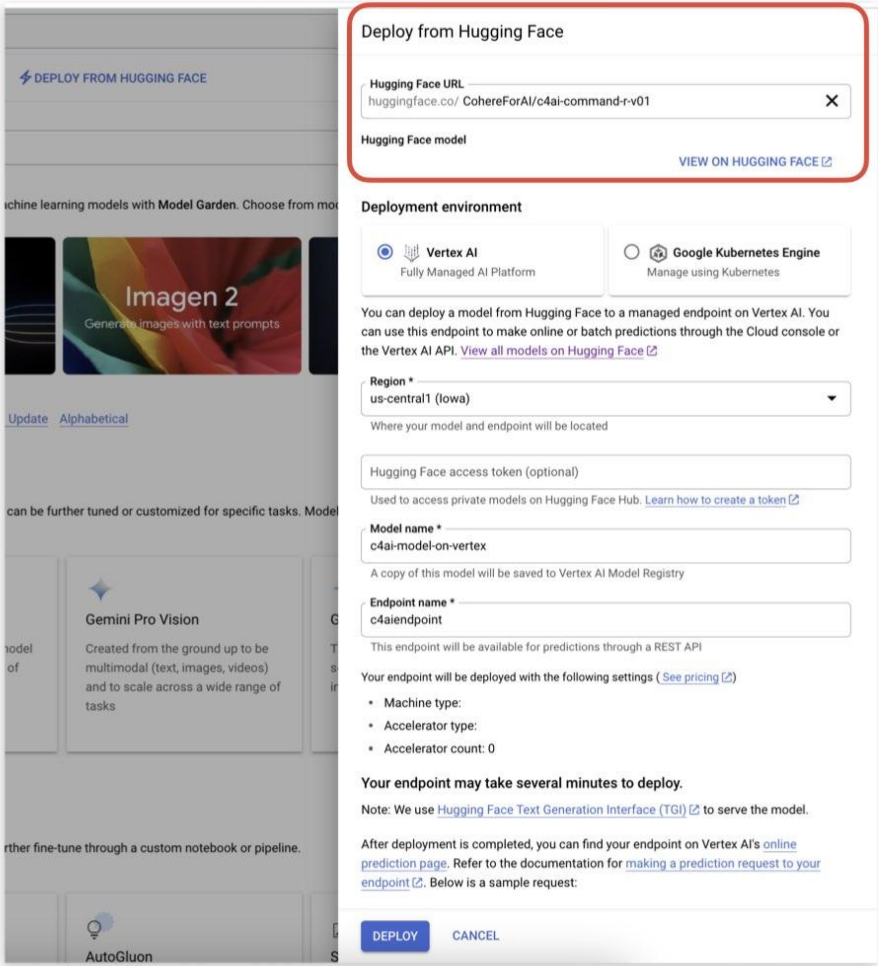
Claude on Vertex AI

Claude 3.0

MISTRAL AI_ Meta

Open Models on Vertex AI

Mixtral 8x7B, Grok, LLaMA-2 や DITO など多数のオープンモデル



Vertex AI

Hugging Face

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 | Training | Feature Store | Pipelines | Monitoring

Vertex AI Model Garden

Google | Open | Partner

Model Builder : Supervised チューニング

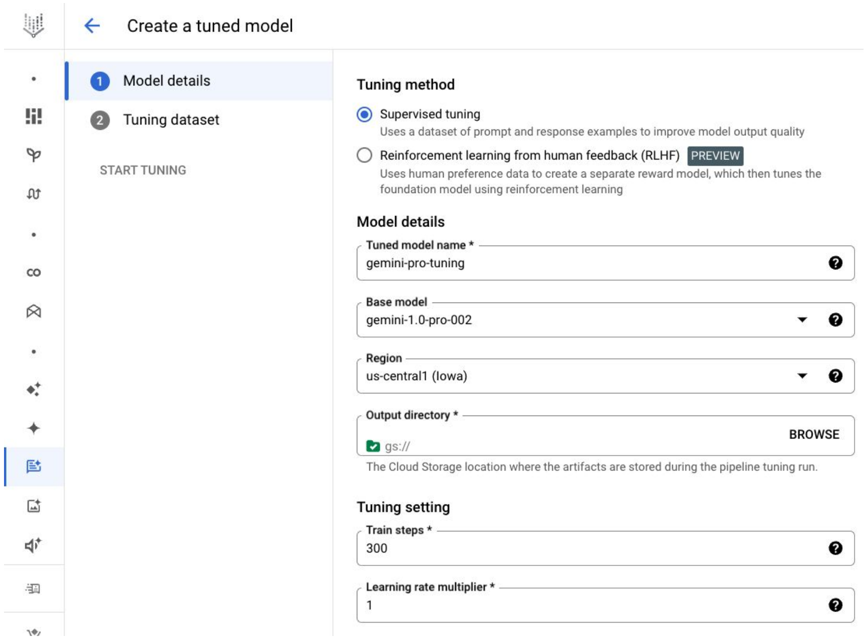
パブリック プレビュー

モデルをカスタマイズ、使いこなす機能が Gemini でもサポート

Gemini モデルの Supervised チューニング

- アダプティブ レイヤーのチューニングが Gemini モデルでも利用可能
- 数百から数千のデータを持ち込むことで、特定のタスクに強いカスタマイズ モデルを獲得

NOTE: Gemini 1.0 Pro (テキストデータ) のみ利用可能



Model Builder : プロンプト管理

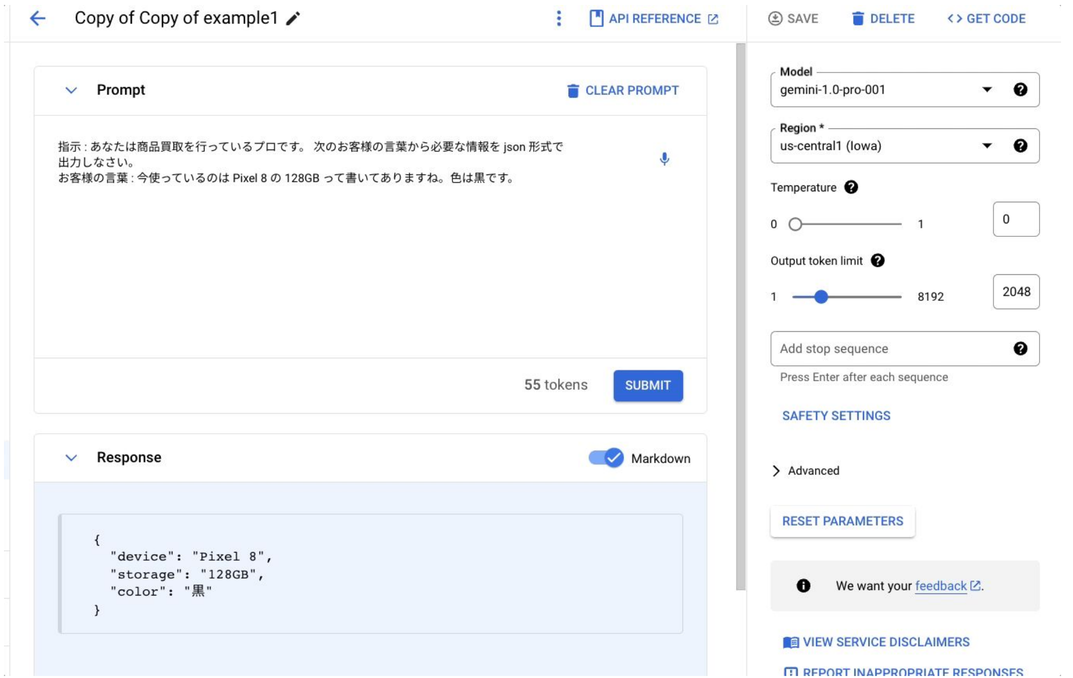
プロンプト デザインをより効率的かつ効果的に行えるツールの提供

新機能 パブリック プレビュー

- **Prompt Version Management**: プロンプトのバージョン管理（パラメータ含む）
- **Prompt Notes**: 個々のプロンプトにノートが残せるようになり、柔軟な管理が可能
- **Side-by-Side Comparison**: コンソール上で生成結果を比較する機能を提供
- **Rapid Evaluation**: Ground truth を持ち込むことで、定量的な評価を迅速に返す機能

近日リリース予定の機能

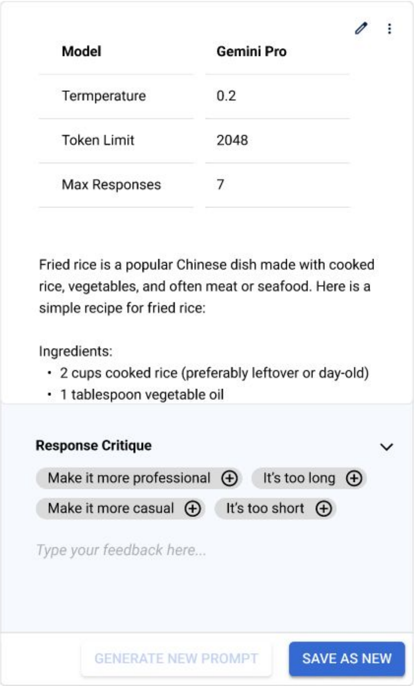
- **Prompt Tags**: タグによって、より柔軟にプロンプトを管理
- **AI-Assist**: より良いプロンプトをデザインするため、AI がフィードバックを返す機能

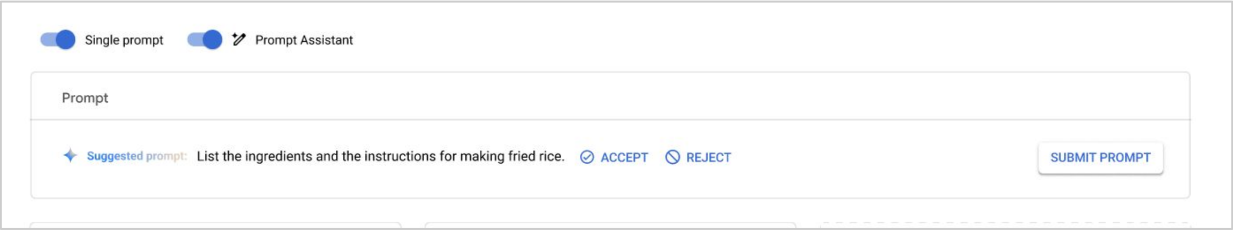


End-to-end model building platform with choice at every level

AI-Assist

近日公開予定





プロンプト改善のための評論を提供

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

AI-Assisted Prompting

近日公開予定

最初のプロンプト

Multi-choice problem: What is the category of the object?
- category_1
- category_2
- category_3
Text: pineapple
The answer is: category_1
Text: sky
The answer is: category_2
Text: car
The answer is: category_3
Text: {var}
The answer is:

入力

banana

出力

category_1

評論

It doesn’t provide reasoning for the chosen category.

直したプロンプト

Multi-choice problem: What is the category of the object?
- category_1
- category_2
- category_3

Text: pineapple
The answer is: category_1
Reasoning: Pineapples are a type of fruit.

Text: sky
The answer is: category_2
Reasoning: The sky is a part of the atmosphere.

Text: car
The answer is: category_3
Reasoning: Cars are a type of vehicle.

Text: {var}
The answer is:

新しい出力

Category_1
Reasoning: Bananas are a type of fruit.

Model Builder : モデル評価

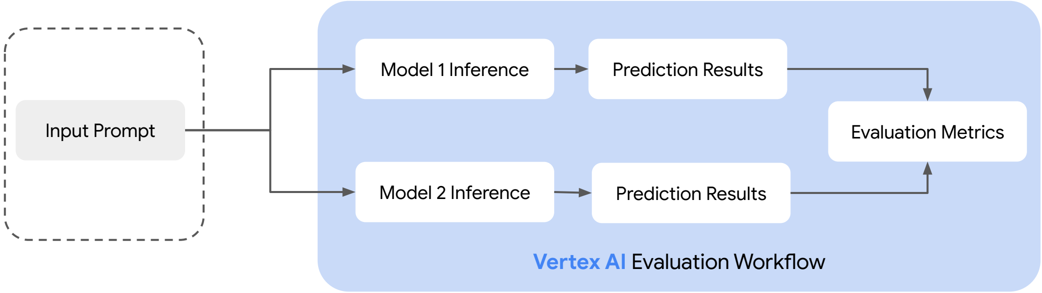
プロンプト デザイン、モデル選定において定量的な評価による意思決定が重要

AutoSxS (オフライン) 一般提供

- 異なるモデル、異なるバージョンのモデルを、Ground truth を用意せずに定量的に 評価する機能

Rapid Evaluation パブリック プレビュー

- 少量のデータセットにより、迅速な評価が可能になる機能
- プロンプト デザインのような探索的なプロセスに有用な機能



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 | Training | Feature Store | Pipelines | Monitoring

Vertex AI Model Garden

Google | Open | Partner

Vertex AI Agent Builder

人間のような
会話とやりとり

会話を制御し、形作る

グラウンディングされた
レスポンスにより
信頼を確保

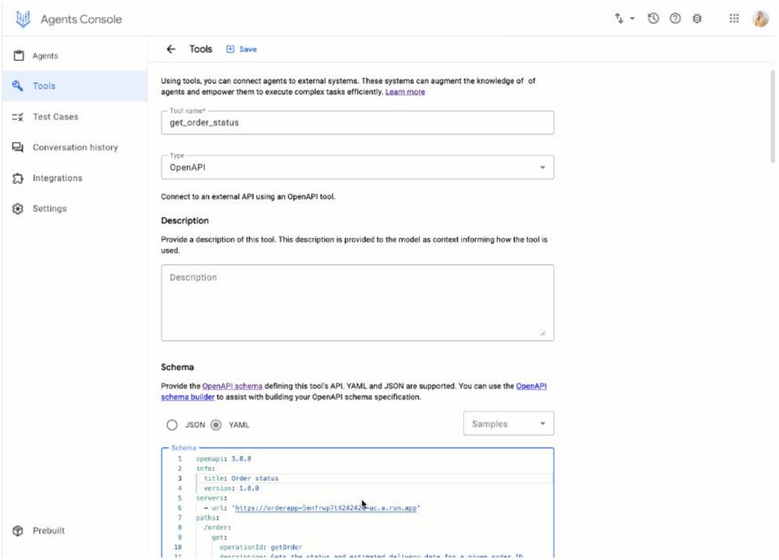
Agent Builder: Playbook Agent

パブリック プレビュー

Agent の会話制御フローを生成 AI で生成

Playbook Agent

- 生成 AI のフローを制御し思い通りのエージェントを作成、安全なエージェントを構築
- エージェントの構築時間を短縮し、メンテナンスの効率化を目指した機能



Vertex AI Agent Builder

エンタープライズ レディな エージェント開発サービス

オーケストレーション

エージェントをあらゆる規模で作成、公開、管理

Vertex AI Agent Builder ではコーディングのレベル（ノーコード[1]、ローコード、コーディング ）をに依らずに構築

LangChain on Vertex を使ってエージェントをオーケストレーション、デプロイする

パブリックプレビュー

グラウンディングと拡張

生成 AI の出力の関連性と精度を向上

最新の高品質な情報にアクセスするための **Ground with Google Search**[3]

パブリックプレビュー

Vertex AI Search によってお客様のデータをグランディングした Out-of-the-box な RAG を実現

一般提供

アクションを起こす

LLM を外部ツールに接続し、API やサービスをコールする

Function Calling で独自のアクションを作成し、API にアクセス

一般提供

Extensions[2] で事前に構築されたモジュールにアクセス

パブリックプレビュー

[1] 自然言語を使用（Public Preview）
[2] Code Interpreter と Vertex AI Search が利用可能

3.Allowlistが必要
https://cloud.google.com/vertex-ai/generative-ai/docs/grounding/overview
Allowlist :
https://docs.google.com/forms/d/e/1FAIpQLSfi7SJNAqXOe7vB9Re-WtwJedTWjHkBKZ6OHaLoGoXRG93ymg/viewform

Infrastructure

Agent Builder: Grounding on Google Search

パブリック プレビュー
With Allowlist

Google 検索とのグラウンディングしたモデルのレスポンスにより、
最新で、高品質な情報でレスポンスの精度を改善

モデルの返答を事実に即させる

- LLM のレスポンスをGoogle 検索の信頼できる結果に接続し、一般知識や公開された事実に根拠付を行う

ソースデータの提供でハルシネーションを防ぐ

- グラウンディングないで引用とアトリビューションを返し、その情報がどこからきたのかを確認し、ハルシネーションであることを簡単に確認可能

一切の開発なくマネージド サービスとして利用可能

- 異なるデータソースに対して、シンプルなコンフィグが可能

Google 検索グラウンディング はLLM の回答にたいして、**ソースのリンク** と **検索結果の推奨** を提供することで、ユーザーに LLM のレスポンスが正しいかを確認し、情報の検索ジャーニーを手助け

