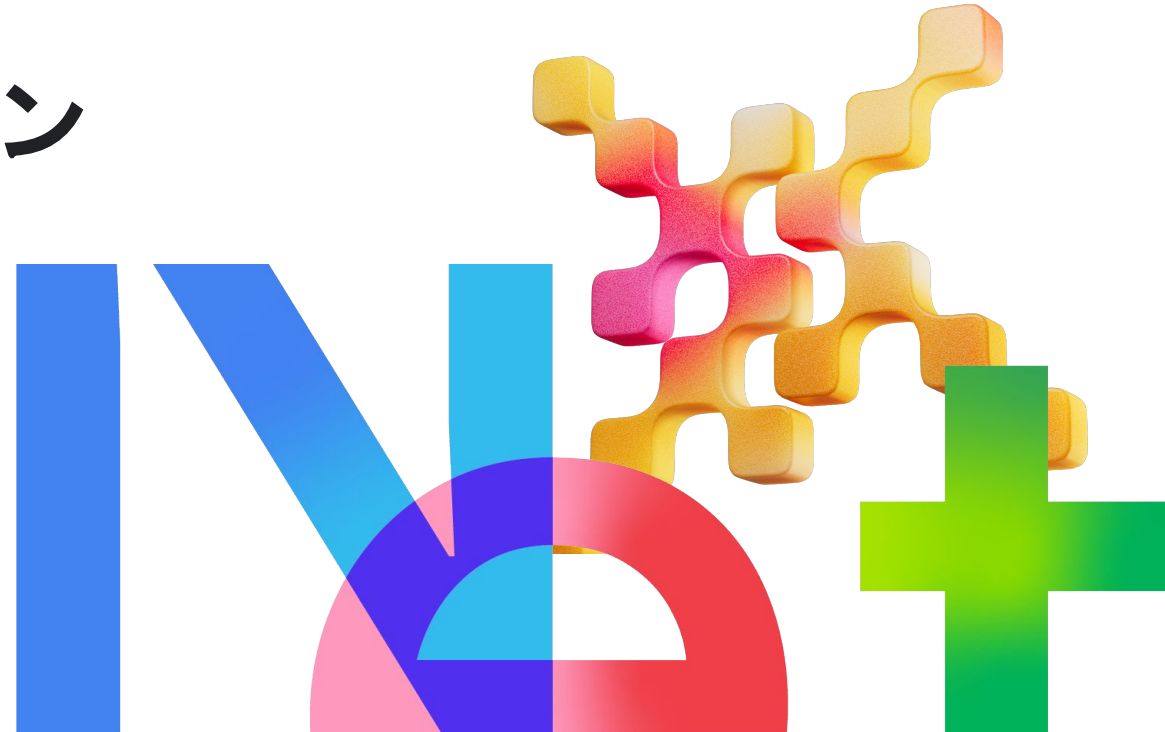
Japan Session and Reception

サマリー セッション

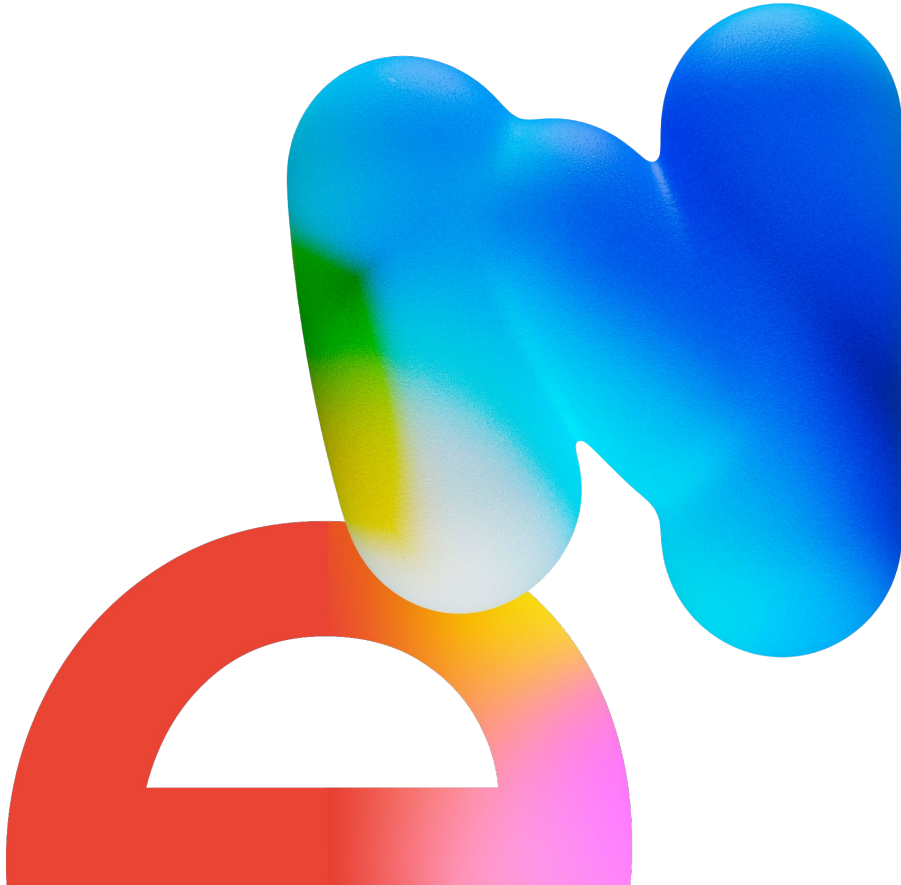
April 24, 2026

Google Cloud Next 26



Google Cloud Next 26

Key Message



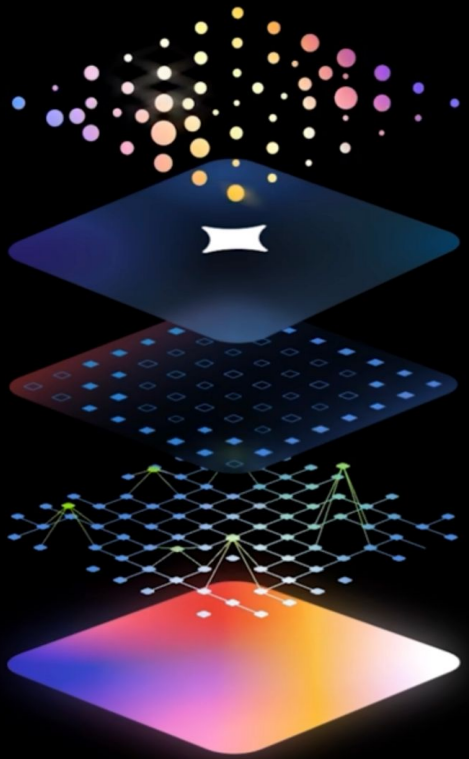
Mission

Accelerate transformation with Google's leading, AI-powered Cloud technology

Agentic Enterprise Blueprint

Only Google

最もセキュアなクラウド上に構築された、完全なデータ & AI スタックを提供します



Agentic Taskforce

Agentic Platform and Models

Agentic Defense

Agentic Data Cloud

AI Hypercomputer

Agentic Enterprise の実現の為に

パートナー様への投資

$750M

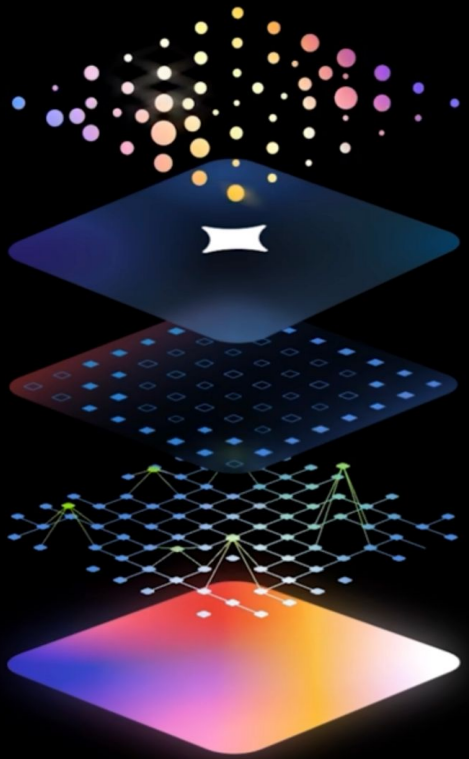
エージェンティック AI の開発を加速

パートナー様への FDE のエンベディング

Google は、主要なパートナー様に Forward-deployed engineers: FDE のチーム立ち上げを支援します

Only Google

最もセキュアなクラウド上に構築された、完全なデータ & AI スタックを提供します



Agentic Taskforce

Agentic Platform and Models

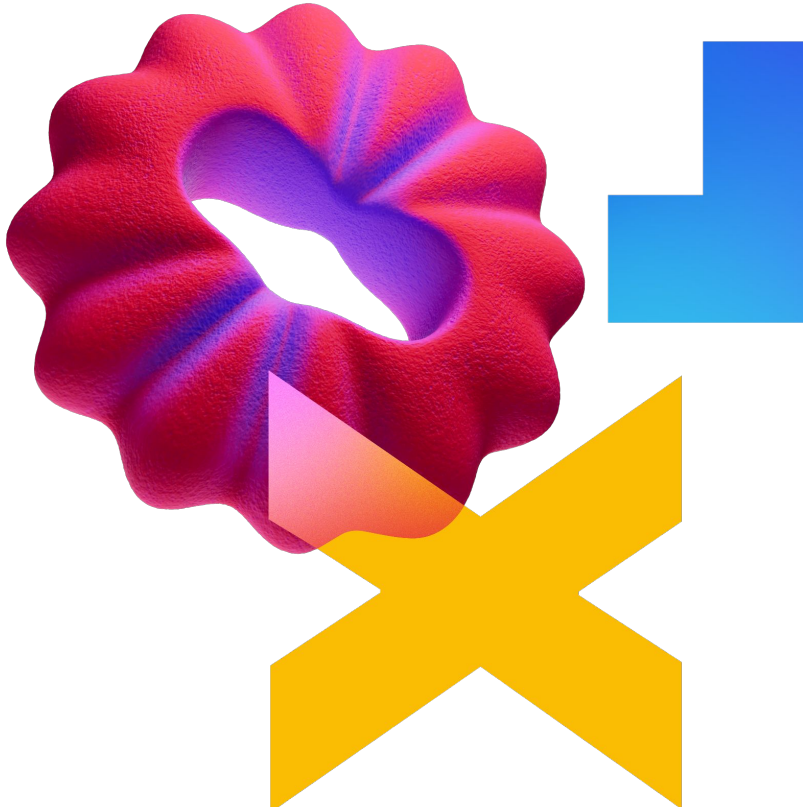
Agentic Defense

Agentic Data Cloud

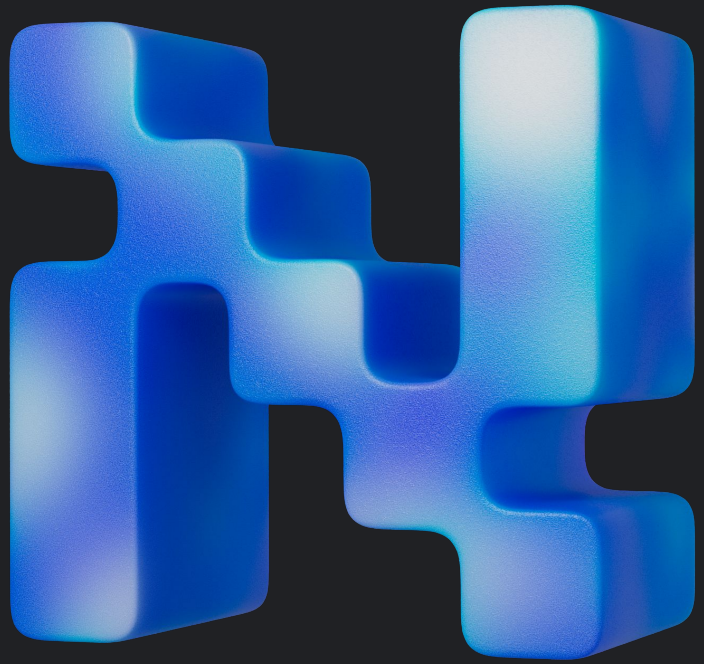
AI Hypercomputer

Google Cloud Next 26

Gemini Enterprise



01.
モメンタムと Update 全体像



単なるチャットボットの時代を終え、自律的に業務を遂行する「エージェント エコシステム」の時代へ

2025 年のモメンタム

変化する AI プラットフォームの要件

質の高い Agent を開発するための ツール要件は？