- **Search Grounding**: インラインの **ソース リンク** をタップすることで、ソースの Web サイトを開ける
- **検索の推奨**: Google 検索の推奨結果を開くことが可能

https://cloud.google.com/vertex-ai/generative-ai/docs/grounding/overview
別の規約への Opt-inが必要となります。
Allowlist :
https://docs.google.com/forms/d/e/1FAIpQLSfj7SJNAqXOe7vB9Re-WtwJedTWjHkBKZ6OHaLoGoXRG93ymq/viewform

Agent Builder : Grounding on Vertex AI Search

一般提供

エンタープライズ データとグラウンディング

エンタープライズ データを含んだ
レスポンスを生成

- お客様の接続したエンタープライズデータを
  グラウンディングする統合された生成AI プラットフォーム
- LLM のレスポンスの品質向上と事実に基づく応答の生成が可能に

Citation によりファクトチェックが可能

- 結果を見つけたソースへのリンクをCitation（引用）として提示。ユーザーがファクトチェックを可能。

[1] Requires Opt-In

Agent Builder : Function Calling 一般提供 と拡張 パブリックプレビュー

拡張(Extension)により、Gemini がモデルの外部システムと連携可能となり、エージェントを素早く実用化する

エージェント型のアプリケーションを素早く構築

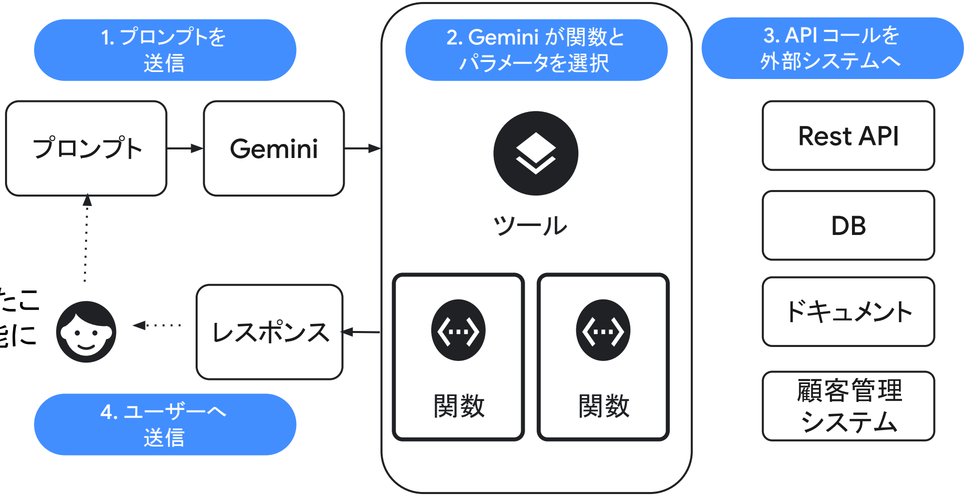
- Extensions によりモデルが外部システムと連携可能に
- Extensions の例
  - Code Interpreter
  - Vertex AI Search Extension
- Function Calling により LLM から関数を定義可能に

Function Calling の拡張 一般提供

- Gemini 1.5 Proモデルで Function Calling が利用可能になったことで、より大きなコンテキスト (最大 100 万 Token) が利用可能に

Function Calling Constrained Decoding

- 出力の構造を保証(e.g. JSON)



Agent Builder : LangChain on Vertex AI

パブリック プレビュー

関数、パラメータを選択する際の推論を最もポピュラーな LangChain で定義、
フルマネージドサービスとしてデプロイ可能に。

エンドツー エンドのエージェント型
アプリケーション開発のフレームワーク

- LangChain を用いた Vertex AI 上でのエージェント側アプリケーションの開発、デプロイ、サービングのライフサイクルとをエンドツーエンドでサポート

Vertex AI エコシステムとの連携による
エンタープライズ レディなエージェントの提供

- Function Calling や拡張をはじめとするVertex AI の様々なエコシステムとの連携
- セキュアでエンタープライズレディなエージェント開発

1. 関数を定義

```
def get_exchange_rate(
    from: str = "USD",
    to: str = "EUR"):

    "Retrieves the exchange rate."

    response = requests.get(
    "https://api.currency.app/",
      params={"from": from,
              "to": to})

    return response.json()
```

2. LangChain のテンプレートを利用

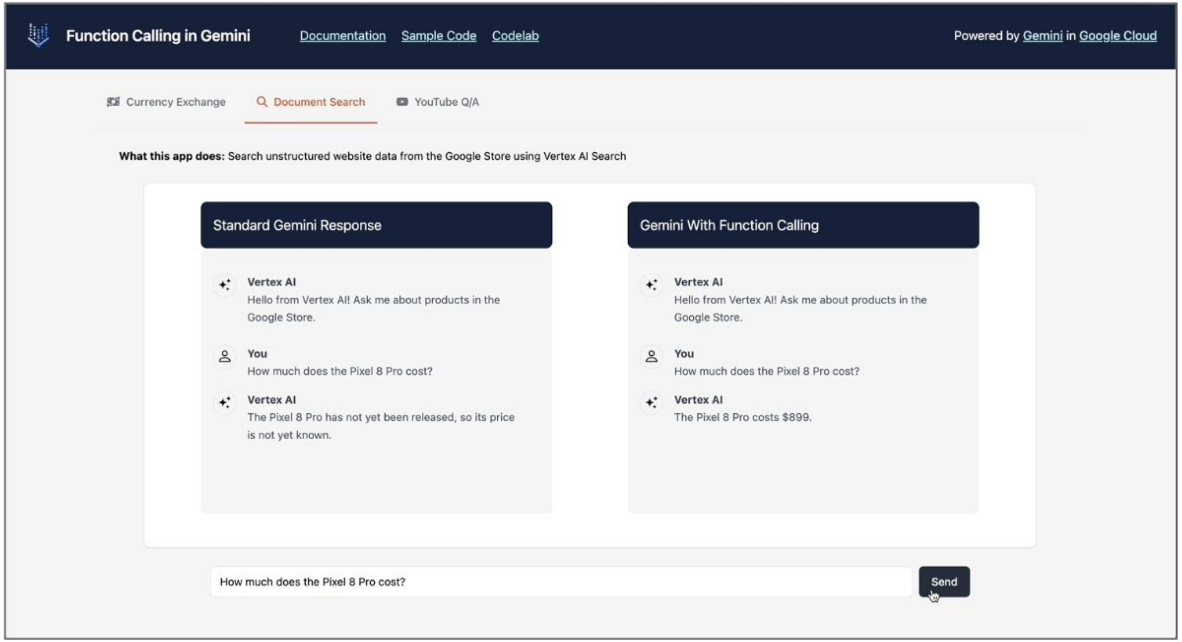
```
app = llm_extension.LangchainAgent(

      tools=[get_exchange_rate],

      model_kwargs={
       "temperature": 0.3,
       "top_p": 1,
       safety_settings: {...},

     }
)
```

3. Vertex AI へデプロイ、マネージドなサービスとして利用

```
remote_app =
reasoning_engines.ReasoningEngine.create(
    LangChainAgent(),
    requirements=[
      "google-cloud-aiplatform",
      "langchain",
      "requests==2.*"])

remote_app.query(
  query="What's the exchange rate
  from US dollars to Swedish
  currency?")
```

Demo

Reasoning Engine を使った Retrieval Augmented Generation（RAG）



Agent Builder: DIY Search and RAG

RAG , Search の仕組みを自分で作る場合の構成要素で Vertex AI Search の仕組みを
マネージド サービスとして提供。必要なものだけ自分で使って組み込みが可能に。

Check grounding API パブリックプレビュー

- 事実を探すのではなく事実の度合いを判定するAPI
- リファレンスからの引用(Citation) の生成、RAGシステムからのレスポンスの信頼性を評価

Ranking API パブリックプレビュー

- 意味の近さでリランキングすることで RAG 全体の精度を改善
- クエリに対してより精度の高い rerank を実行しセマンティック検索をドキュメントに対して実施

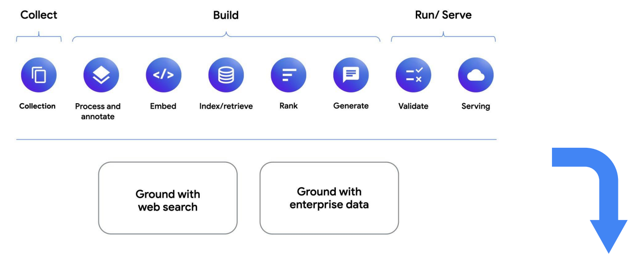
Document AI Layout Parser パブリックプレビュー

- 様々なフォーマットのドキュメントを、その内容やコンテキストを加味しながら構造化データへと変換

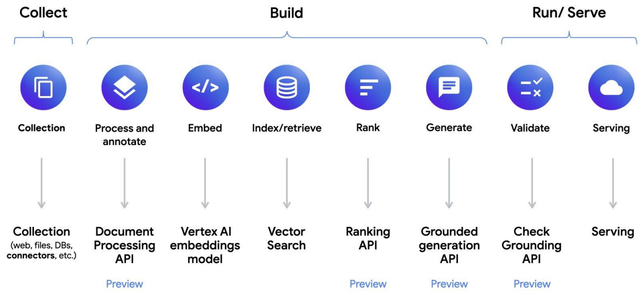
Grounded generation API プライベートプレビュー

- グラウンディングしたレスポンスを生成。Vertex AI Search やカスタムデータを利用可能

Grounding in Vertex AI



Do-it-yourself RAG systems



Agent Builder: Extensions

エージェントを構築する際に、実装をより迅速化するために
用意されたモジュール

Code Interpreter 一般提供

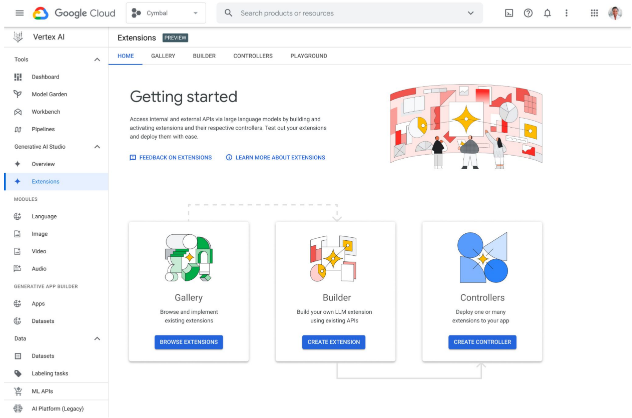
- データ分析・可視化、算術計算など、即座に LLM がタスクを実行できるようなモジュール

Vertex AI Search Extension パブリック プレビュー

- Vertex AI Search で構築されたインデックスを使い、独自データへのグ ラウンディングを実装する機能
- Search 体験を Agent の一機能として組み込むことで、より広範なシナリオに対応できるようになる

BigQuery Extension 近日リリース予定

- BigQuery Vector Search で構築されたインデックスを使い、独自データへのグランディングを実装する機能



Agent Builder: Vertex AI Search

様々なアップデート

Connector

- Google Drive 一般提供
- Bigtable, Spanner Cloud SQL, Firestore

Media Search

- Media コンテンツに特化した Vertex AI Search
- 日本語のサポートが追加
- Search-as-you-type 機能が追加され、クエリの補完が可能に

Blended Search 一般提供

- 複数の検索ソースをまとめて検索が可能に

構造化データストア

- 一般提供開始
- Geolocation, datetime をサポート

サーチ結果のブースト パブリック プレビュー

- Media コンテンツに特化した Vertex AI Search
- 日本語のサポートが追加
- Search-as-you-type 機能が追加され、クエリの補完が可能に

Answers with summaries and follow-ups パブリック プレビュー

- 検索時のサマリとフォローアップの方法をよりよい結果で返せるように

Data Analytics

BigQuery

全てのデータ分析を実現する統合プラットフォーム

600%

Object Tables を活用した画像、音声、ビデオ形式を含む非構造化データ管理の成長

400%

BQ Migration services の成長

350+

新しい製品ケイパビリティ-
前年比 60%+

54%

マーケットプロダクトと比較してより安価な **TCO**

データ アナリティクスのイノベーション

Tier one launches at Next '24

01

データからAIまで “One unified” プラットフォーム

BigQuery Studio (GA) データ分析、データ エンジニアリング、ML 向けの単一インターフェイス

BigQuery metastore (Preview) SQL、オープンソース エンジン、AI ML 用のフルマネージド、サーバーレス、スケーラブルなランタイム メタデータ サービス

Real-time intelligence (Preview) BigQuery ワークフロー、BigQuery 用の Apache Kafka、サーバーレス Composer を含んだイノベーション

Data to AI governance (Preview) AI powered なガバナンス ルール、自動化されたデータ プロファイリング、データと Vertex AI にわたるデータ品質を実現

02

AI を自社データと “つなぐ”

BigQuery multimodal AI with Gemini 1.0 integration (Preview) をビジネスデータに

Embeddings, vector indexes & search (Preview) BigQuery で LLM、セマンティック検索、近傍検索全体のユースケースを強化

Continuous Queries (Preview) BigQuery の AI 機能を備えた自動化パイプラインにより、リアルタイムのユースケースを強化

Doc AI and Speech to Text (Preview) BigQuery の非構造化データの音声とドキュメント AI に LLM を使用

03

全てのデータに関わる人たちへ “Always-on intelligence”

Gemini in BigQuery (Preview) 生成 AI の支援、レコメンデーション、最適化、新製品エクスペリエンスにより、データ チームの生産性と価値実現までの時間を短縮

Gemini in Looker (Preview) 会話型 BI を実現し、ビジネス ユーザーとアナリストの洞察を得るまでの時間を短縮

01

データからAIまで
“One unified” プラットフォーム

One unified platform from data to AI with BigQuery

BigQuery Studio

一般提供

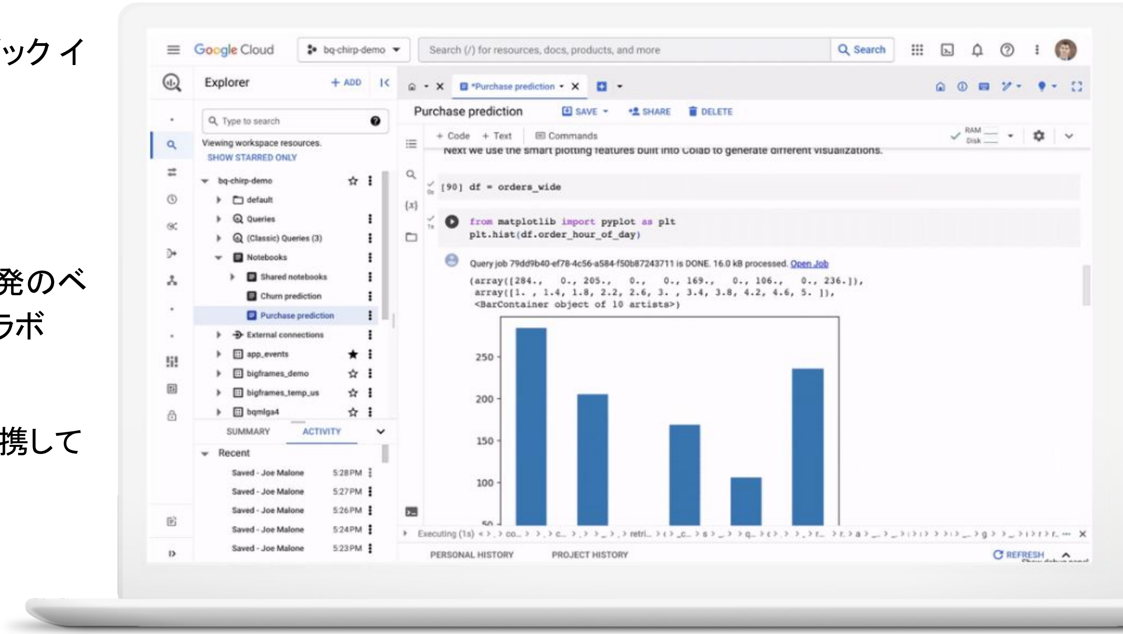
データに関わる全ての人々が、AI を活用したデータ分析を行うための
コラボレーティブなワークスペース