企業環境を安全に接続するにはどうすればよいか？

どうすれば容易に費用対効果の高い方法で AI をScaleできるか？

企業は、概念実証（ POC）や単純なチャットボット、 RAG の段階を超え エージェント システムの導入を支援する機能を必要としている

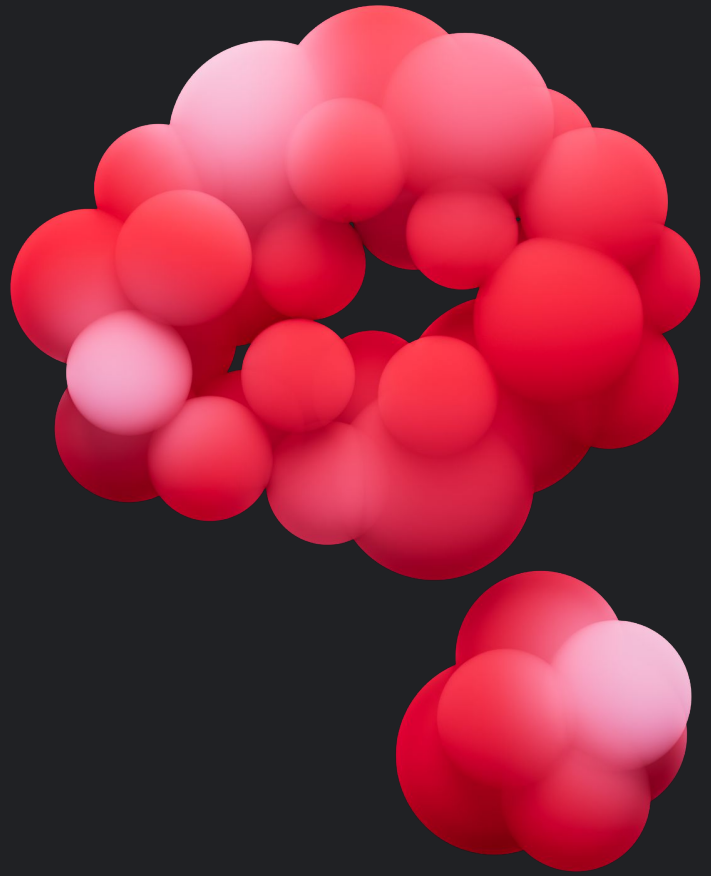
ブランドの進化と「 Gemini Enterprise Agent Platform」の誕生

エージェント時代を全社規模で支えるためにGoogle Cloud の
AI ポートフォリオは「Gemini Enterprise」という統合ブランドに進化

Gemini Enterprise

Gemini Enterprise App （旧称 :Gemini Enterprise）	Gemini Enterprise Agent Platform （旧称 :Vertex AI）	Gemini Enterprise for Customer Experience （旧称 : Customer Engagement Suite）
従業員向けアプリ	開発者向けプラットフォーム	顧客体験の向上

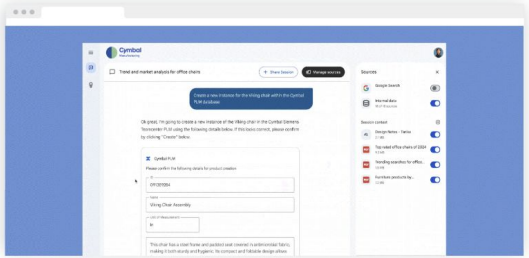
02. Gemini Enterprise App



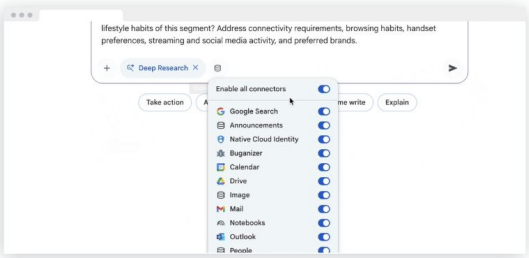
Gemini Enterprise app

職場における AI への「玄関口（フロントドア）」

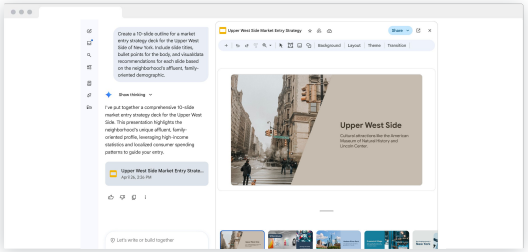
厳選されたデジタル　タスクフォース（専門チーム）と共に、
確かな成果を生み出します



永続的なチームメモリ
（共有蓄積された記憶）を活用し、
連携を深めます



すべての従業員が
独自のカスタム　エージェントを
構築できるようにします



デジタル ワークフォースを一元管理

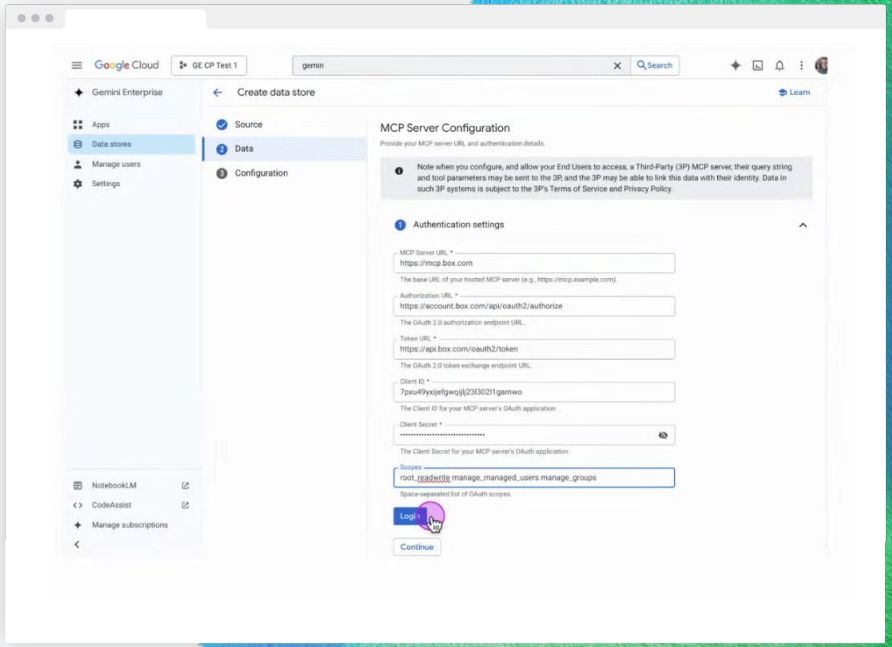
プレビュー

MCP サーバー コネクタ

独自の MCP サーバー（BYO-MCP）へ直接接続し、エージェントのデータソースや実行ツールとしてシームレスに活用

この機能の利点：

- データを複製・集約する手間が省けるため、コストを抑えつつ常に最新のビジネスデータへアクセスできます
- 管理者が一度サーバーを登録すれば、一般の従業員でもノーコードで自作のエージェントに社内データを組み込めます
- データを外部へコピーせず、元のシステムに留めたまま連携するため、厳格なセキュリティ要件を満たすことができます



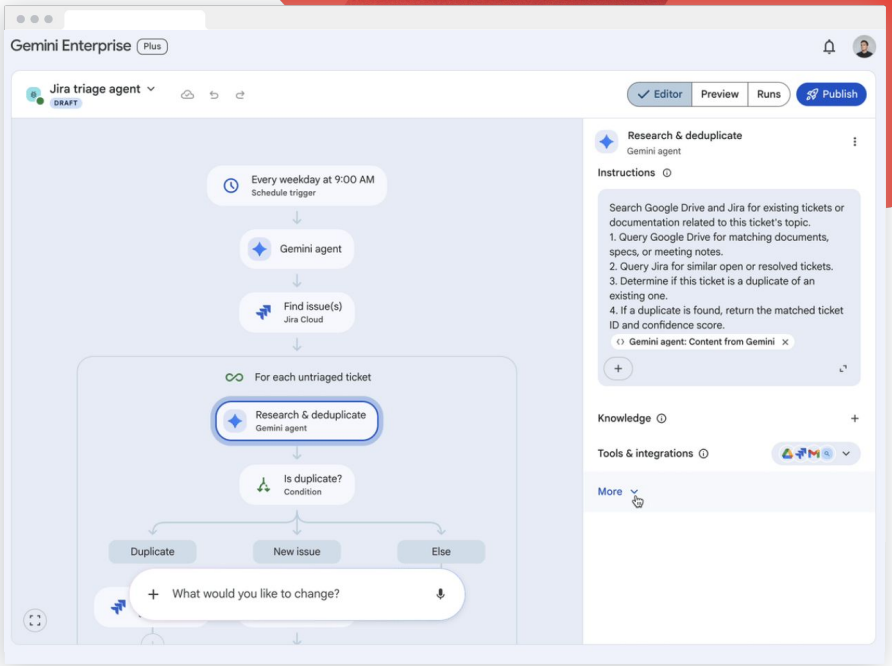
プレビュー

エージェント デザイナー v3

コードを書かずに従業員がエージェントを作成、テスト、公開できる

この機能の利点:

- 非エンジニアが自然言語でエージェントやワークフローを作成できる
- 実行順序の確定や条件分岐などにより、より決定的にタスクをこなせる
- Gmail や カレンダーのイベント及びスケジュール実行を組み合わせた自立型ワークフローをサポート
- MCP サーバーをコネクタとして使用し、企業のデータを使用したエージェントを構築できるように



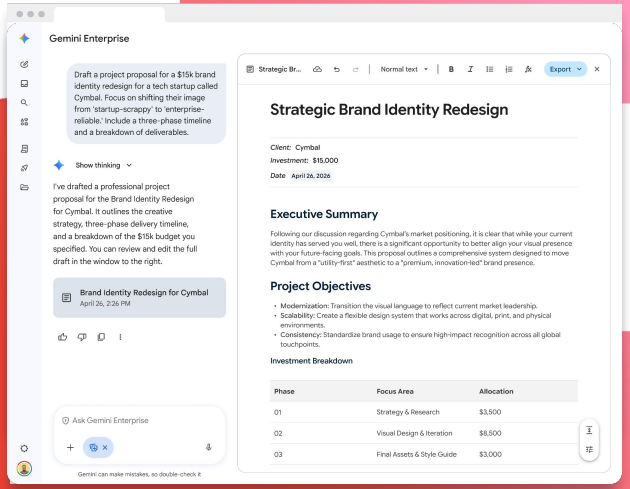
プレビュー

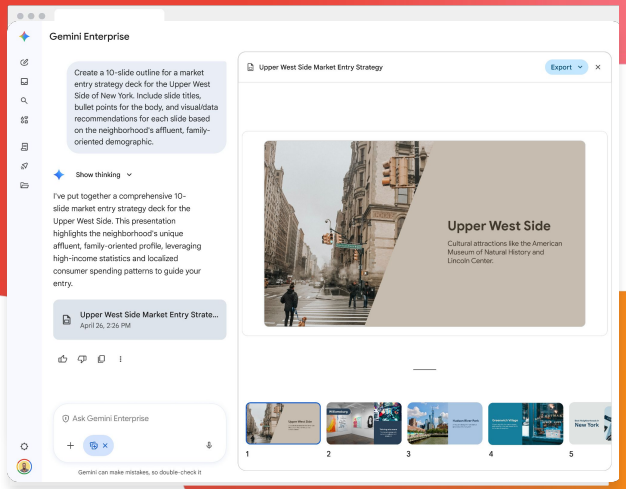
Canvas によるドキュメント、スライドの作成
GWS および M365 での編集

スライドや文書の作成・アウトプットの品質が向上

この機能の利点：

- 編集可能なスライドの生成と Google スライド及び Powerpoint 形式でのエクスポートに対応しました。
- 編集可能なドキュメントの生成と Google ドキュメントと Word のエクスポートに対応しました





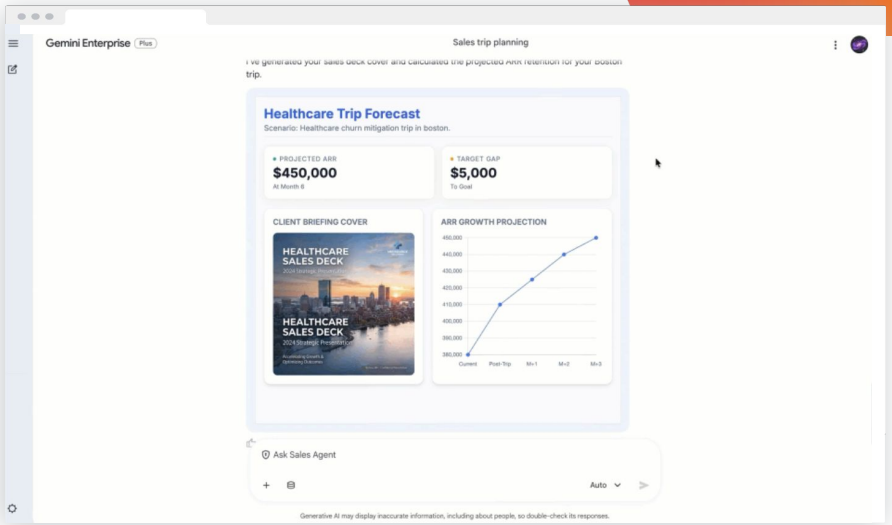
プレビュー

インタラクティブな UI(A2UI)の対応

Agent-to-User Interface (A2UI) プロトコルに対応し、単なるチャットを越えたリッチなUI 体験を提供

この機能の利点:

- 単なるテキストにとどまらず、ドロップダウン、入力フォーム、日付ピッカーなどのカスタム UI ウィジェットをチャット画面に表示できるように
- グラフ描画のコンポーネントを使用して単なる画像ではない、動きのあるリッチな表現が可能に



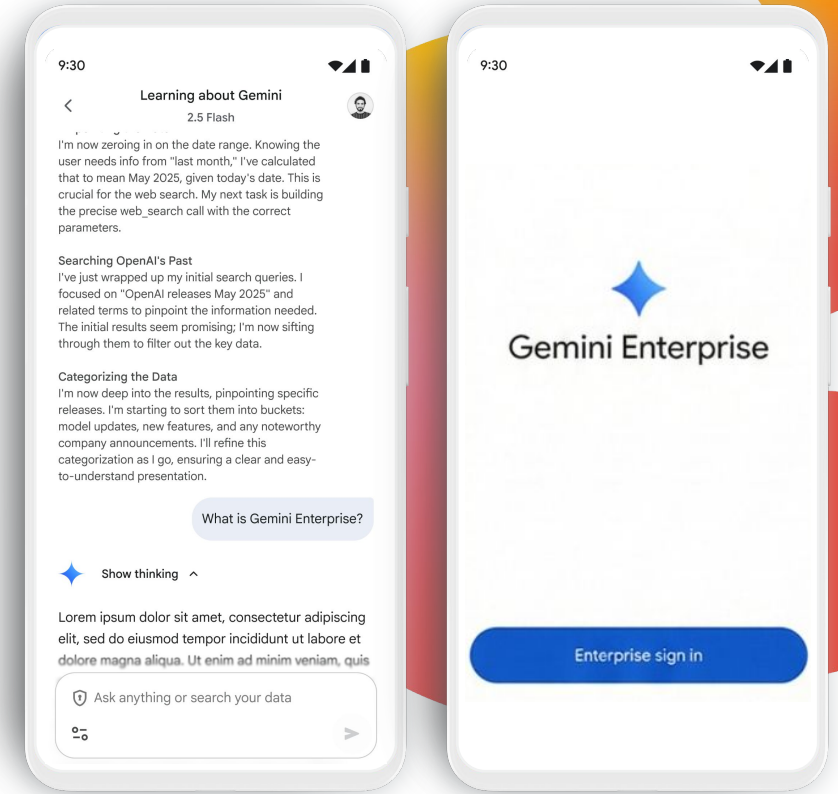
プレビュー

Gemini Enterprise モバイルアプリ

iPhone または Android 端末で、外出先から Gemini Enterprise アプリにアクセス可能

この機能の利点：

- 音声入力に対応し、文字を打つことが難しい現場作業員の方でも簡単にアクションや企業データの情報収集が可能になります
- 電車やタクシーなどでの移動中に資料の調査や Agentic なタスク実行が可能になります