すべてのデータ ユーザー向けの SQL およびノートブック インターフェイスを含む単一の統合ワークスペース

データの規模、形式、場所に関係なく、分析に SQL、Python、Spark、JavaScript、自然言語を使用

一元的なソース管理とリビジョン履歴: ソフトウェア開発のベスト プラクティスをデータとコードに拡張することでコラボレーションを向上

Gemini 搭載のチャットおよびコード アシスタントと連携して生産性を最大化



Managed lakehouse with BigLake

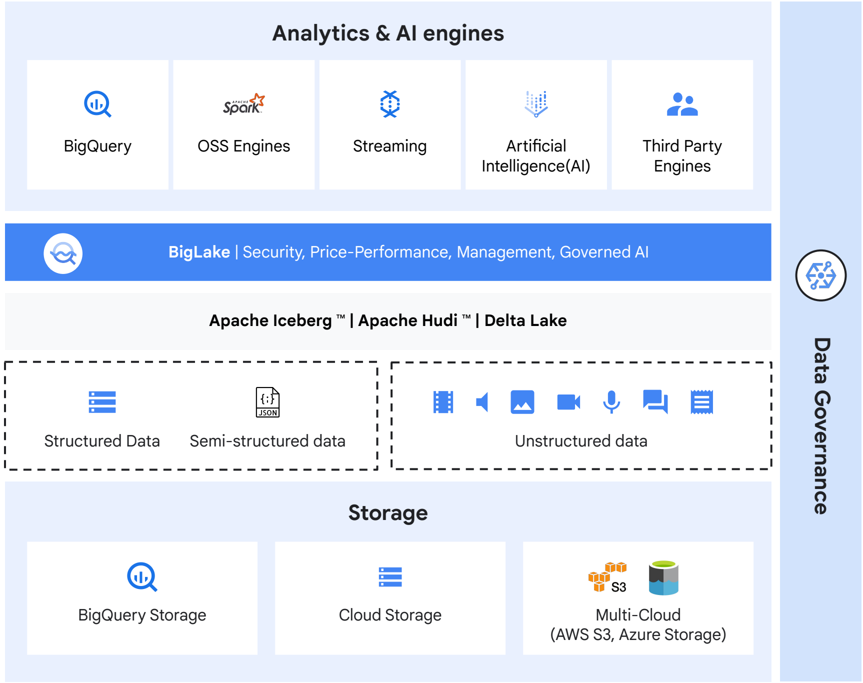
Google スケールをオープンフォーマット データに導入
高スループットのストリーミングと自動管理を実現

一般提供 ビルドインされたきめ細かいアクセス制御によりApache Iceberg、Delta、および Hudi 形式をクエリ

一般提供 大規模なインジェッションのためのApache Iceberg の**パフォーマンスアクセラレーション**

プレビュー **BigLake Managed Tables** : DML と高スループットストリーミングをサポートするオープンフォーマット(Iceberg) でのフルマネージドエクスペリエンス

プレビュー **Spanner external datasets:** Cloud Spanner スキーマを外部 BigQuery データセットとして使用



BigQuery metastore

プレビュー

SQL およびオープンソース エンジンのための統合されたランタイム メタデータ

単一の 共有型メタストア

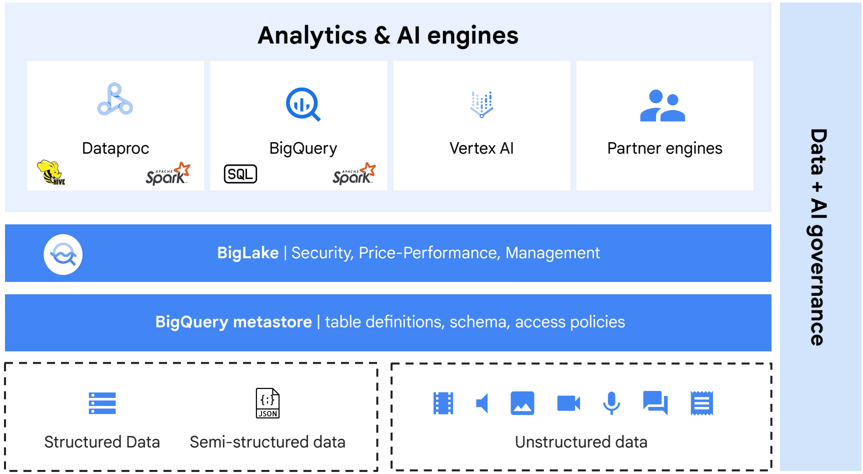
SQL、オープンソースエンジン、AI/ML 用のランタイム メタデータとコネクタ

構造化データと非構造化データをサポート

同じランタイム メタストアを使用して構造化テーブルとオブジェクトテーブルの両方をクエリする

ビルドインされたガバナンス

大規模なポリシー管理に利用可能なすべてのメタデータ



Governance at scale with trusted data

Data Profiling 一般提供

BigQuery Studio 内で詳細な統計プロファイリング情報を使用して、**データをより深く理解**し、データのドリフトを追跡

Automatic Data Quality 一般提供

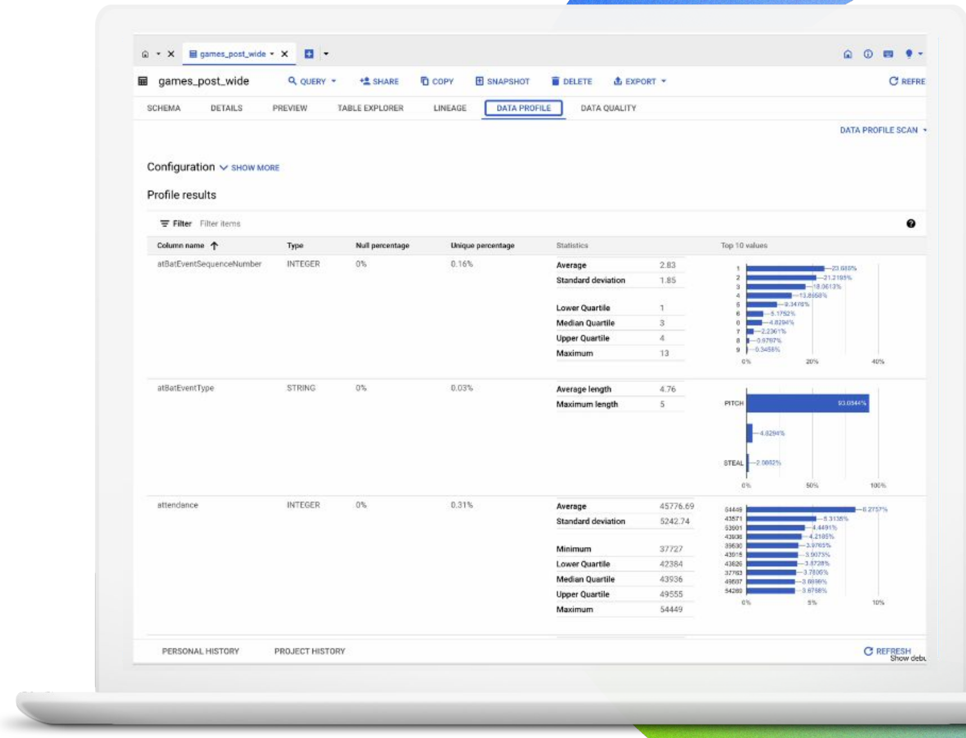
組み込みのルール推奨、サーバーレス実行、およびアラートにより、大規模なデータの信頼を構築します。BigQuery Studio でテーブルの品質指標を表示

Data Lineage 一般提供

データの移動を理解するための**End-to-End のリネージ**を自動追跡。データの問題に関する下流の影響、上流の原因、およびポリシーへのコンプライアンスを追跡

Governance Rules プレビュー

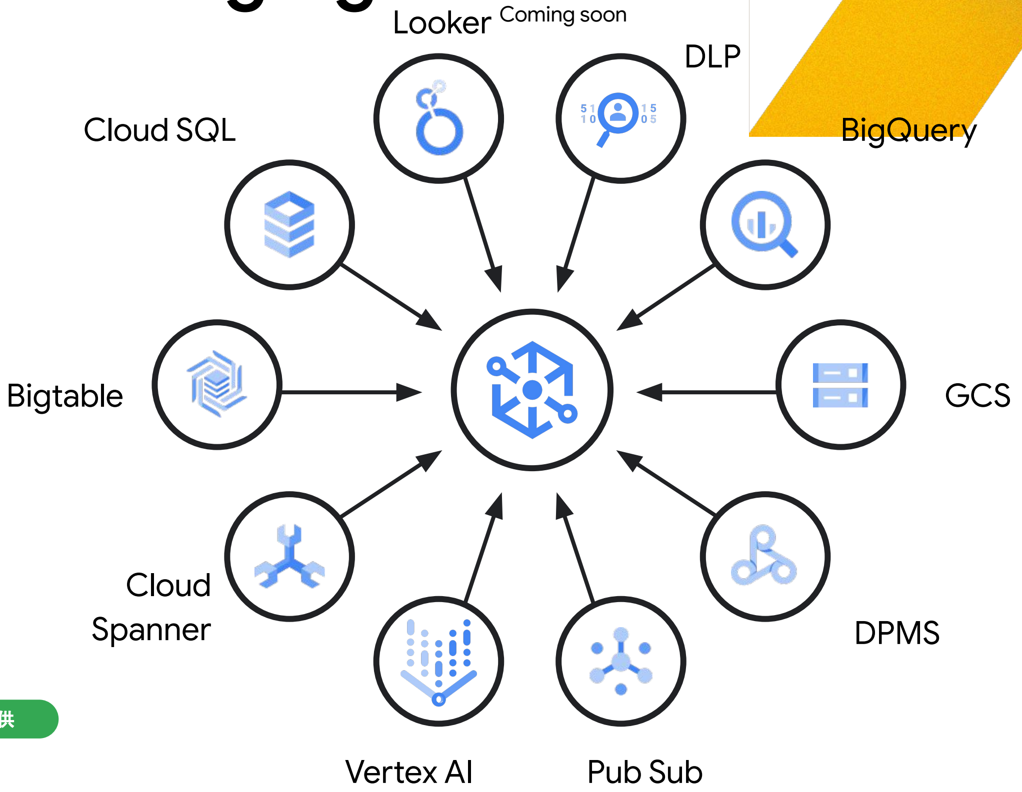
BigQuery データに大規模に適用される**ポリシー**を定義し、きめ細かいアクセス制御を通じて自動的に適用



Data to AI automatic cataloging

- **必要なデータを簡単に発見 -** アナリティクス、レイク、データベース、AI、BI サービスにわたる自動メタデータ取り込みとインデックス作成機能
- **分散データ資産を明確に管理 -** Dataplexの包括的なメタデータグラフ、データリレーションマッピング、重要なビジネスコンテキストの提供
- **セルフサービスのデータ アクセスを強化 -** Dataplex の支援により、データ ガバナンス実践のための堅牢な基盤を構築

Cataloging of metadata from **Bigtable** 一般提供
Cataloging of metadata from **Spanner** 一般提供
Cataloging of metadata from **Cloud SQL** プレビュー
Cataloging of metadata from **Looker** Coming Soon
Cataloging of models and datasets from **Vertex AI** 一般提供
Cataloging of features from **Vertex AI** プレビュー

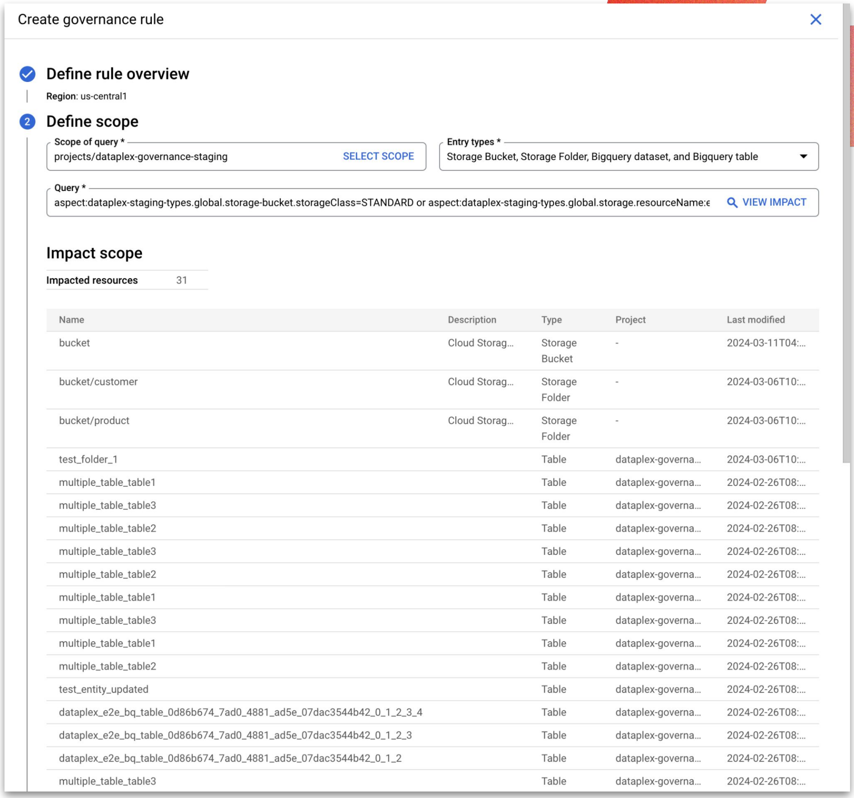


プレビュー

Governance rules

ガバナンス ルールは、データ アクセスの管理方法を合理化するのに役立ちます。これにより、メタデータ経由で取得されたビジネス セマンティクスを使用して、適切なユーザーが適切なデータに大規模にアクセスできるようになります

- コンプライアンス指示を強制可能なセキュリティ ポリシーに変換するために費やす時間と労力を削減
- 誰が何に、なぜアクセスできるかを定義するアクセス スペックを使用して、Dataplex 検索クエリとしてインテントを定義
- クロスシステム アクセスルールをサポート
  - **BigQuery** Dataset、Table、Column
  - **Cloud Storage** Bucket、Folder (via **managed folders**).



Data sharing

データおよび AI の資産を組織の壁を越えてセキュア・効率的に交換

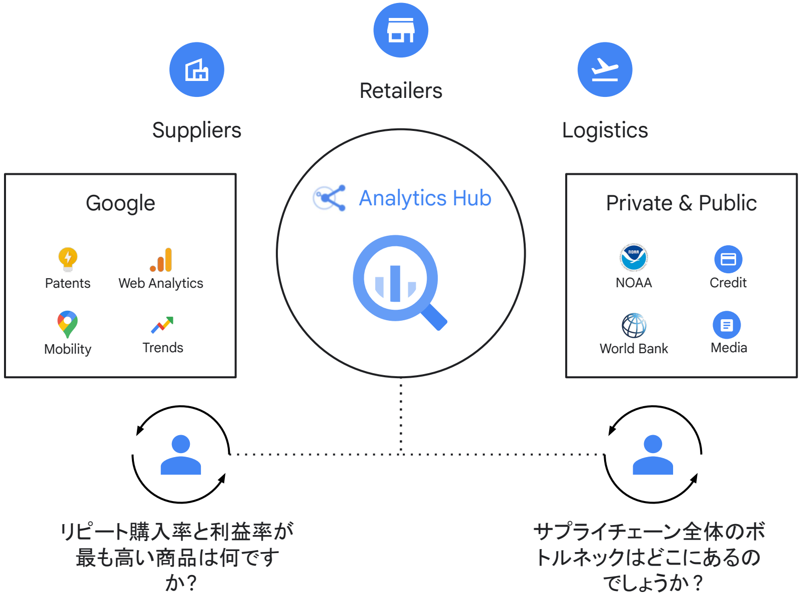
一般提供
Subscriber による共有データの使用状況を測定および監視する**使用状況メトリクス**

パブリック プレビュー
サブスクリプション管理により、データプロバイダーは自社の Listing へのサブスクリプションを追跡および管理

パブリック プレビュー
Data Egress Control により、サブスクライバーが Analytics Hub を通じて共有されたデータをコピーまたはエクスポートできないように制限

パブリック プレビュー
プロバイダー定義関数のサポートにより、データに関連付けられた複雑なビジネスロジックまたは変換

一般提供
一部リージョン
BigQuery Omni を使用した**クロスクラウド**（AWS/Azure）データ共有

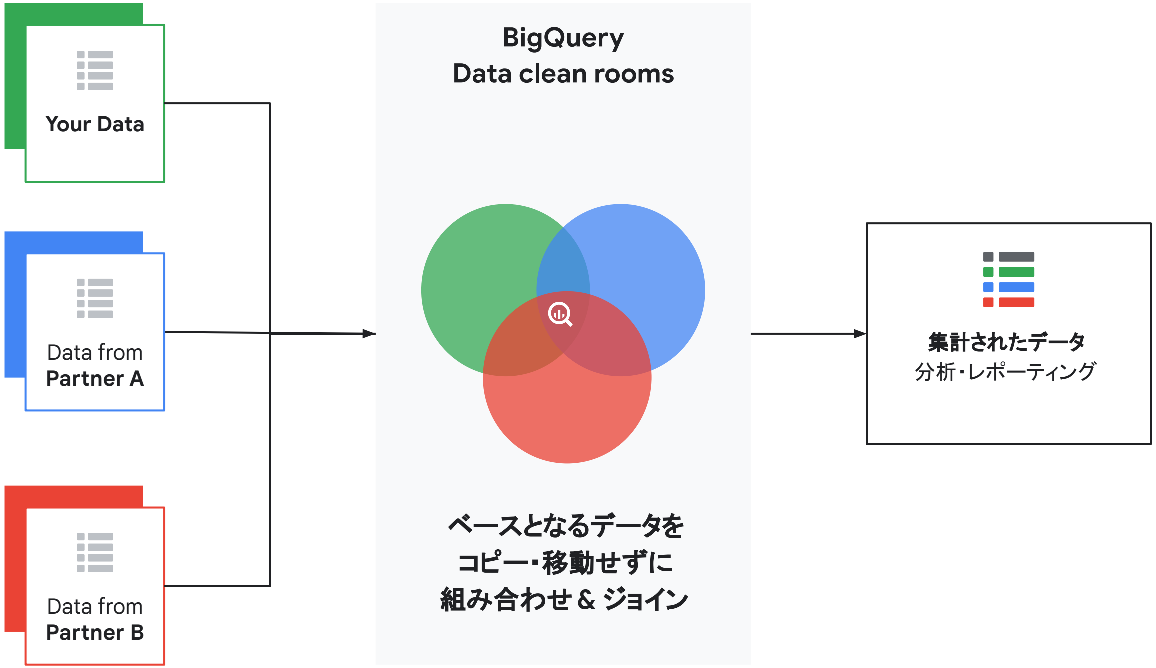


BigQuery data clean rooms

一般提供

センシティブなデータを共有、組み合わせ、分析する

- ベースとなるデータをコピーしたり移動したりせずに、お客様とパートナーがデータ資産で共同作業できるように、**セキュアでロートラスト**な環境を作成
- 最先端のデータサイエンス技術と**プライバシー ポリシー**を活用してデータを保護
- 使用状況を監視して共有データの**プライバシーの脅威**を検出