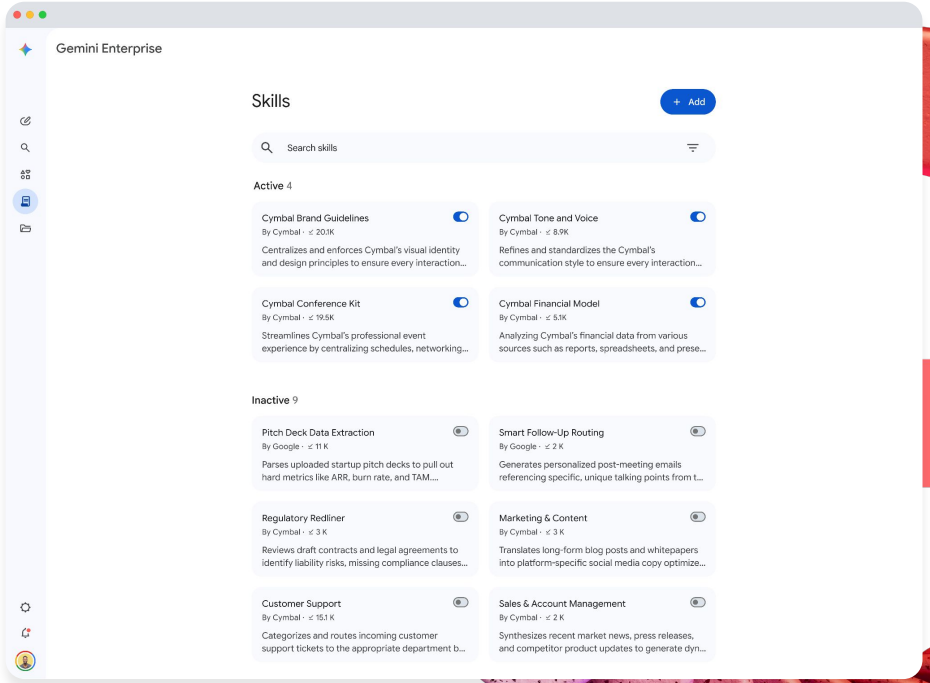
プレビュー

スキル

反復的なプロンプト入力の手間を省く「プロンプトとエージェントの中間」に位置し、特定のタスクを標準的かつ一貫した方法で実行させるための機能を導入

この機能の利点：

- コアのシステム プロンプトを複雑にすることなく、タスクに必要な時だけ深い専門知識や指示を呼び出して再利用
- 作成した指示セットは他のユーザーと簡単に共有できるため、個人のノウハウをチームや組織全体へ Scale



03. Gemini Enterprise Agent Platform

Gemini Enterprise Agent Platform

プレビュー

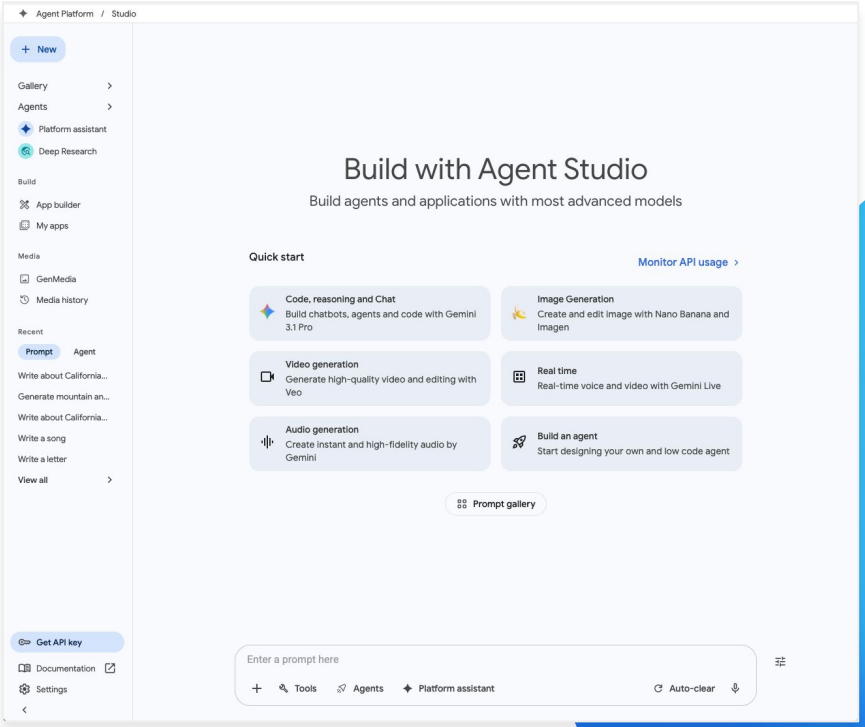
Build

Agent Studio

Agent Platform に直接組み込まれ、仕様（Spec）ベースで AI エージェント開発・共同作業を可能に

この機能の利点：

- Agent Platform 上で直接、エージェント駆動型アプリケーションの構築やチーム共同作業が行えます
- アプリケーションの仕様に合わせて、組み込むモデルやツールを自由に選択
- VS Code などへのお好みの IDE へプロトタイプのエクスポートが可能



ベータ

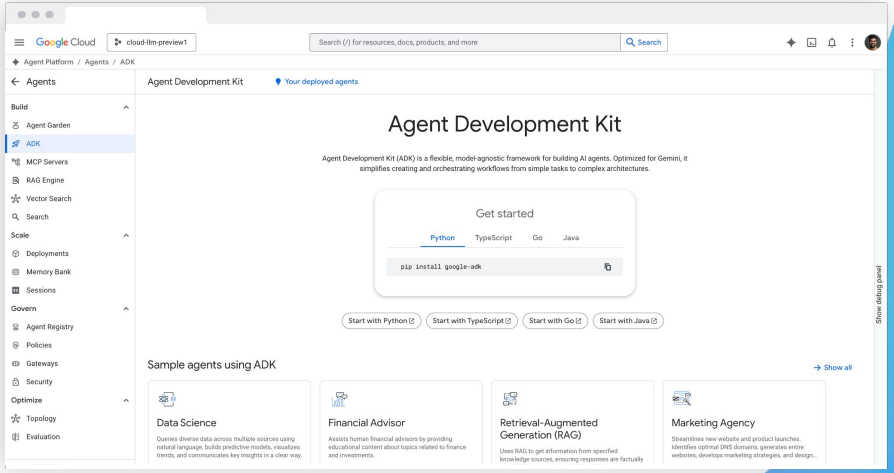
Build

Agent Deployment Kit 2.0

エージェント スキルのサポート、
グラフベースのオーケストレーション
などが可能に

この機能の利点：

- グラフベースのワークフロー：
  タスクのルーティングと実行方法をより
  詳細に制御できる、決定論的なエージェント
  ワークフローを構築できるように
- エージェント スキルのサポート：
  エージェントはドメイン固有のスキルを
  動的に検出し実行できるように



GA

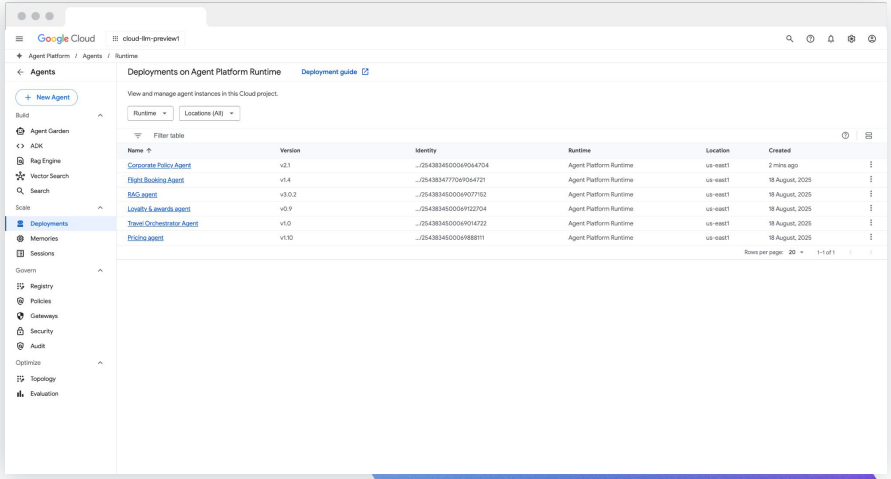
Scale

Agent ランタイムの改善

開発効率と実行パフォーマンスを劇的に向上させ、大規模な AI エージェントの運用を可能にするアップデート

この機能の利点：

- 即時起動と大規模拡張：
  1 秒未満の高速起動とプロジェクトあたり
  最大 3,000 エージェントの同時稼働を実現
- 柔軟な開発環境の提供：
  BYOC や主要 4 言語に対応し、ローカル環境とクラウドの完全な同期により開発を効率化
- 高度な運用管理と安定性：
  最大 7 日間の長期実行、詳細な権限設定



顧客企業：Comcast、CVS、AT&T、Paypal、United Health Group、楽天、マッコーリー銀行

一般提供開始

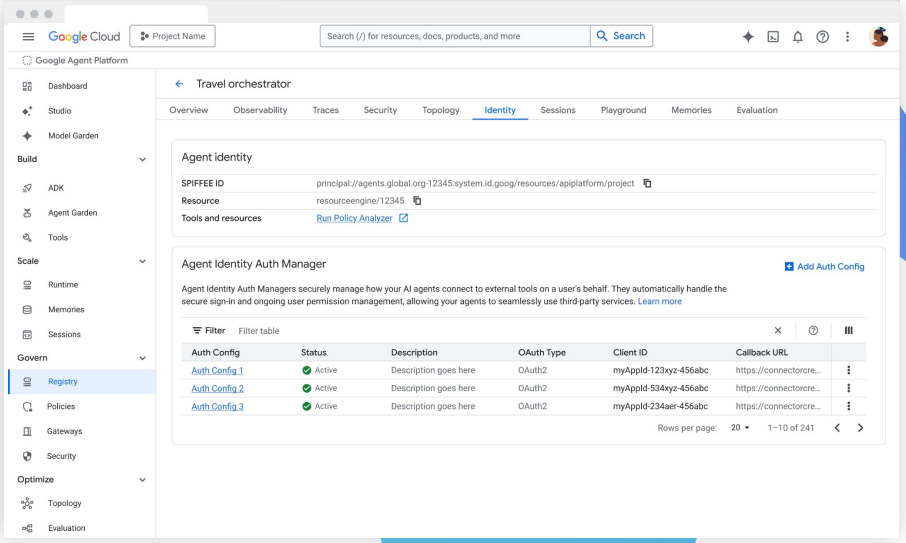
Govern

Agent Identity

GCP のアクセス制御とネイティブに
統合され、エージェント固有の
識別子によりセキュアな権限管理を実現

この機能の利点：

- IAM の許可 / 拒否、VPC-SC、文脈認識アクセスなど GCP の主要セキュリティ機能とシームレスに連携
- SPIFFE ID による暗号化認証を自動化し、Agent Engine を含む多様な環境で安全な識別を可能にします
- トークンをランタイムに直接紐付けることで、万が一の流出時にも不正利用を効果的に防止



プレビュー

Govern

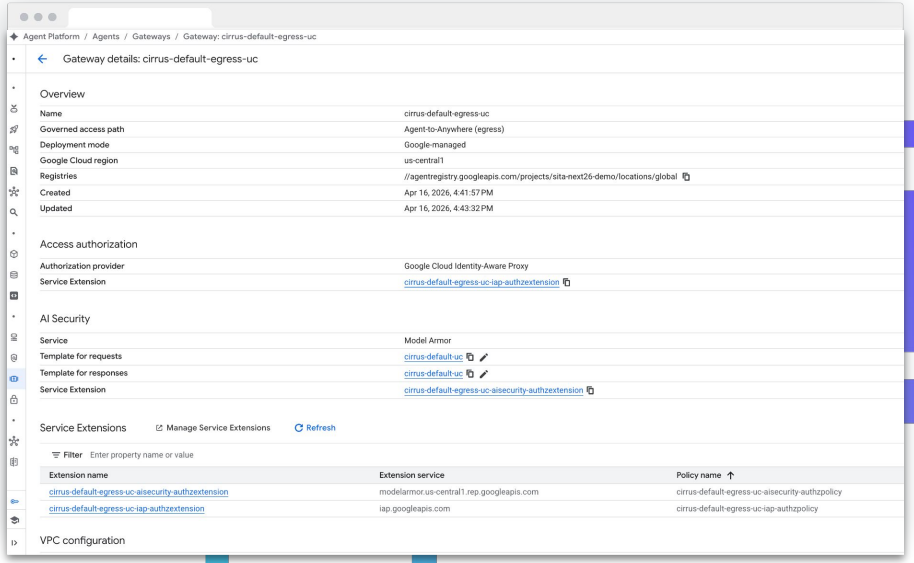
Agent Gateway

組織内のすべての AI エージェントの
トラフィックや権限、アクセス制御を
単一のポイントで中央管理

この機能の利点：

- Google Cloud IAM などと連携し、
全エージェントの通信や外部ツールへの
アクセスに対して、きめ細かく一貫した
セキュリティ ポリシーを強制・適用

- トラフィックがゲートウェイを通過するため、自動化
されたログと Trace ID によって
エージェントの全てのアクションを可視化



プレビュー

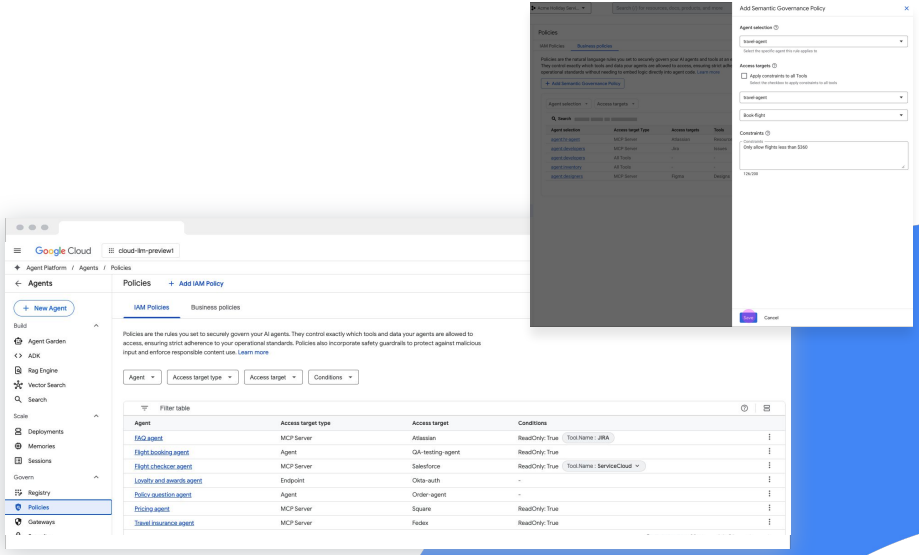
Govern

Governance Policies

エージェントのアクセスとアクションを
統制するための、意図を考慮したポリシー。
Agent Gateway を通じて適用

この機能の利点：

- IAM ポリシーにより承認済み
  エージェントのみを許可し、
  未登録の接続をブロック
- 自然言語で定義したビジネスルールを、
  ユーザー操作の文脈に合わせて
  動的に適用



プレビュー

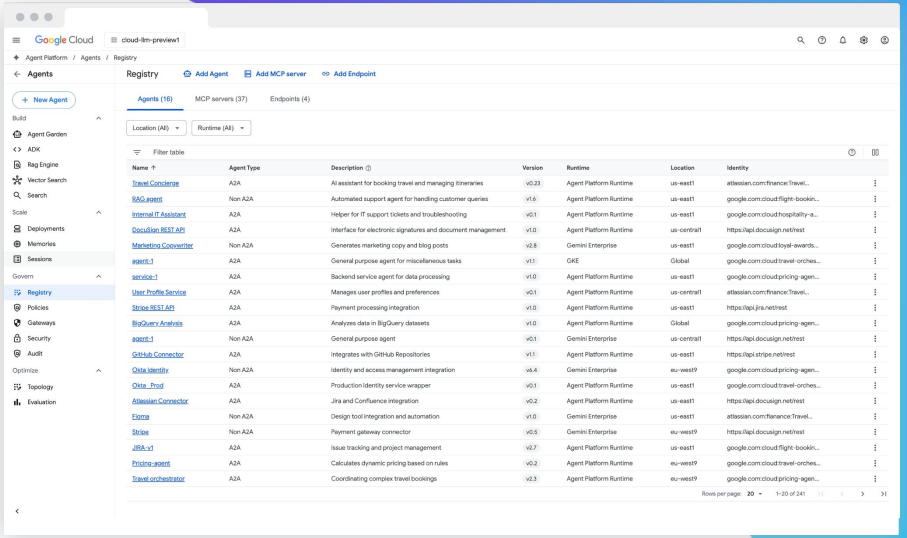
Govern

Agent Registry

組織内のすべての AI エージェント、ツール、MCP サーバーを登録して一元管理する中央カタログ機能

この機能の利点：

- 3P を含む既存のエージェントやツールを容易に発見し、再利用を促進できる
- AI資産の利用状況、コスト、コンプライアンスを横断的に把握
- ポリシーと統合し、登録状況等に基づく厳密なアクセス制御を適用



エクスペリメンタル

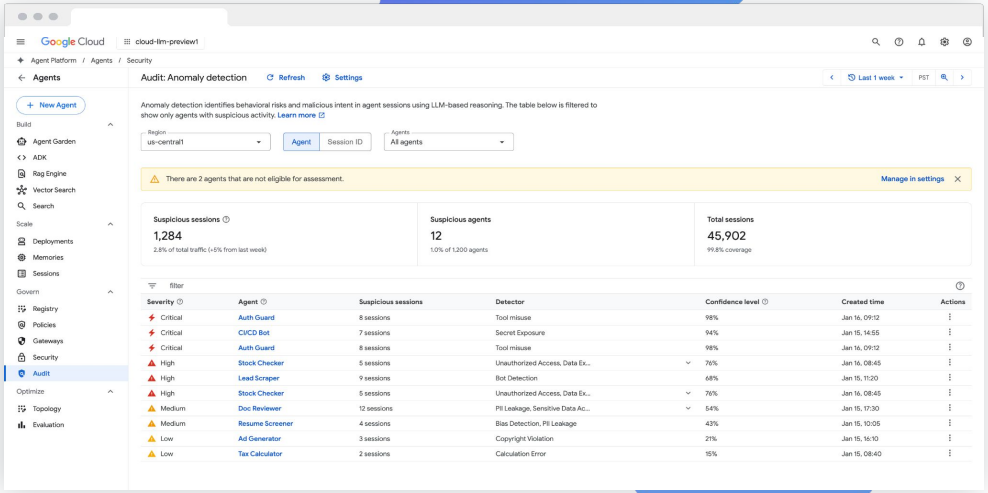
Govern

Agent Anomaly Detection

エージェントの実行時動作を監視し、ツールの不正使用を検出

この機能の利点：

- 統計解析と LLM の論理的判断により、エージェントの推論と活動をリアルタイムに監視
- 複数セッションにわたる動作を時系列で分析し、不審な挙動やツールの悪用を特定
- 実行時のアクティビティを常に先回りしてチェックし、実害が出る前に異常を自動で検知