BigQuery workflows

プレビュー

BigQuery Studio でオーサリングされた ビジュアル ワークフロー

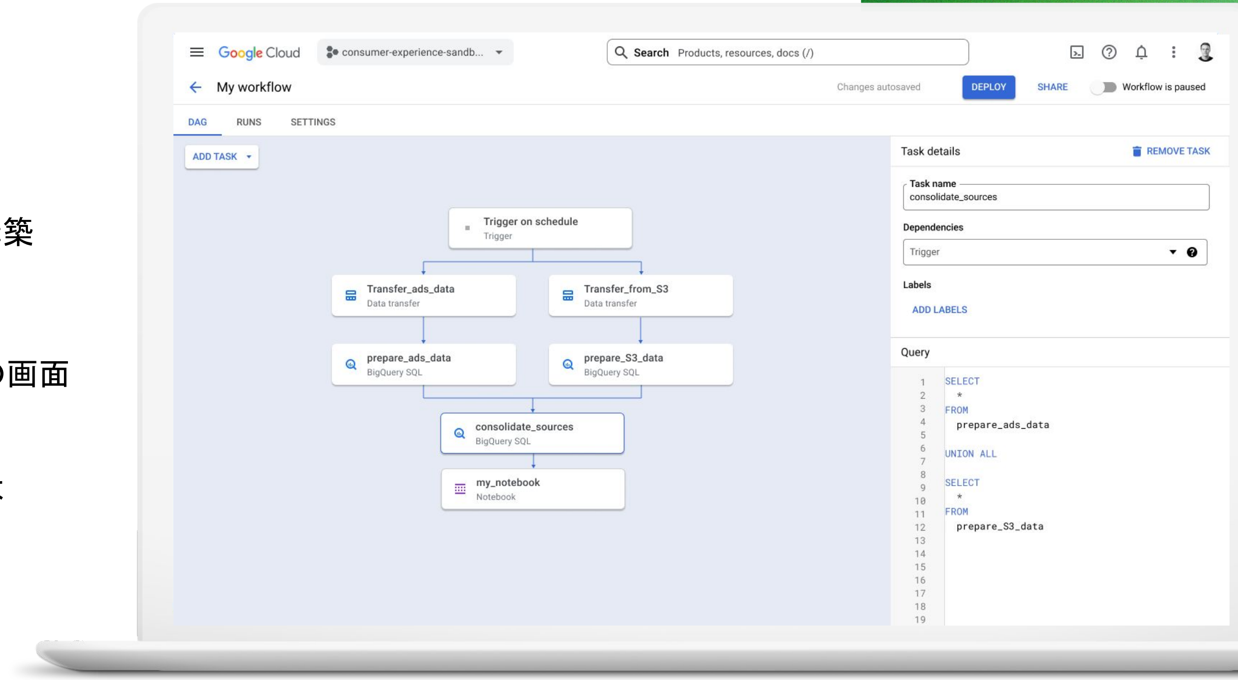
ワークフローでの SQL Scripts、Python Notebooks、Data Canvases、および Data Preparation を**スケジュール**

コードを書かずに**ワークフローを視覚的**に構築

BigQuery Studio に埋め込み

すべての BigQuery スケジュールを 単一の画面から監視

ビジュアル ワークフローを Dataform または Composer DAG に**エクスポート**



Continuous real-time analytics in SQL

プレビュー

継続的な SQL 処理により “インテリジェントな” データパイプラインを解き放つ

継続的な分析処理

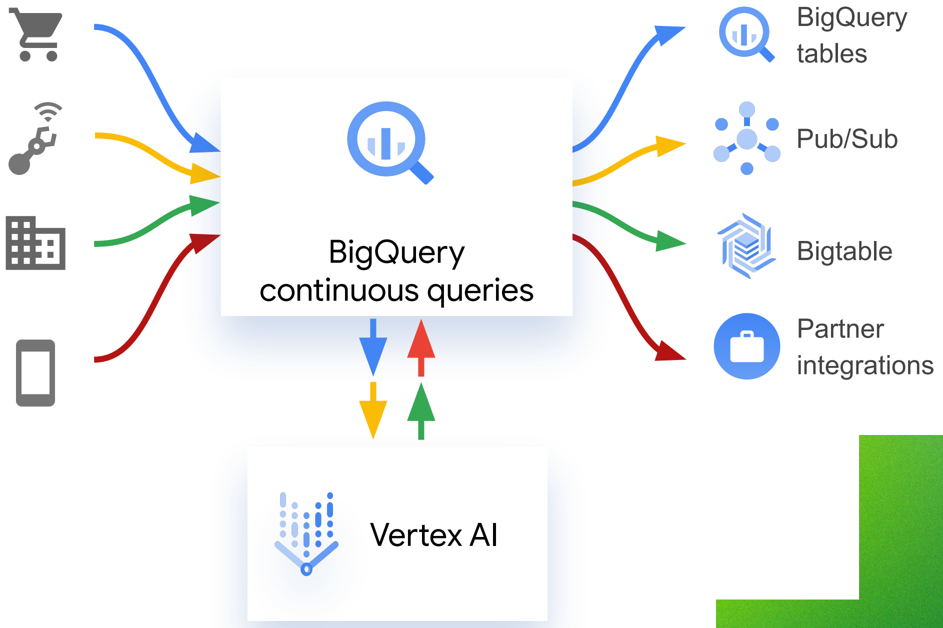
SQL を使用して、受信データのストリームに対して無制限のサーバーレス分析を実行

リアルタイムアプリ開発 + GenAI

リアルタイムの異常検出、センチメント分析、レコメンデーションなどの AI+ML 機能を備えたパイプラインを開発

Reverse ETL を オペレーションシステムに

BigQuery から Pub/Sub、Bigtable、または別の BigQuery テーブルにリアルタイムでデータを変換および複製



Query acceleration improvements

ワークロードを変更することなくコストパフォーマンスを最適化

プレビュー

自動化された自己学習型 **履歴ベースの最適化** により、クエリ パフォーマンスが最大 5 倍高速化

プレビュー

小規模なクエリ (スキャン対象 1 GB 未満) の実行が最大 50% 高速化。マルチユーザーワークロード向けの高スループット (‘000 秒 qps)

一般提供

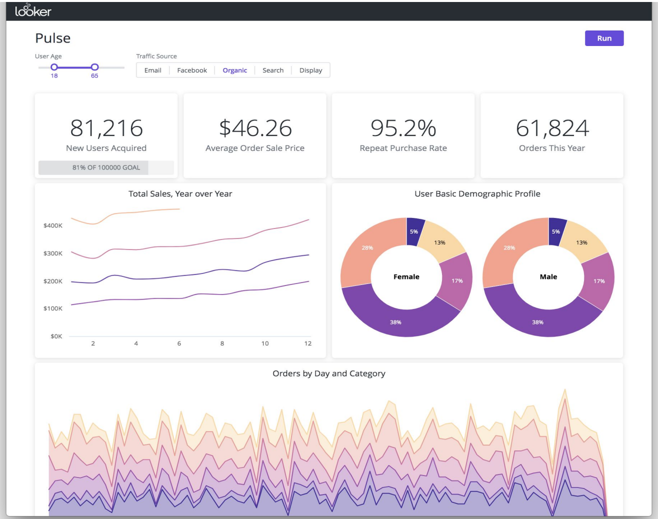
自動スマート キャッシュを備えた適応型ベクトル化 BI エンジン。キャッシュによって満たされるクエリは 1 秒未満で実行

一般提供

あらゆるツールからの透過的なクエリ アクセラレーション：BigQuery Studio、Looker、Looker Studio、Tableau、Power BI など

プレビュー

Looker Studio のエクスペリエンスの向上 (ジョブの詳細、モニタリング、コスト追跡、クエリ キャッシュ、情報スキーマ、マテリアライズド ビュー)



02

AI を自社データと “つなぐ”

Connect AI to all your data

BigQuery ML: Bringing AI to your data

BigQuery と Vertex AI の一体化により、エンドツーエンドの機械学習サポート

一般提供

データとAIの統合

SQL経由のモデルトレーニングと推論

Vertex AI統合 - モデルレジストリ、機能ストア、リモートモデル

一般提供

進化したモデルケイパビリティ

多変量、階層的時系列予測、休日モデリング

説明可能なAI

高度な機能エンジニアリングとサービス提供

プレビュー

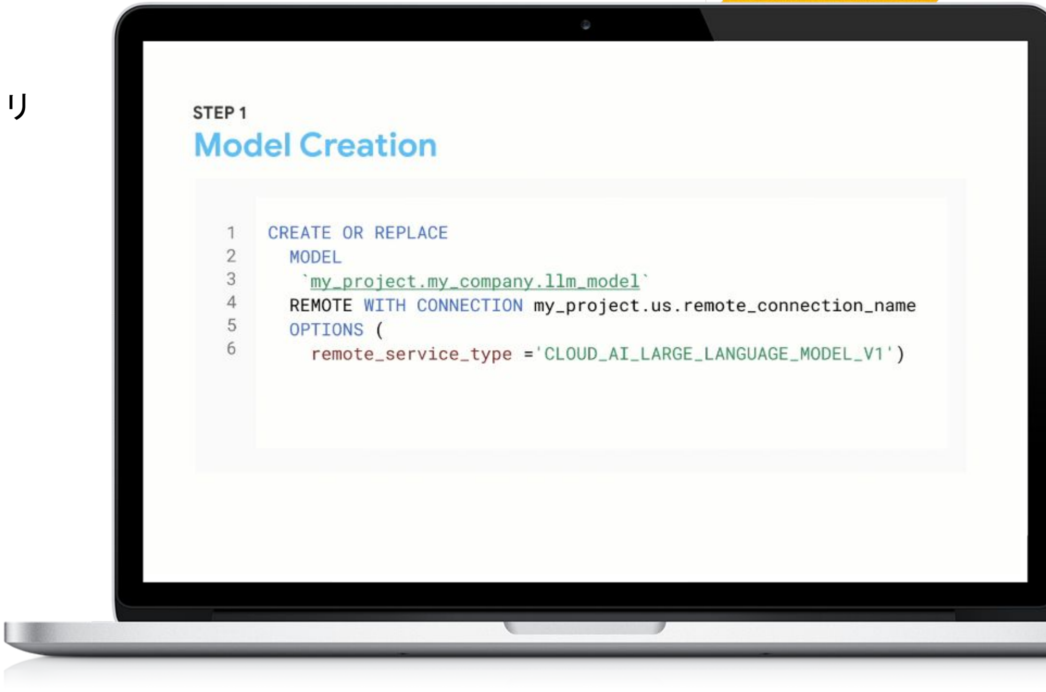
LLMs / Cloud AI APIs へのリモート推論

generate_embedding

generate_text (text, image, video input),

process_document, transcribe および

LLM ファインチューニング



BigQuery vector embeddings, search, and support for LLM frameworks

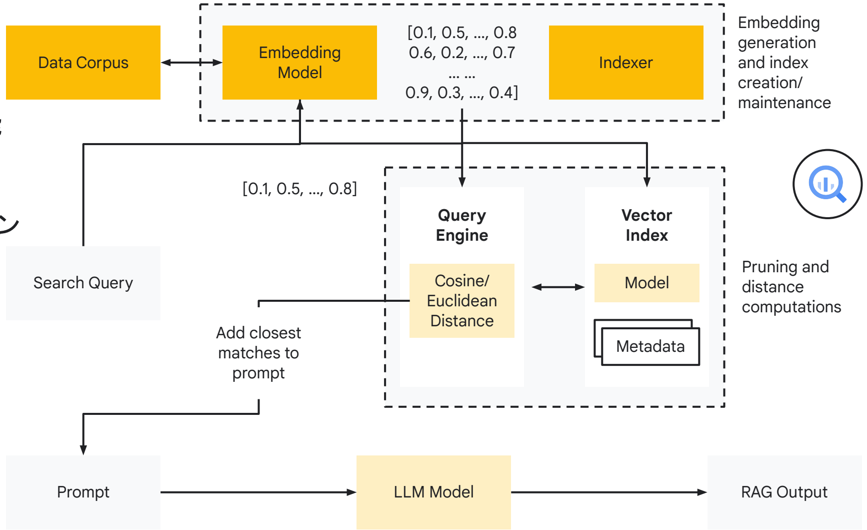
プレビュー

RAG と Long term memory で LLM を 一歩先へ

- BigQuery のスケールを使用して、複数のオペレーショナルデータベース、ベクトル データベース、Cloud Storage にわたるベクトルを収集し、標準化
- データ ウェアハウス上で直接RAG 設計パターンを開発し、LLM の機能とビジネス データを組み合わせて精度を向上
- フルマネージドインデックスにより、エンベディングが自動生成され、オンライン予測のためにVertex と同期
- Langchain などの一般的なLLM フレームワークのサポートにより、BigQuery データが LLM と通信できるように

ベクトルインデックスが必要となるケースは?

- RAG デザインパターン
- LLMs のための Long term memory
- セマンティック検索 - 意味ベース検索 vs キーワード検索
- 近傍検索 - 他のテキストと意味の近いテキストを特定
- レコメンデーション



RAG Workflow in BigQuery

03
全てのデータに関わる人たちへ
“Always-on intelligence”

Gemini in BigQuery

Always-on intelligence

データ & アナリティクス ジャーニー全体に融合

探索, ディスカバリ および 分析
Explore, Discover & Analyze

エンジニアとトランスフォーム
Engineer & Transform

ワークロードの最適化
Optimize Workloads

データ セキュリティ
Secure Data

ファンダメンタル
Fundamentals

Gemini in BigQuery

プレビュー

すべてのデータ プラクティショナーのインサイトを加速させる　生成 AI との "統合"

コンテキスト アシスタント - GoogleSQL と Python 向けに最適化

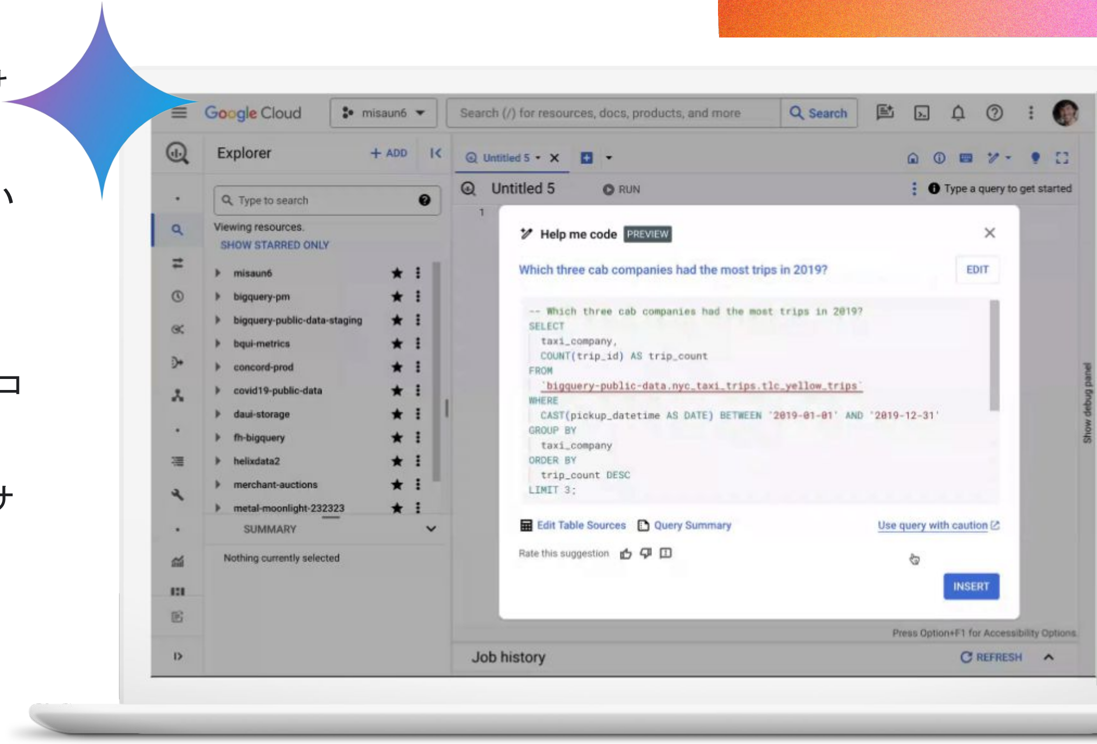
自然言語から SQL 生成 - 自社のデータセットに基づいて新しいクエリをブートストラップ

自然言語から Python 生成 - Colab notebooks

インライン SQL 補完 - クエリコンテキストに基づいたレコメンデーション

統合されたチャット - 複雑なクエリを理解し、分析者をサポート

セキュアモデルの統合 - データとプロンプトは Google Cloud 上で実行

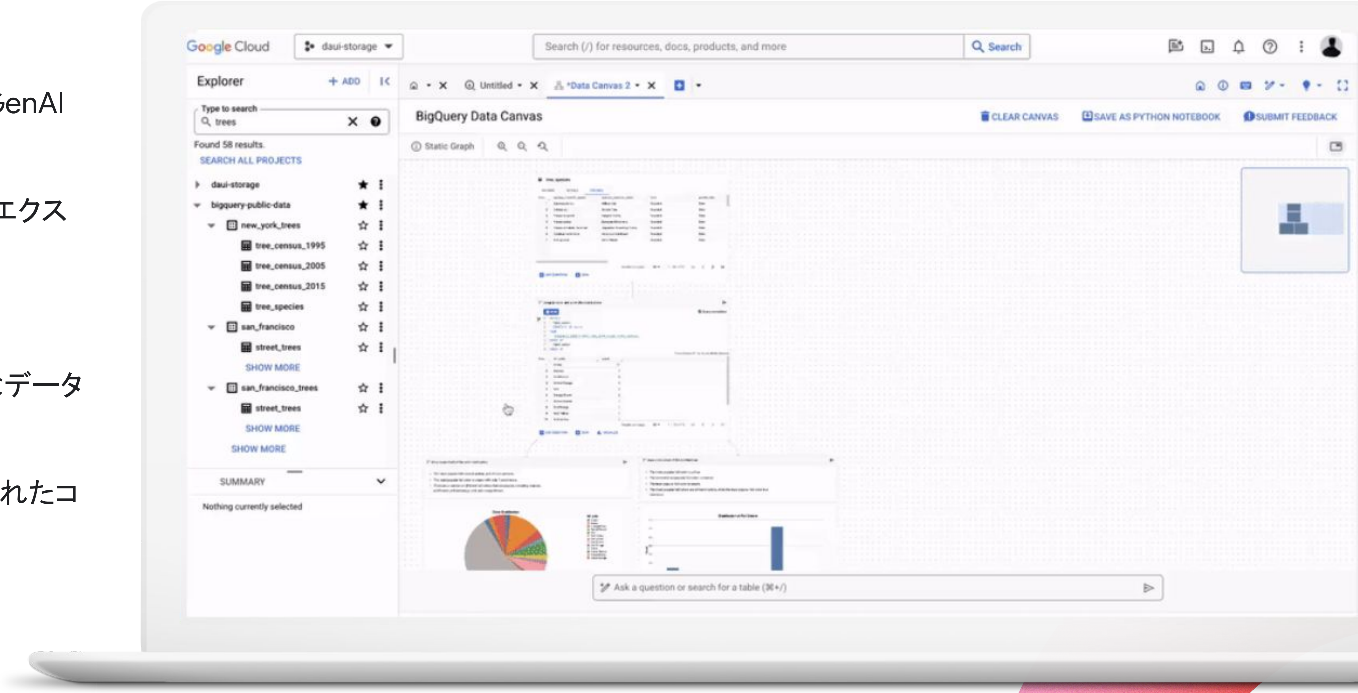


BigQuery data canvas

プレビュー

ユーザの思考をインサイトに変える "生成 AI 中心" の新しいエクスペリエンス

- データの探索と可視化を実現する GenAI 中心のエクスペリエンス
- ユーザを支援するインタラクティブなエクスペリエンス
- BigQuery Studio との統合
- Dataplex catalog のセマンティックなデータディスカバリ
- データアナリスト向けの ビルドインされたコラボレーション機能



BigQuery data preparation

プレビュー

生成 AI アシストで進化した直感的な画面操作可能なデータ プリパレーション

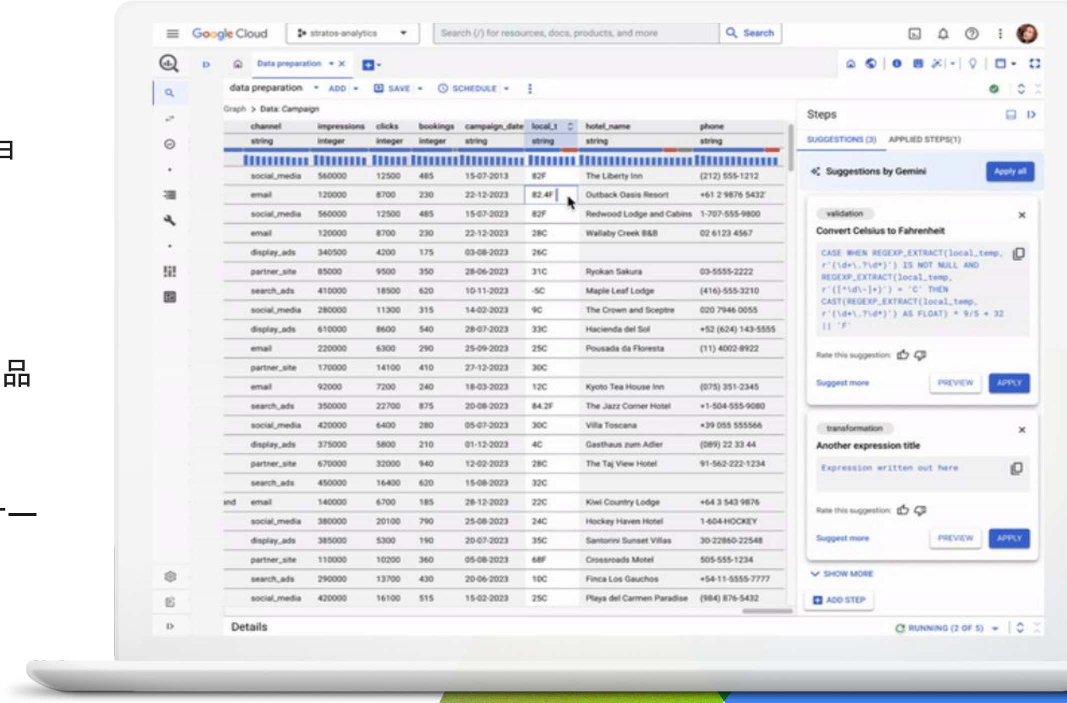
BigQuery でのデータのクレンジング、変換、準備

生成 AI によりコンテキストを認識した変換レコメンデーションにより、データ開発を加速

合理化された取り込みと変換で時間を節約

自動化されたスキーマ マッチング、ビジネスおよびデータ品質ルールの識別と検証

生成された SQL パイプラインを BigQuery にデプロイ、オーケストレーション、モニタリング



Gemini in Looker

ビジネス データと会話し
レポートと分析情報を作成および自動化

プレビュー

会話型アナリティクス
Conversational Analytics

レポート自動生成
Report Generation

自動スライド生成
Automatic Slide Generation

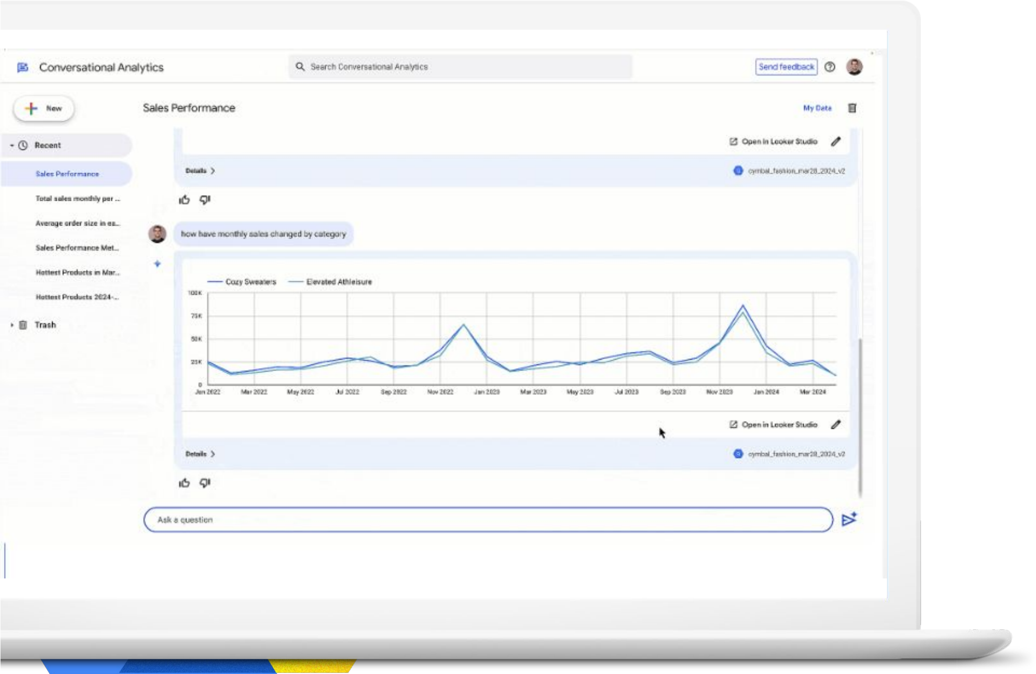
フォーミュラ アシスタント
Formula Assistant

可視化アシスタント
Visualization Assistant

LookML アシスタント
LookML Assistant

AIドリブンな会話型ビジネスインテリジェンス

Looker Conversational Analytics



プレビュー

- 自然言語を使用してチャートやビジュアライゼーションを即座に作成
- マルチターンのサポートと完全なクエリの透過性
- Studio ダッシュボードにエクスポートして、ドラッグアンドドロップで簡単に鮮明なレポートを生成
- BigQuery、Google スプレッドシート、CSV ファイルに接続してアドホックデータ探索
- Google ドキュメントのような共有とコラボレーション
- 管理された Looker モデルへの簡単なアクセスを拡張 *(Coming Soon)*

Data Management

Google の AI レディな Data Cloud

あらゆるユーザーに

- データチーム
- ビジネス部門の ユーザー
- 意思決定者

あらゆるユースケースに

- ソリューションの 加速
- ビジネス 部門
- パートナー 連携

Databases	Data analytics	AI & ML	Business insights
AlloyDB	BigQuery	Vertex AI	Looker

Vertex AI 基盤モデル (text, chat, code, video, image, audio), オープソース, サードパーティー AI モデル

Google Cloud GPUs/TPUs | Run across clouds / on prem | Govern data & AI | Most open ecosystem

AI でビジネスを変革

AI に対応したデータ基盤を活用

あらゆる業務データ・分析データを包含する、信頼性と安全性の高い統合データ基盤

AI 機能をデータやアプリケーションへ持ち込む

組み込みの機械学習、LLM、エンベディング、ベクトル検索があらゆるデータストアで利用可能となり、アプリケーションや分析を強化

Gemini の活用 AI による支援・効率化のために

Google Cloud の Geminiを活用して生産性向上、リスク低減、ワークロードの最適化を実現

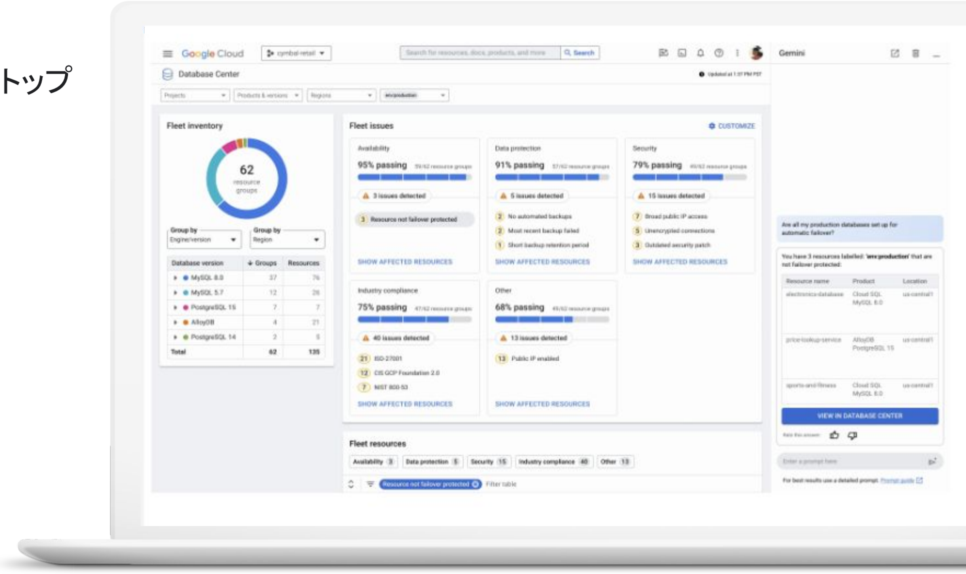
プレビュー

Gemini in Databases

AI を活用してデータベース開発・管理を劇的に効率化

- **Database Studio** で、インテリジェントなコード支援、コード補完、ガイダンス機能を使って、SQL コードを簡単に生成、要約、修正
- **Database Center** で運用中のデータベース群をワンストップで最適化
- 潜在的な問題をスマートな推奨機能で先回り
- AI を活用したコード変換でデータベース移行を加速

> “Gemini in Databases を導入することで、データベース全体の健全性に関する回答を数秒で得ることができ、アプリケーションに対する潜在的なリスクをこれまで以上に迅速に予防的に軽減することができます。”
>
> Bogdan Capatina, Technical Expert in Database Technologies, Ford Motor Company



オペレーショナル データが、基盤モデルと企業向け 生成 AI アプリケーションのギャップを埋める。

LangChain があらゆる Google Databases で利用可能に

高速な開発 | 相互運用性 |
エンタープライズグレードな機能

サポートされる Google Databases

- Cloud SQL for MySQL
- Cloud SQL for PostgreSQL
- Cloud SQL for SQL Server*
- AlloyDB
- Spanner
- Bigtable*
- Memorystore for Redis
- Firestore

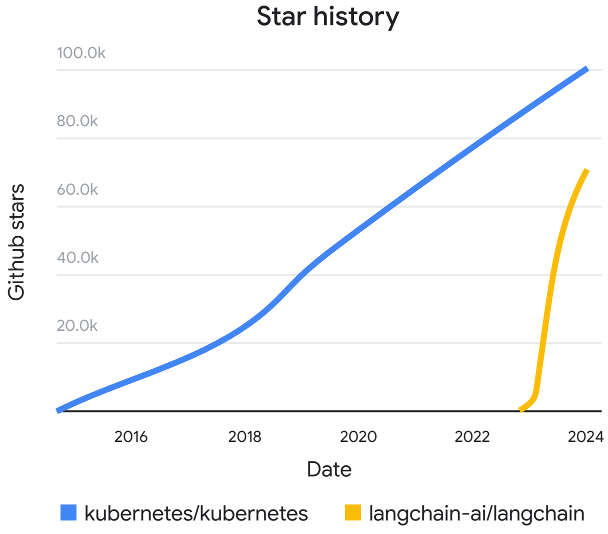
LangChain:
最も人気のある OSS 生成 AI フレームワーク

GitHub repo

+700 different integrations

+2k contributors

~70k stars



* no vector store

star-history.com

Vector support everywhere

生成 AI アプリケーションを任意の Google Cloud データベースで開発

In-memory	Relational			Document	Key value
Memorystore for Redis[1, 2]	Cloud SQL[1, 2]	AlloyDB[1, 2]	Spanner[2]	Firestore[2] (NEW)	Bigtable[2]
一般提供	PostgreSQL 一般提供 MySQL Preview	一般提供	プレビュー	プレビュー	プレビュー coming soon