プレビュー

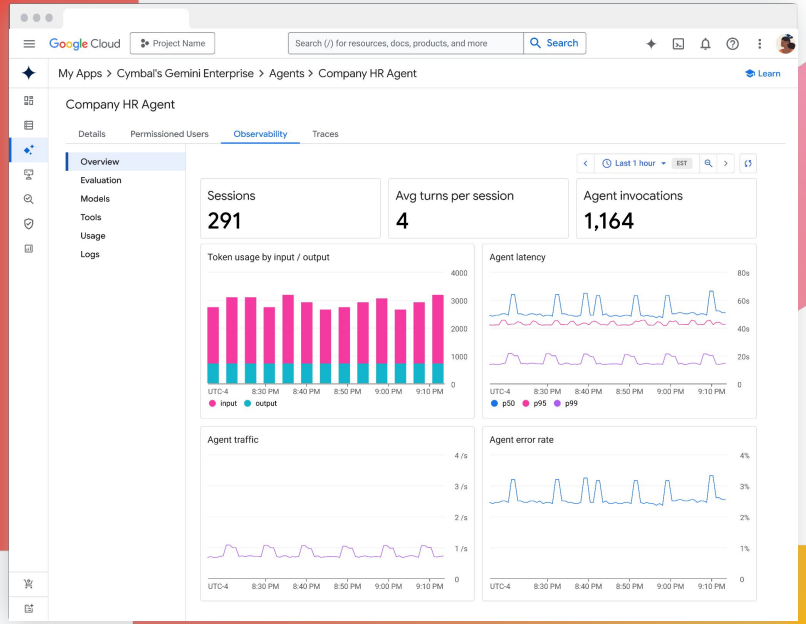
Optimize

Agent Observability

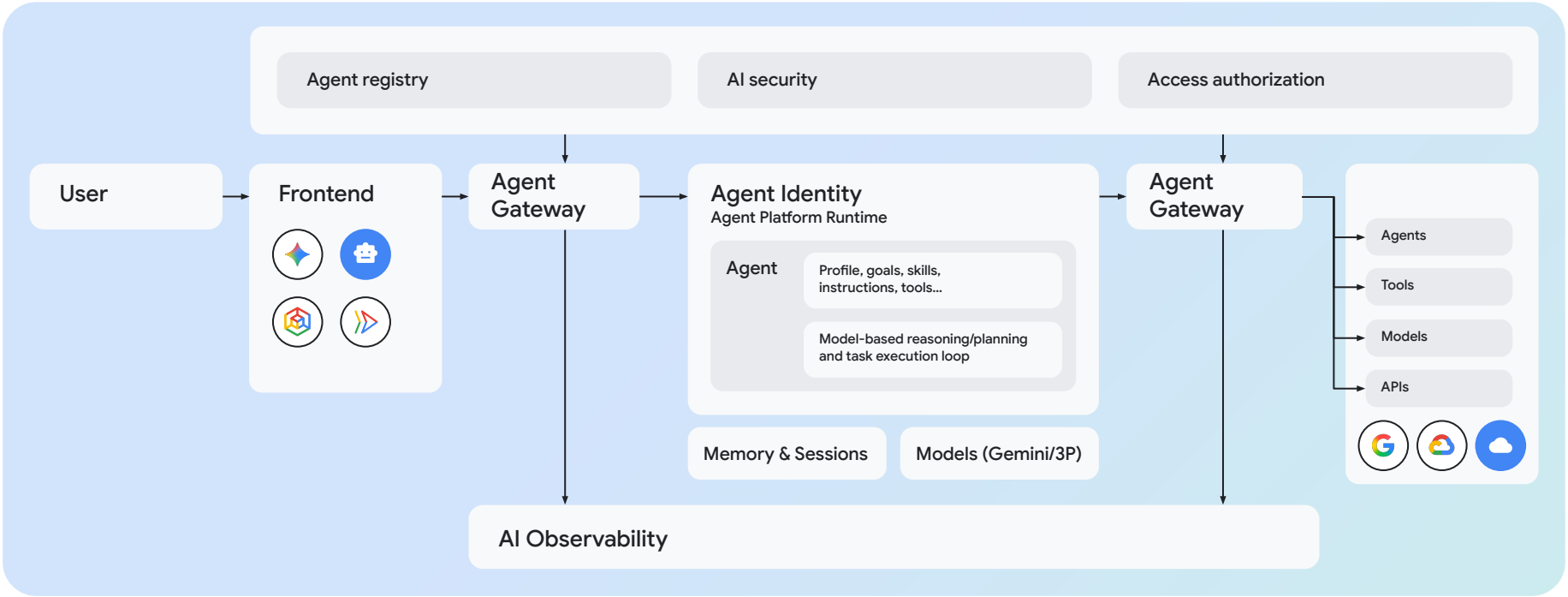
エージェントの挙動を可視化・追跡してパフォーマンスを最適化するための監視機能

この機能の利点：

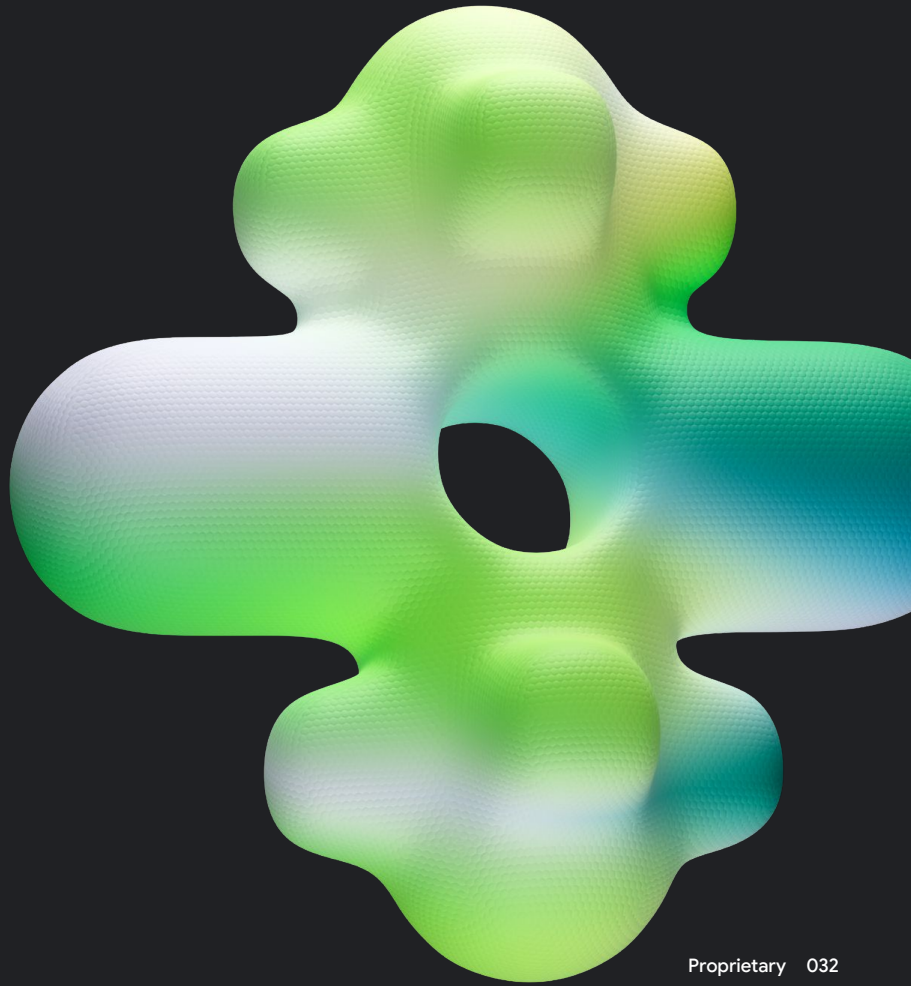
- エージェントの呼び出し、クエリ、レスポンスを記録し、コンプライアンスと安全性を担保
- 使用モデルやツール、稼働状況をエージェント単位で集約・可視化
- DAG 形式でステップごとの処理やレイテンシを追跡し、ボトルネックの迅速な特定を可能に



アーキテクチャ


Agent registry
AI security
Access authorization
User
Frontend
Agent Gateway
Agent Identity
Agent Platform Runtime
Agent
Profile, goals, skills, instructions, tools...
Model-based reasoning/planning and task execution loop
Memory & Sessions
Models (Gemini/3P)
Agent Gateway
Agents
Tools
Models
APIs
AI Observability

04.
エージェントの頭脳とツールの進化



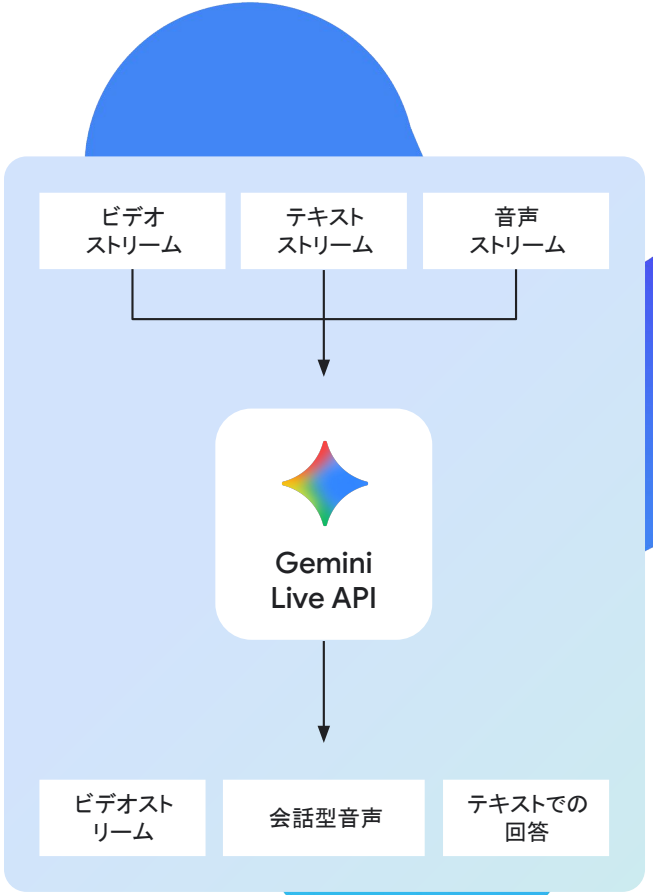
プレビュー

Gemini 3.1 ライブ APIと ライブアバター

音声・ビデオを直接処理することで、
人間のように自然かつ低遅延な対話を実現する
次世代のマルチモーダルモデル

この機能の利点：

- テキスト変換を介さず音声・映像を直接処理し、
  90 以上の言語で即応性の高い対話を可能に
- 声のトーンや感情の機微を認識し、
  人間同士のような流暢でリアルな会話体験を提供
- リアルタイムのビデオ生成と正確なリップシンク
  により、視覚的にも自然なライブアバターを実現