1: ANN—大規模データセットに最適な近似最近傍探索
2: KNN—小規模なデータセットでの精度の高い結果のための厳密な最近傍探索

AlloyDB AI - 生成 AI アプリケーション構築をより強力にサポート

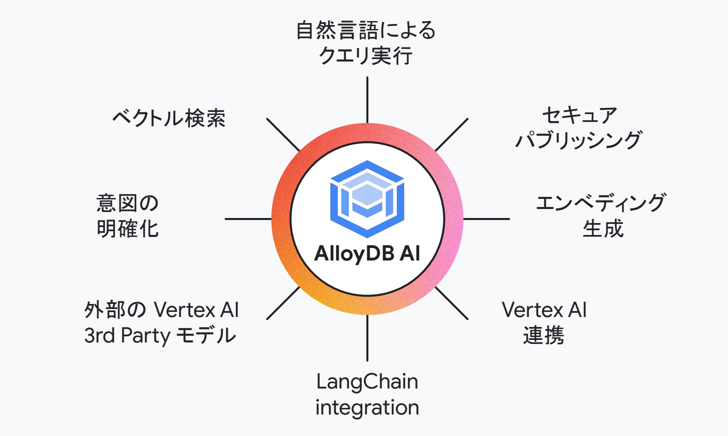
プレビュー

- **postgres_ann による高速なベクトル検索** - pgvector と互換性のある拡張機能で標準 PostgreSQLの `hnsw` インデックスよりも最大 4 倍高速、かつメモリ使用量は通常 1/3-1/4に改良

- 自然言語サポート - 自然言語(英語)を用いたクエリ実行が可能になり、柔軟性と表現力の最大化

- Secure publishing - パラメータ化されたセキュアビューにより、生成 AI アプリケーションからの柔軟でセキュアかつ正確なデータベースアクセスを簡単に設定可能に

- モデルエンドポイント管理 - Vertex AI 上、あるいは他のプラットフォーム上にあるリモートモデルの登録と呼び出しを簡単に。どこに配置されたモデルでも管理が容易に



> "Nuro は自動運転車企業で、ベクトル類似性検索を活用して、走行中に自動運転車が遭遇する物体(歩行者、車、信号など) を分類し、適切な行動を取らせるシステムを開発しています。現在Nuro は数億個のベクトルデータをAlloyDB AIに移行しており、アプリケーションアーキテクチャの簡素化を目指しています。"
>
> Fei Meng, Head of Data Platform, Nuro

AlloyDB AI によるベクトル検索のパフォーマンス

AlloyDB AI の ScaNN インデックス性能 - 標準の PostgreSQL の HNSW インデックスと比べて

最大
4倍
高速なベクトルクエリ

最大
8倍
高速なインデックス作成

最大
10倍
高い書き込みスループット

通常
3-4倍
少ないメモリ使用量

Source: Google Cloud performance tests, March 2024

AlloyDB AI Innovations @ Next'24

New | Available

自然言語

- 意図の明確化
- 自然言語によるSQL 実行
- パラメータ化された
セキュアビュー

モデルアクセス

- モデルエンドポイント
マネジメント
- エンベディング生成
- Vertex AI 連携

ベクトル検索

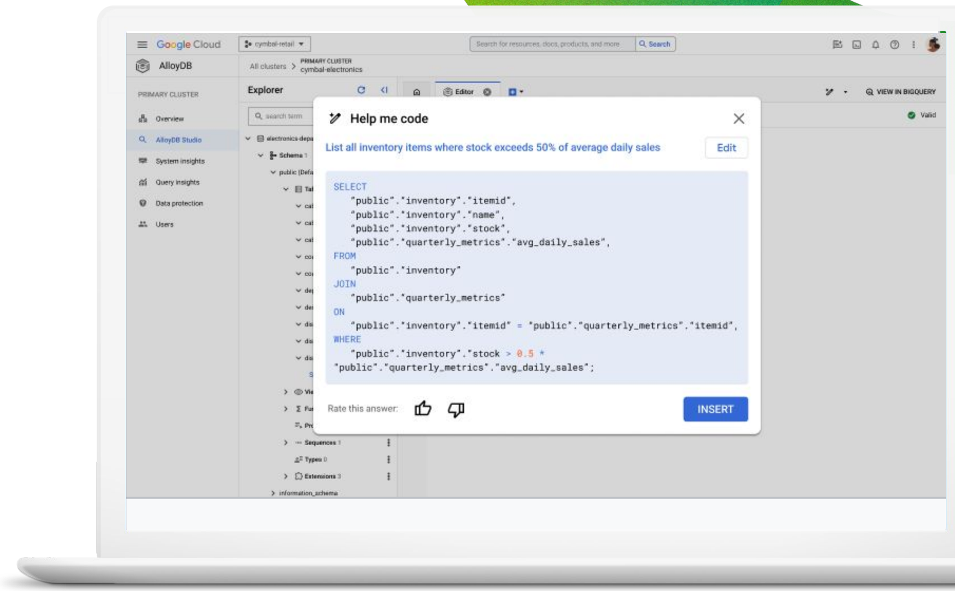
- Google 独自の ANN - SCaNN
（近似最近傍探索）
- pgvector サポート

AlloyDB & AlloyDB Omni

LLM Orchestration with frameworks: LangChain, Vertex AI

AlloyDB @ Next

- AlloyDB AIで、より高速なベクトル検索のためのpgvector互換の"ann"拡張、アプリケーション開発における自然言語サポート、正確なデータベースアクセスのためのセキュアパブリッシング、および自社モデルとサードパーティ製モデルの呼び出しをサポートするモデルエンドポイント管理機能 **プレビュー**
- Gemini in AlloyDB により、AlloyDB Studioを利用し、高度なコードアシスト、コード補完、ガイダンスを利用可能に。Google Cloud console上から、SQLではなく自然言語(英語など)を使って、データの検索・更新といった操作が可能 **プレビュー**
- AlloyDB Omniは、AlloyDB Omni Kubernetes OperatorのGA版リリースにより、データベースのプロビジョニング、バックアップと復元、高可用性といった機能強化 **一般提供**



“株式会社AbemaTVでは、AlloyDB AI のエンベディング生成およびベクトル検索を活用し、「ABEMA」の動画レコメンデーションを行っています。AlloyDB のモデルサポートやベクトル機能の迅速な拡張に期待しており、新たに提供されるモデルカタログを活用してGemini Pro のような最新エンベディングモデルを手軽に活用したいと考えています。また、新たに発表されたscann index によるベクトル検索の高速化を試すことを楽しみにしています。”

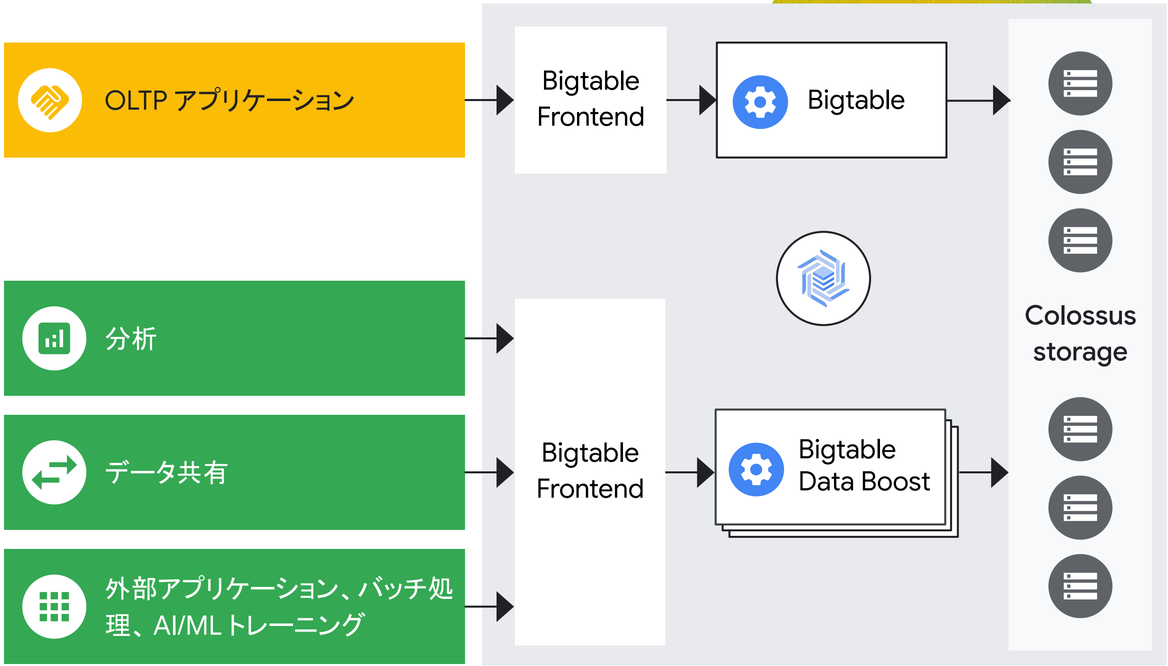
菅 俊弥, Engineering Manager, 株式会社AbemaTV.

ABEMA TV for the Future

Bigtable Data Boost

プレビュー

- 分析、バッチクエリの実行、トランザクションワークロードから完全に分離された状態でMLモデルをトレーニング
- キャパシティ管理不要でオーバーヘッドなし、従量課金制
- 自動的な最適化および常時稼働するコンピューティングで長時間実行されるクエリの高速化
- データの集約を可能にし、ガバナンスとデータ共有を簡素化



Additional Bigtable launches

承認済みビュー
一般提供

行/列レベルのアクセス制御により、マルチテナントテーブルをサポート

ビューを通じた読み取り、書き込み、削除に対応

Data Boost の活用により、連携先やパートナーとデータを共有。共同分析やML機能の開発を簡素化

分散カウンタ
プレビュー

リアルタイムな指標を大規模に扱うための、インクリメントが可能な集計型

シングルノード運用から、大規模分散システムまで、スループットやパフォーマンスを妥協する必要なし

広告クリック、いいね、シェア、センサー情報、API リクエストなどの大量のストリーミングデータを機械学習や運用レポート作成のために処理可能

リクエストの優先度
一般提供

異なる優先度でワークロードを実行することで、バッチ処理がレイテンシーに厳しいトラフィックに与える影響を最小限に

Reverse ETL
プレビュー

BigQuery から継続してクエリを実行して、変更データのストリーミングを取り込み可能

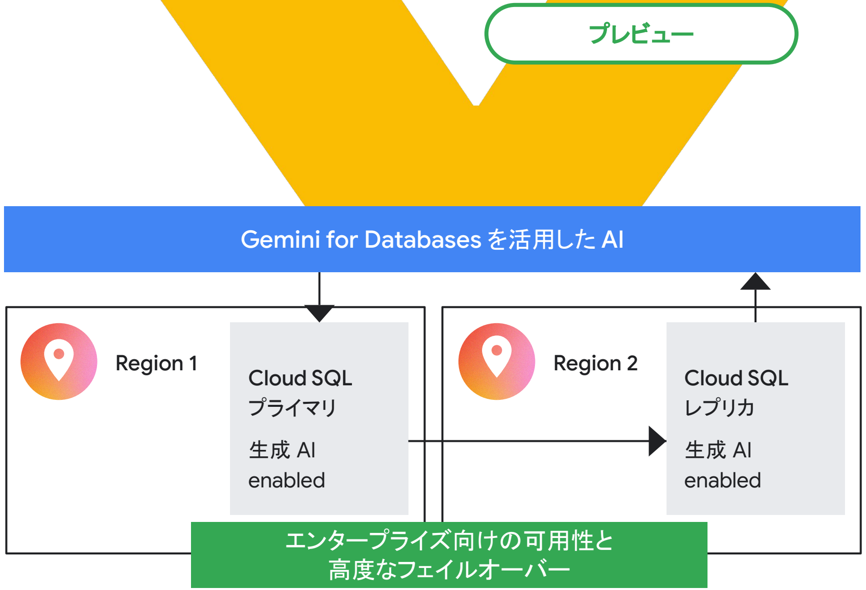
> “私たちにとって、Bigtable は単なる巨大なデータ保管庫ではありません。バッチクエリとストリーミングクエリを大規模に実行できる高性能な分析エンジンです。これは、データ処理能力のあり方を大きく変えるものです。”
>
> Mike Wiacek, Chief Executive Officer & Founder, Stairwell

Cloud SQL の最新情報

ベクトル機能（Cloud SQL for PostgreSQL および MySQL）、LangChain および Vertex AI との統合など、更新された機能を使用して、**生成 AI アプリを簡単に構築**できます。

Gemini in databases は、データベース利用における最高の助手であり、Studio でのコード開発、Database Center でのフリート管理、移行の専門知識など、データベースジャーニーのあらゆる側面を簡素化する、統合されたインテリジェントなエクスペリエンスを提供します。

Enterprise Plus の新たな改善により、エンタープライズの事業継続計画の重要な部分である、オーケストレーションされたスイッチオーバーとスイッチバックなどの高度な機能が実現します。



> “Linear では、既存のデータ量とコストパフォーマンスを両立させるためpgvector のサポート開始をきっかけにCloud SQL for PostgreSQL の採用を決定しました。決め手となったのは、その高い拡張性と信頼性です。また、既存システムで利用していたデータベース、モデル、ORM などと互換性があるため、チームはすぐに対応・運用できました。”
>
> Tom Moor, Head of US Engineering, Linear

Cloud Spanner 最新リリース情報

プレビュー

KNN ベクトル検索

最近傍（KNN）検索のために、コサイン距離関数とユークリッド距離関数を使用した柔軟なベクトル検索を実行

プレビュー

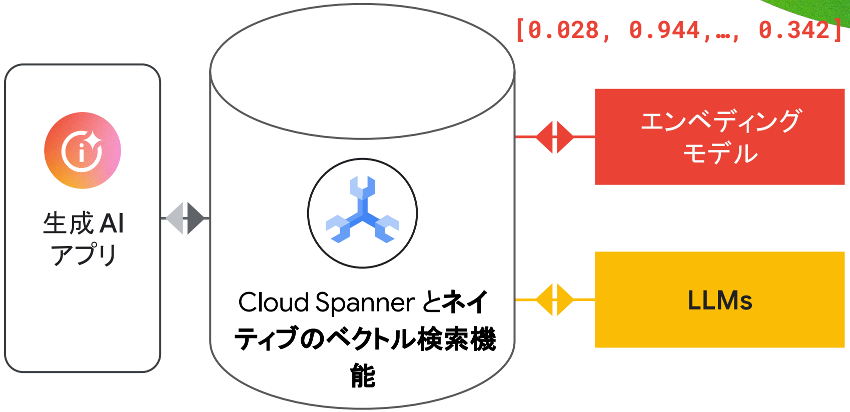
BigQuery でのゼロ ETL クエリ

Spanner のトランザクションデータと BigQuery の分析データを、2 つのデータソース間でETL によるレイテンシなしで直接シームレスにクエリ

一般提供

スループットを最適化した書き込み

スループットとレイテンシのトレードオフを調整可能なパラメータを介して、スループットパフォーマンスを2〜5 倍に最大化



> “Cloud Spanner 上に構築したものは、機敏性、スピード、安全性、そして経済的な健全性のために設計された、モダンで柔軟な銀行プラットフォームです。これは、我々のお客様のニーズを満たし、現在そして将来にわたってビジネスの成功を確実にするものです。”
>
> Connor Squires, Senior Software Engineer, ANZ Bank

Memorystore @Next
ベクトル検索、永続化機能とクラスタサイズの柔軟性

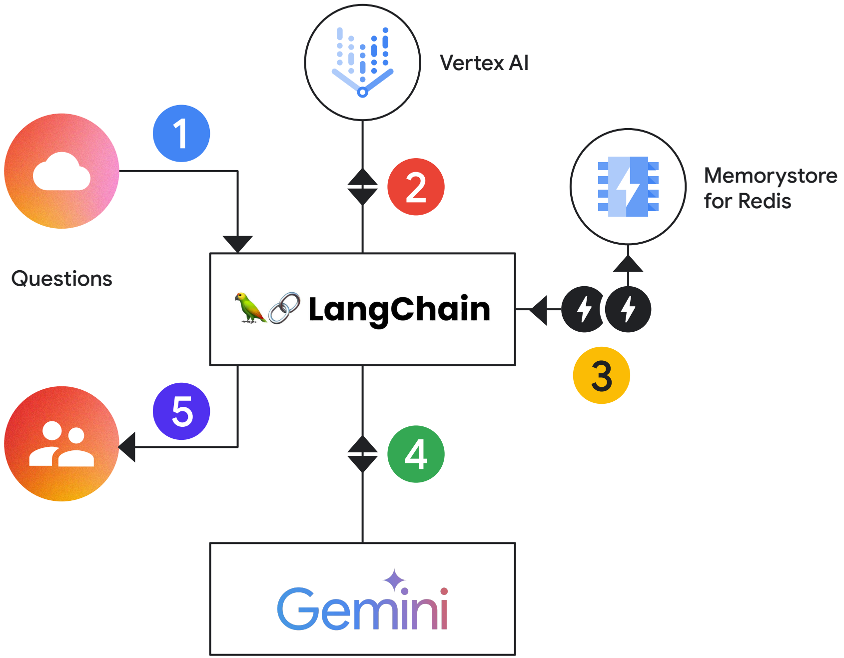
Ultra-Fast Vector Search 一般提供
Memorystore for Redis インスタンスを使用すると、数千万個のベクトルに対して 1 桁ミリ秒のレイテンシでベクトル検索を実行可能に

Memorystore for Redis Cluster AOF (Append Only File), RDB (Redis Database) プレビュー
データ永続化により、インメモリデータを保護

Flexible Cluster Sizing 一般提供
3 つの新しいノードタイプを使用した、柔軟な新しいクラスタサイズ設定。単一クラスタインスタンスは 4.2 GB から開始でき、最大で 14.5 TB のメモリにスケールアウト可能

> “Memorystore for Redis Cluster は、ワークロードに必要なスケーラビリティと管理性を提供してくれるため、より意義のある方法で Memorystore を活用できます。”
>
> Stanley Feng, Senior Engineering Manager at Reddit

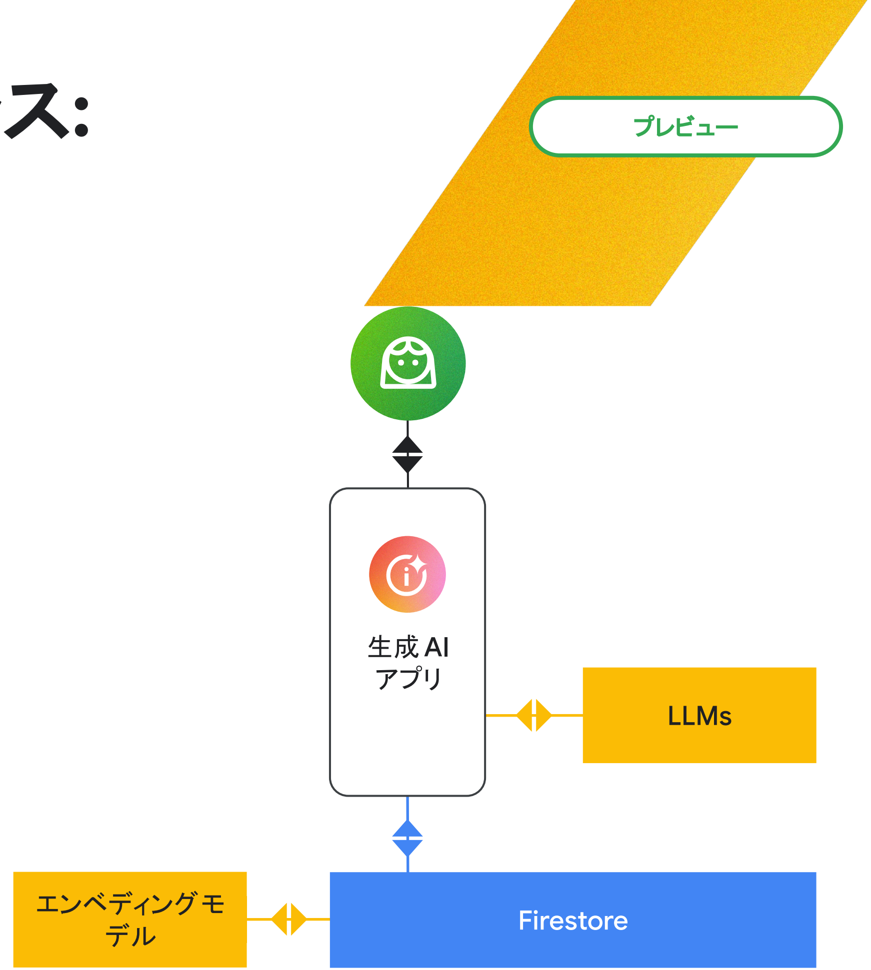
reddit



Firestore の Next アナウンス: 生成 AI、クエリ、CMEK

- Firestore 上で 生成 AI ソリューションを構築 - ベクトル検索、自動化 & 拡張可能なエンベディング生成、LangChain、LlamaIndex & Google Genkit のネイティブ統合を活用
- 柔軟なクエリフィルタ機能 (複数範囲など) とクエリ プロファイリング ツールにより、コストとパフォーマンスに関する洞察を得ながら、より多くのことを達成
- 顧客管理の暗号鍵により、コンプライアンス要件を満たすプライバシーとセキュリティの層を追加

> “MongoDB Atlas から Firestore に移行した後、Firestore のシンプルさにより開発者の生産性が55% 向上し、サービスの信頼性が向上し、データベースDevOps から解放され、顧客にとって重要な製品イノベーションに集中できるようになりました。”
>
> Karan Agarwal, Director of Engineering at HighLevel



Database Migration Service の最新情報

一般提供

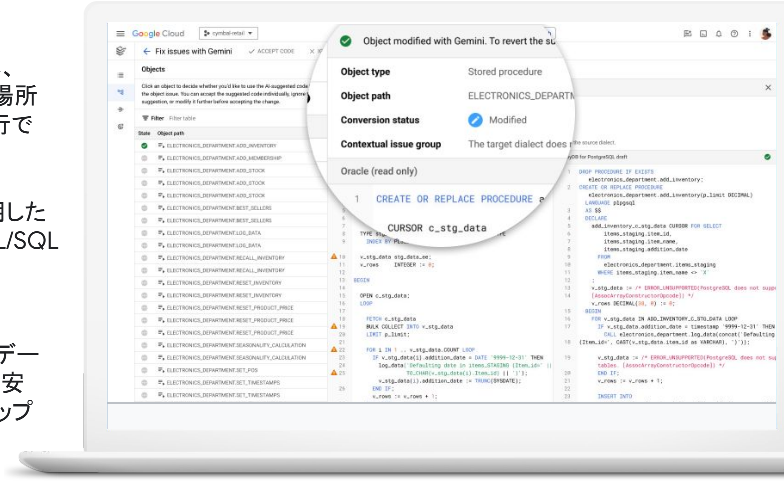
Oracle から PostgreSQL への移行は、Oracle ワークロードの移行を自動化し、コードとスキーマの変換を実行します。オンプレミスまたはクラウドのいずれの場所からでも、Cloud SQL for PostgreSQL および AlloyDB for PostgreSQL に移行できます。

プレビュー

Gemini 支援によるコードとスキーマの変換は、移行の最終段階を LLM を活用した変換で行うことで、データベース移行に革命を起こします。スキーマと複雑な PL/SQL コードを PostgreSQL に自動的に変換できます。

プレビュー

SQL Server の同種 DB 移行では、オンプレミスまたはクラウドの SQL Server データベースを Cloud SQL for SQL Server に移行します。DMS はサーバーレスで安全であり、高い忠実度とパフォーマンスを実現するためにネイティブのバックアップと復元を利用します。



> Google Cloud の Database Migration Service と Nordcloud のおかげで、PoC を成功させることができました。10TB の販売時点情報管理データベース全体をOracle から AlloyDB に移行および変換できました。バックフィルのプロセスは、高い効率で完了しました。全体的に、変換とデータ移行の両方における高度な自動化に驚きました。"
>
> Olivier Bandle, CEO Nordcloud Switzerland and Mario Schaniel, Head Software Engineering & Executive Board Member

Google Workspace & AppSheet

Customer Momentum

昨年 100 万人以上の人々と数千の企業が Gemini for Google Workspace を活用

1. **小売業者 : Sports Basement**
   Gemini を使用して、顧客回答の効率を向上し、メール作成に費やす時間を **30-35% 削減**

2. **小売業者 Woolworths**
   1 万人を超える従業員が Gmail でGemini を活用して、 業務の品を向上

3. **大手住宅ローン会社 PennyMac**
   Gemini を使用して、**採用、雇用、新入社員のオンボーディングを加速**

4. Docs や Gmail の「Help Me Write」機能を利用するエンタープライズ ユーザーの **70% が Gemini の提案を実際に使用**、スライドの「Help me visualize」で画像を生成したユーザーの**75% 以上が画像をプレゼンテーションに挿入**

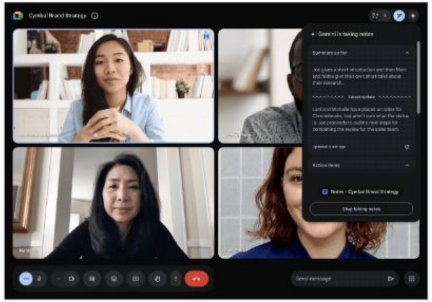
発表ハイライト

お客様の選択肢を広げるセキュリティ、会議に特化したAI アドオン製品の提供
Gemini を中心とした 3rd Party を含めたワークフローの自動化機能を新たに発表

1 新しいアドオン製品と新機能

AI Security add-on
- データの自動分類と保護
- プライバシー保護AI モデル

AI Meeting & Messaging add-on
- 議事メモ取得と翻訳
- Screenshare watermark
- Adoptive audio



2 ワークフロー自動化の新機能

HubSpot
- Gemini for Google Workspace の HubSpot 拡張機能

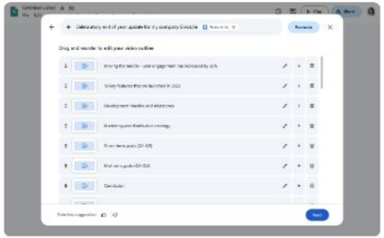
Vertex とGoogle Workspace の連携

AppSheet
- Google フォーム連携
- AppSheetのための統合コネクタ

3 コミュニケーションの革新

Google Vids
- 簡単な指示でDrive 上のコンテンツを利用して業務用ビデオを作成

Chat
- スペースの最大メンバー数を50 万に拡張
- Mio によるSlack やTeams との相互運用性向上



AI Security add-on for Workspace

一般提供

セキュリティの向上を目的としたAI 活用を推進するお客様向けのアドオンを新たに提供

アドオンに含まれる機能

- ドライブにAI 分類機能を搭載し、お客様独自のニーズに合わせてAI のモデルをトレーニング
- AI分類は、新規および既存のファイルを自動的かつ継続的に分類
- 分類されたファイルは、既存のデータ損失防止（DLP）コントロールで保護

AI Security add-on for Workspace: 1,130 円

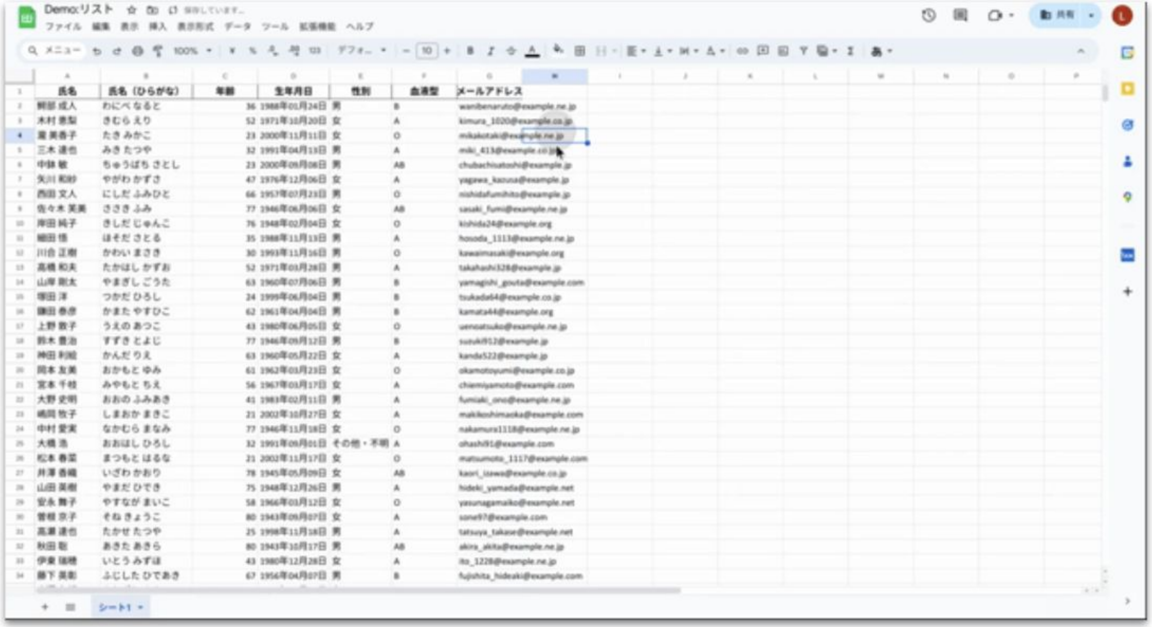
管理コンソールでAIモデルのトレーニングを設定する

Demo : AI による自動分類と保護

一般提供

従来ルールベースで対応できなかった企業固有の機密情報保護のニーズに対応

- 従来は機密ファイル分類には正規表現や定義済みのクレジットカード番号検知などルールベースで行われていた
- 新機能によりお客様独自の機密ファイルをベースにお客様固有の自動分類モデルをトレーニングし、Google ドライブ上の機密ファイルを自動的に分類して保護



AI Meetings and Messaging add-on

一般提供

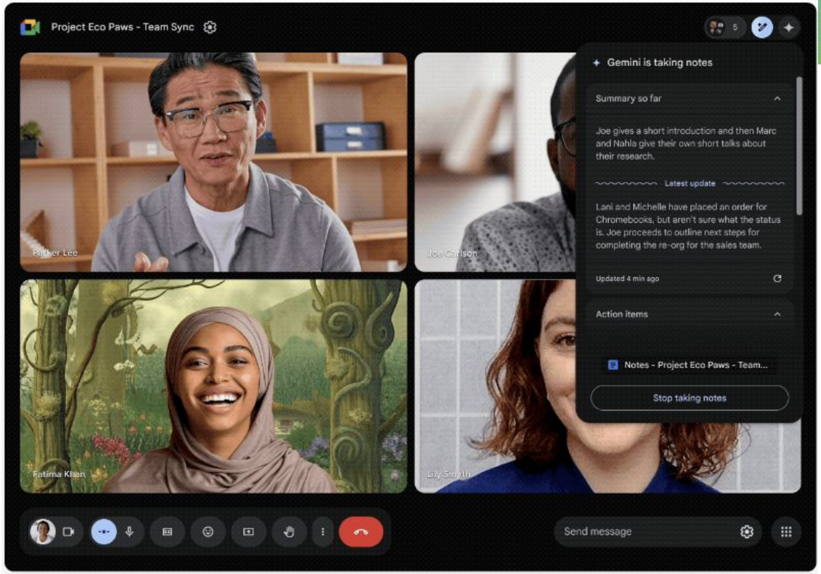
ミーティングの生産性向上を目的としてAI 活用を推進する
お客様向けのアドオンを新たに提供

アドオンに含まれる機能

- Studio look, studio sound, studio lighting
- 議事録の取得を自動化する Take notes for me
- 多言語の翻訳字幕
- 安全に画面共有するためのScreenshare watermark 機能
- 仮想背景の作成
- オーディオ最適化機能