プレビュー

Gemini 3.1 TTS Flash

高度な感情表現と200 種以上の音声タグにより、
多言語で人間味のある自然な対話を高速に生成する
音声合成モデル

この機能の利点：

- 200 種類以上の音声タグ（笑い声、ささやき等）に
より、感情やペースを自然言語で精密に制御できます
- モデルの進化により、長文や複数回の対話でも
声質が変化せず、安定した高品質な音声を維持します
- 日本語を含む 24 言語で最高品質を実現し、
計 70 以上の言語で超低遅延な音声出力を提供します

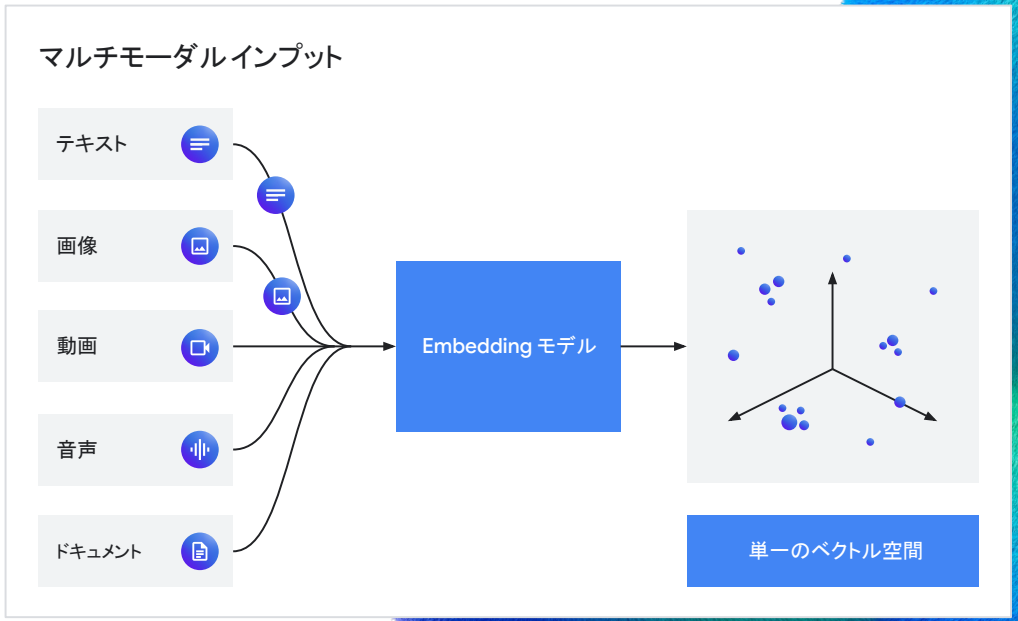
一般提供開始

Gemini Embedding 2

テキスト、画像、動画、音声、PDF の全 5 モダリティをネイティブに統合し、単一のベクトル空間で検索可能

この機能の利点:

- 音声や動画、最大 6 枚の画像を中間変換なしで直接取り込み、モダリティを跨いだ高度な検索や分析を可能にします
- 最大 8,192 トークンのテキストや 120 秒の動画、6 ページの PDF を一度に処理し高精度な特徴抽出を行います



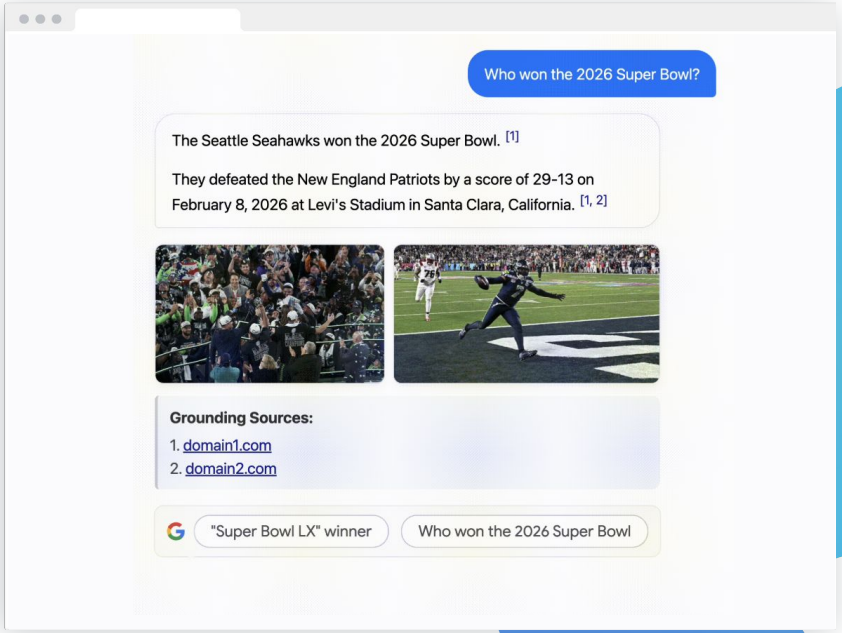
プレビュー

グラウンディング（画像付き Google検索）

Agent が画像付き Google 検索を
使用可能になり、よりグラフィカルな
表現ができるようになりました

この機能の利点：

- Google ウェブ検索は、
  ウェブページ リンクやスニペットに加えて、
  関連画像を活用できるようになりました
- 検索結果には、テキストと画像、
  そして引用情報が含まれます



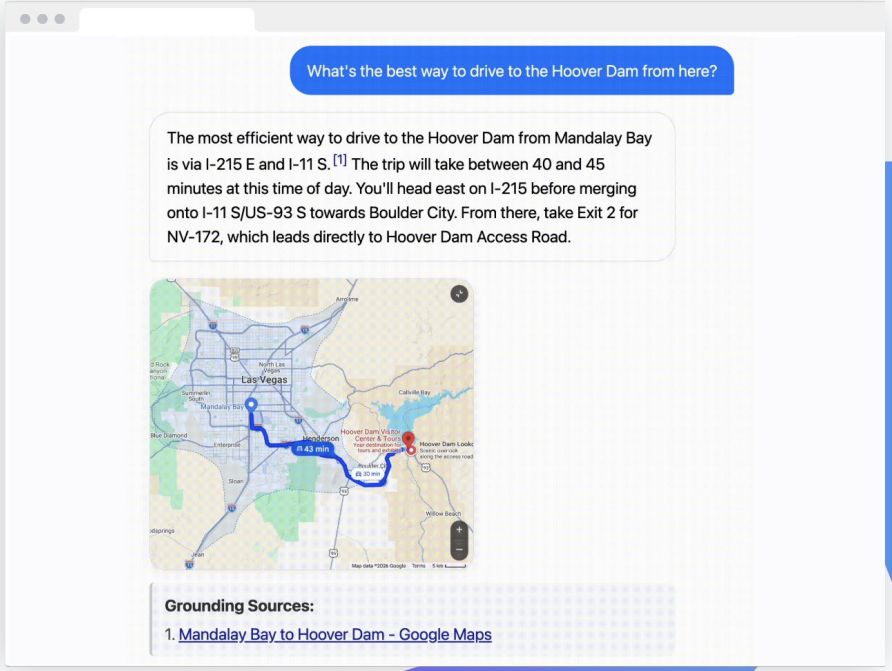
プレビュー

Google マップのルート案内

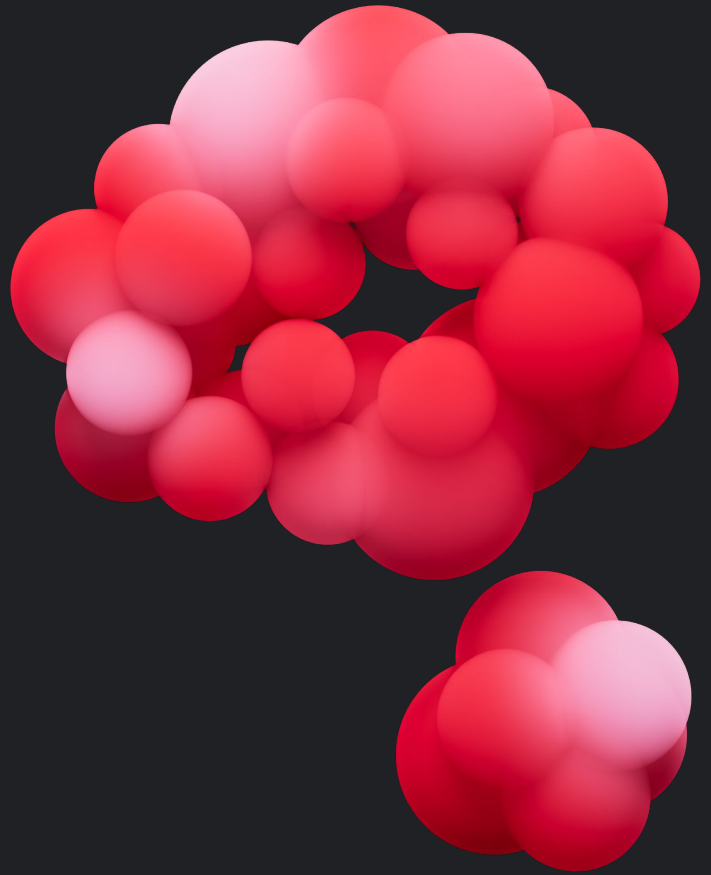
Agent がユーザーの質問に応じて
ルート案内ができるようになりました

この機能の利点：

- Google マップのルート案内機能で、
ユーザーからの質問に対応
- 興味のある場所を検索した後、
ユーザーはそこへの行き方を質問し、
様々な交通手段に対応した最新の
ルート案内を受け取ることができます
- 開発者は、視覚的なルート案内を
アプリケーションに統合できます