など（後段にて解説）

AI Meetings and Messaging add-on: 1,130円

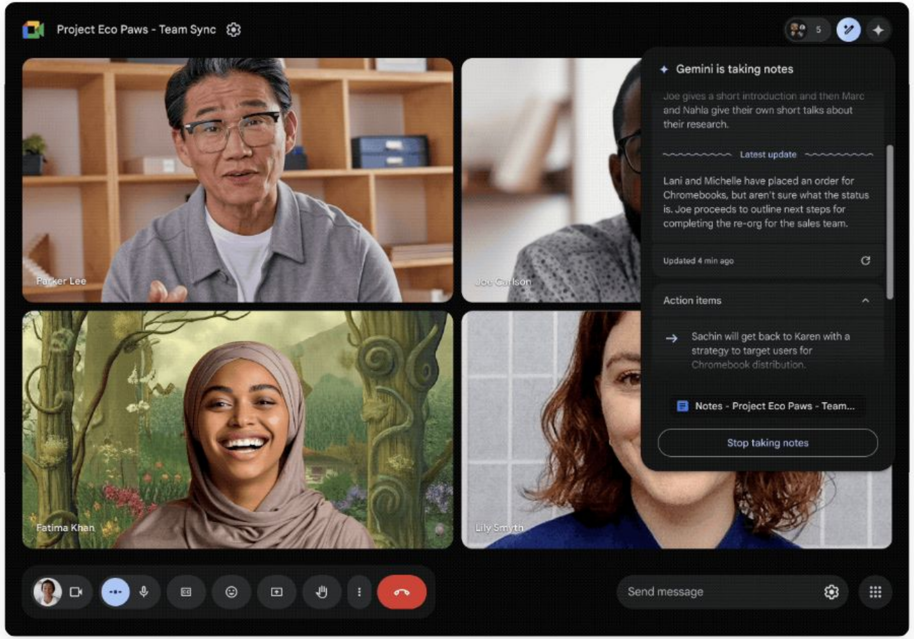


Take notes for me

プレビュー

議事録の取得を自動化し生産性を向上

1. 会議に集中する
2. 会議を中断させない
3. 何が話されたかを記録
4. Gemini に業務をオフロード

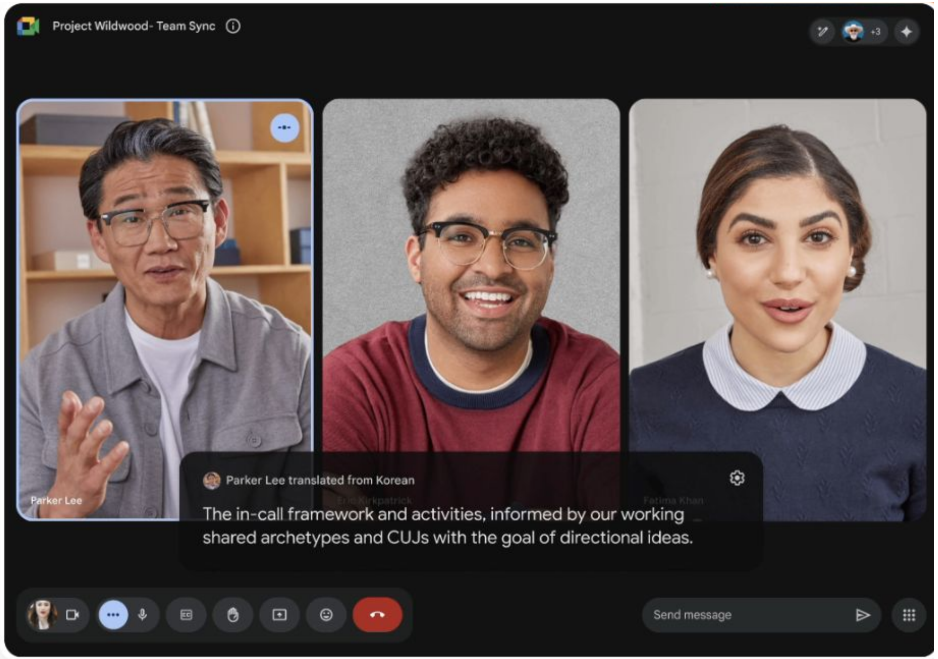


Translate for me

6月一般提供予定

新たに日本語、韓国語などを含む 69 の言語に対応した翻訳字幕の表示が可能に

1. 自動言語検出と翻訳
2. 69 言語に対応

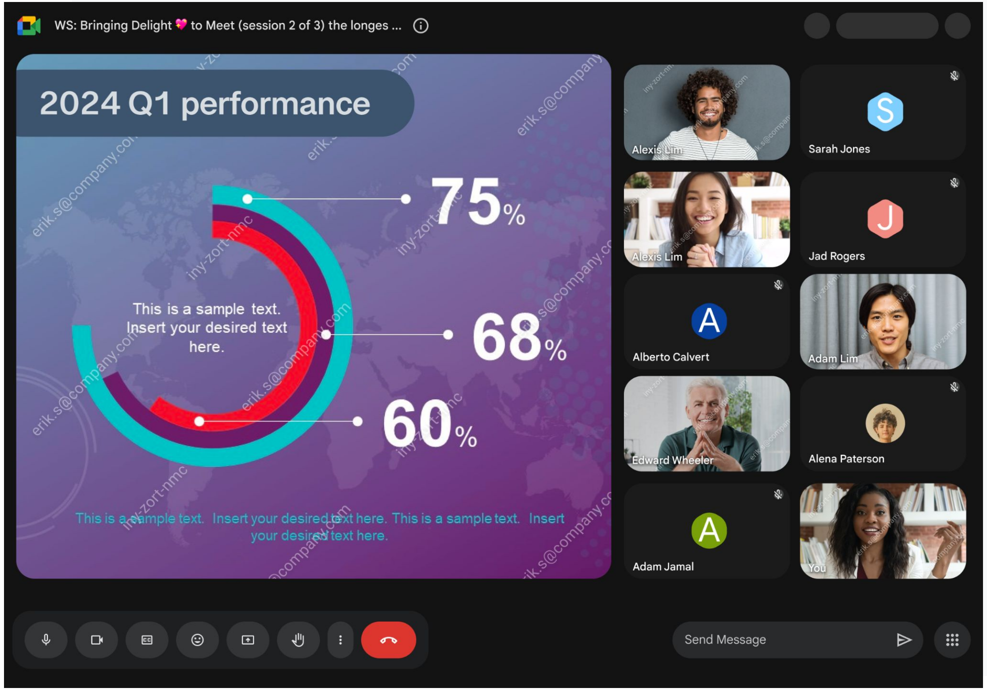


Screenshare watermark

ロードマップ

新たに透かしを導入することで会社の情報を安全に保ち、組織外への不適切な共有を防止

1. Meet の画面共有される情報を保護
2. スクリーン共有されたコンテンツに重ねて表示されたテキストにより、共有した人を追跡可能



Vertex とGoogle Workspace の連携

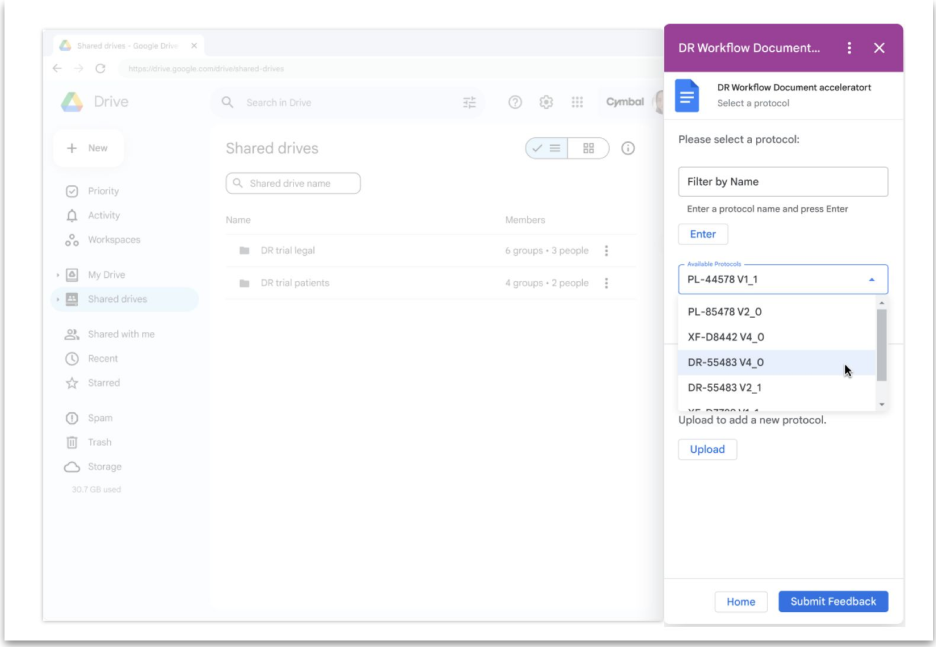
一般提供

時間を要していた業界固有のコンテンツ作成業務をVertex + Workspace で自動化

医療業界：

ブリストル マイヤーズ スクイブ

- Google Workspace の拡張機能として Vertex AIと連携するアドオンを構築
- インフォームド コンセント フォーム (ICF) をはじめとする 80 種類の異なる文書に対する取り組みを自動化
- 数週間かかっていた ICF 作成作業が、ほんの数分でそれができるように



HubSpot Gemini Integration

ロードマップ

Gemini for Google Workspace が初となる 3rd Party とのコネクタを発表、
業務プロセスのより多くの範囲の自動化、効率化が可能に

Gemini for Google WorkspaceのHubSpot 拡張機能は、マーケティング担当者と営業担当者が Workspaceからより多くのことを行えるようにします

ユーザーは、Gemini for Google Workspace を使って、自然言語を使用して、HubSpot Sales Hub データと Marketing Hub プロジェクトを Docs または Gmail で更新できるようになります。

Gemini for Google Workspace

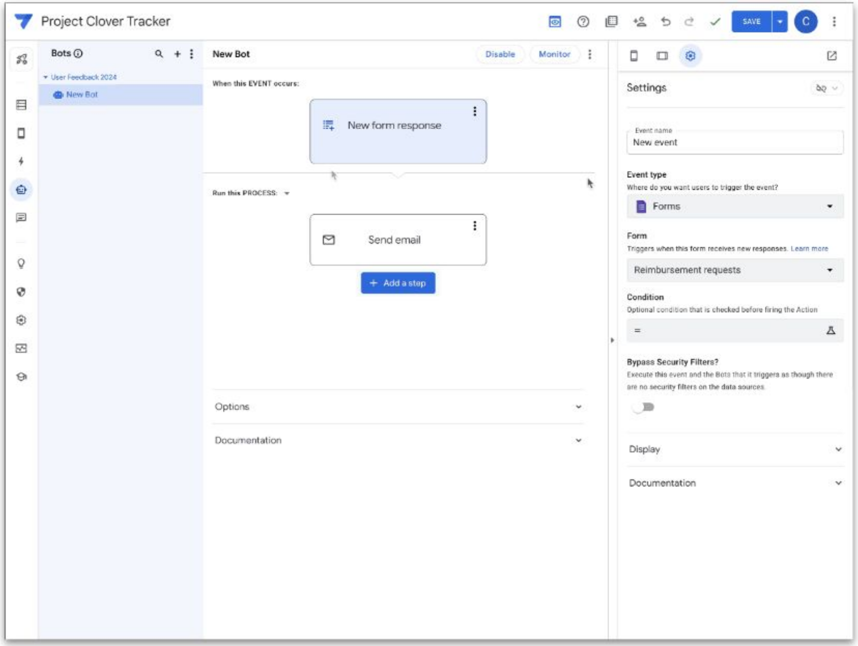
HubSpot

Google Forms を AppSheet に統合

プレビュー

以前は Google Apps Script を使用して構築する必要があった Google Forms と連携したワークフローをノーコードで構築可能に

Google Forms の入力をトリガーとした、Gmail や Google チャットにアクション可能な通知を送信するオートメーションをノーコードで構築することで、承認、チケット、その他さまざまなシナリオを自動化するワークフローを簡単に作成できるようになります

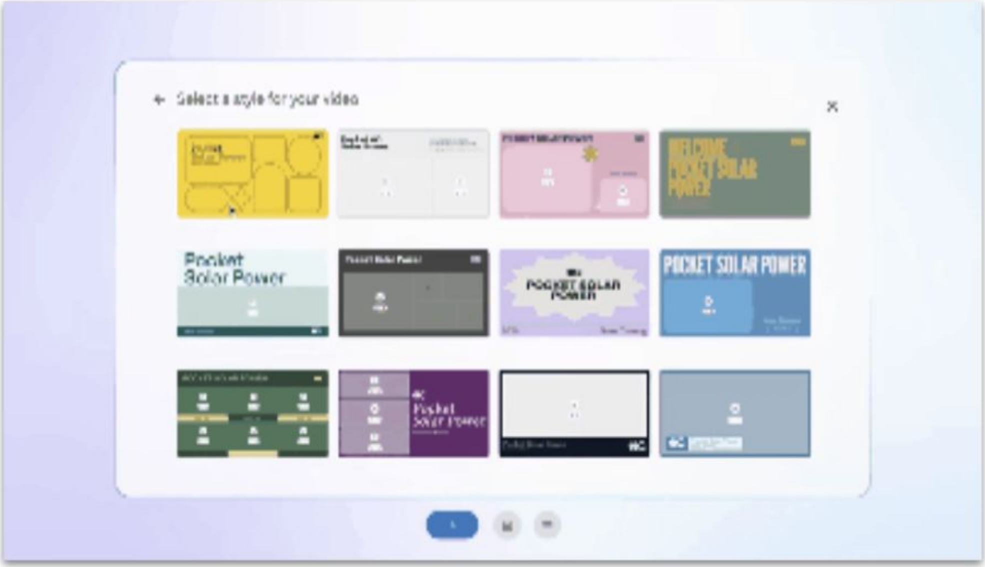


Google Vids

6 月より
プライベート プレビュー

Google ドライブ上のファイルとテキストによる指示で、並外れたビデオ プロデューサーや脚本家として業務用ビデオを作成

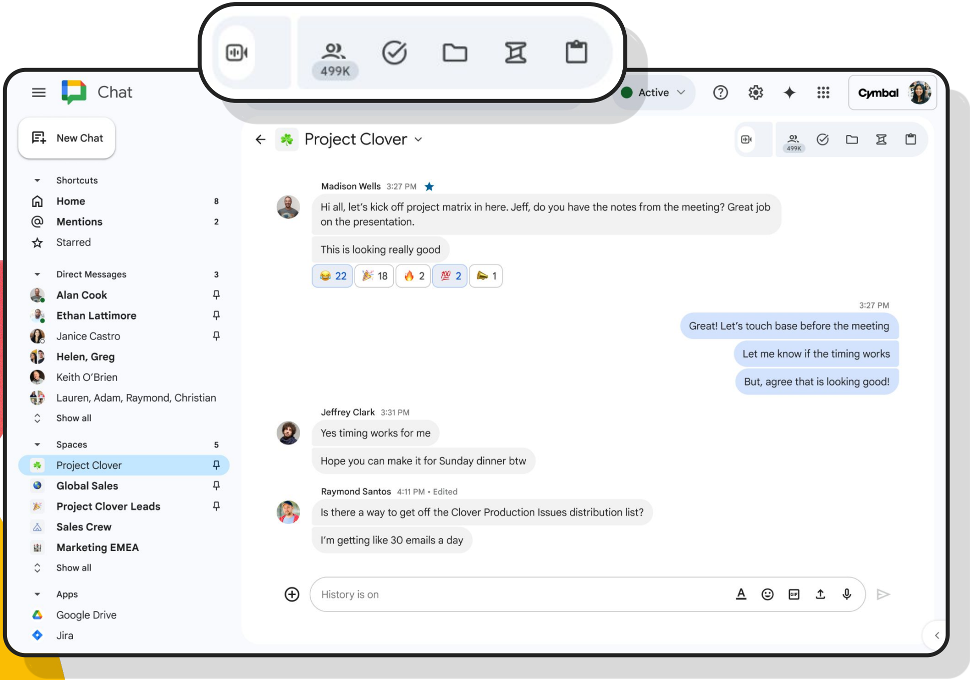
- 簡単なプロンプトでスタート
- Recording Studio でパーソナライズ
- 何百万もの高品質なストックビデオ、GIF、音楽などを使ってシーンを充実させる
- AI を活用した編集機能で制作をスピードアップ



最大 50 万人を Space に追加できるように

スペースの参加者の上限が 5 万から 50 万に拡張され、
Google Chat で全社的なコミュニケーションを実施できるようになります

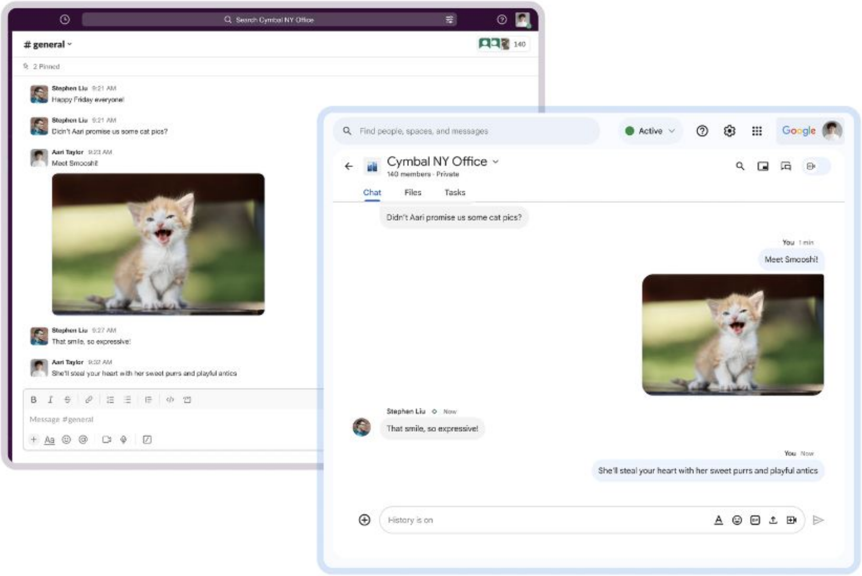
数週間以内に一般提供



チャットの相互運用性

一般提供

Mio との提携により、Chat、Slack、Teams 間のメッセージ相互運用を実現しシャドー IT とデータサイロを削減



Security

発表まとめ

Gemini in Security

- Gemini in SecOps[一般提供]
- Gemini in Threat Intelligence[プレビュー]
- Gemini in Security Command Center[プレビュー]
- Gemini Cloud Assist[プレビュー
  - IAM Recommendations
  - Key Insights
  - Confidential Computing Insights

Security Operations

- Applied Threat Intelligence in SecOps [一般提供]
- Mandiant Hunt for Security Command Center Enterprise [一般提供]

Chrome Enterprise

- Chrome Enterprise Premium [一般提供]

Trusted Cloud

- Cloud NGFW Enterprise [一般提供]
- Cloud Armor Enterprise [一般提供]
- Privileged Access Manager [プレビュー]
- Principal Access Boundary [プレビュー]
- Confidential VMs on Intel TDX[プレビュー]
- Sensitive Data Protection Integration with Cloud SQL [GA]
- Autokey [プレビュー
- Regional Controls for Assured Workloads [プレビュー
- Audit Manager [プレビュー