05.
顧客接点 AI に関するアップデート



‘Gemini Enterprise for Customer Experience’ Product Suite

Managed Solution

Omnichannel Gateway (オムニチャネル ゲートウェイ) [Preview in May]
Web | Mobile | WhatsApp | Instagram | Voice

OOTB Agents（業界特化型エージェント テンプレート) [Preview]
Shopping | Food Ordering | Auto

Custom CX Agents（カスタム エージェント）

AI Commerce Search
(旧 VAIS:C)

CX Agent Studio
Native Audio | AI Augmentation | Evaluations | 100+ connectors | Visual Builder

Agent Assist
Live Translation | Summarization | Companion Agent | Supervisor Assist

CX Insights and Quality AI
Performance Monitoring | Custom Dashboards | Discovery Engine

Google’s AI Foundation

- Gemini Model Family
- Agent Development Kit (ADK)
- Google DeepMind
- Security and Compliance
- RAG & Search
- 100+ Connectors

5 月にプレビュー版

Omnichannel Gateway

会話をあらゆるチャネルで

消費者は「チャネル」単位で物事を考えず、「関係性」で考えます。私たちのゲートウェイは、会話の文脈（コンテキスト）をタイムラグなしで完璧に引き継ぐことを保証します

- **Multimodal Streaming:** 双方向のオーディオおよびビデオをサポート
- **Seamless Transitions:** WhatsApp から音声通話、そしてモバイルアプリへと、途切れることなく移行可能
- **Persistent Context:** コンテキストを保持し続けることで、「もう一度言っていただけますか？」というやり取りを排除
- **Zero Friction:** あらゆるチャネル間でコンテキストを共有

As featured at NEXT keynote

プレビュー

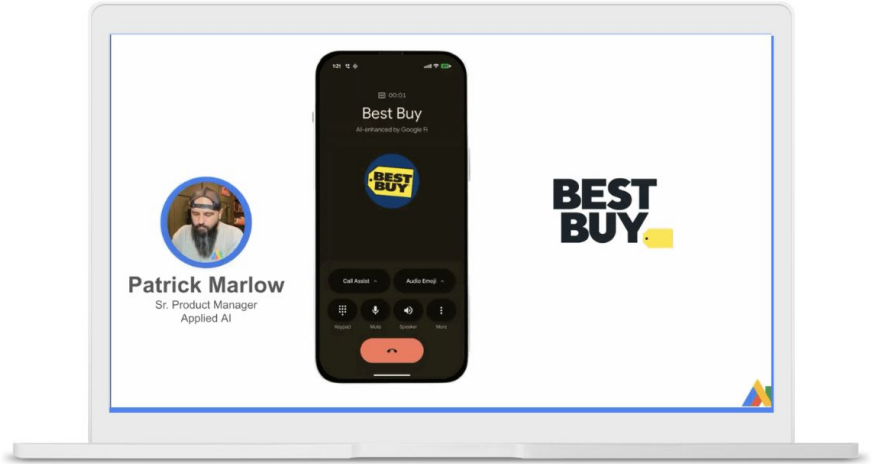
音声モデル Gemini 3.1 Flash Live 搭載

Gemini 3.1 Flash Live オーディオ モデルを搭載し、音声を直接処理することで超低遅延を実現

即時の診断：技術的な問題をリアルタイムで検出し、解決

感情的インテリジェンス ：顧客の不満や焦りを真摯に認識し、対応

人間レベルの精度：自然な割り込み（バージイン）や、リアルタイムでのトーン調整が可能

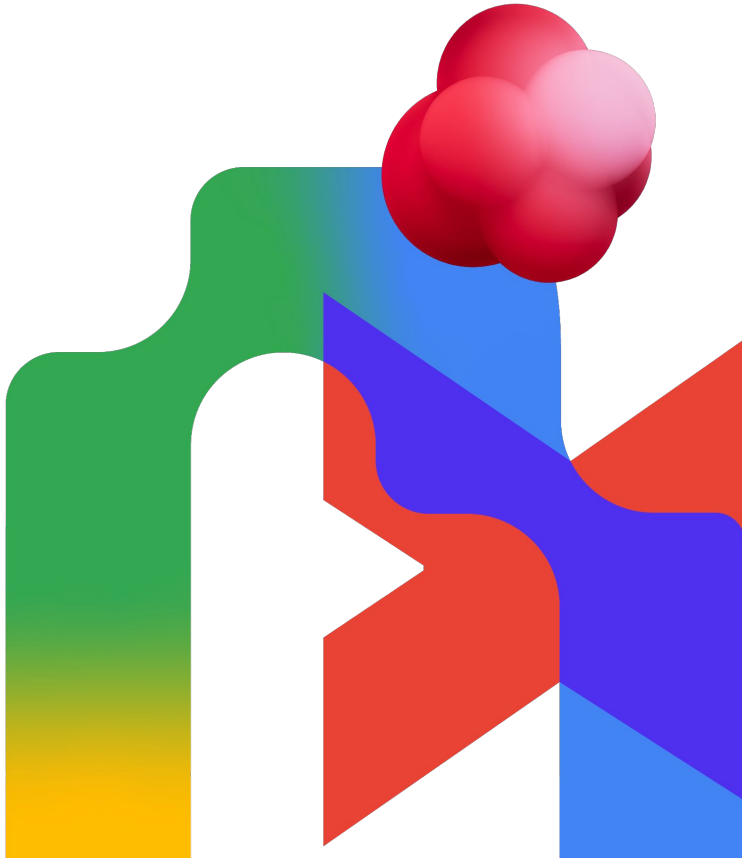


関数呼び出し精度　90.8% | 音声推論精度　97%

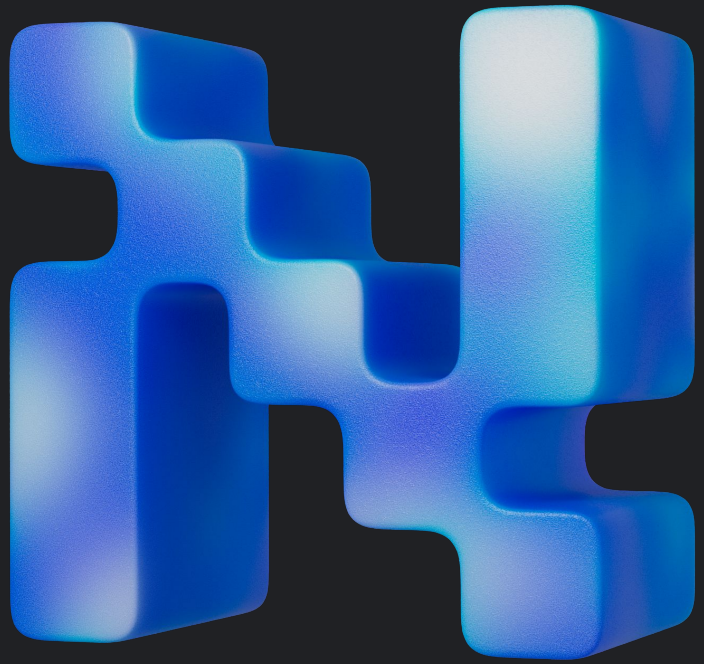
As featured at NEXT keynote

Google Cloud Next 26

AI Hyper Computer

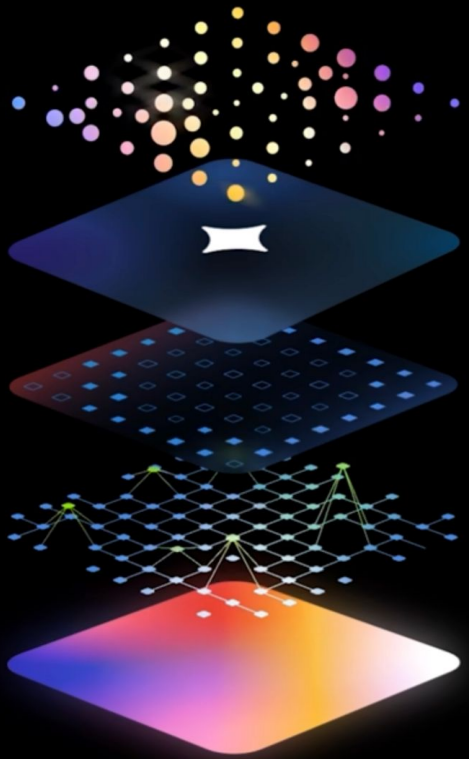


01.
AI Infrastructure



Only Google

最もセキュアなクラウド上に構築された、完全なデータ & AI スタックを提供します



Agentic Taskforce

Agentic Platform and Models

Agentic Defense

Agentic Data Cloud

AI Hypercomputer

NVIDIA on Google Cloud

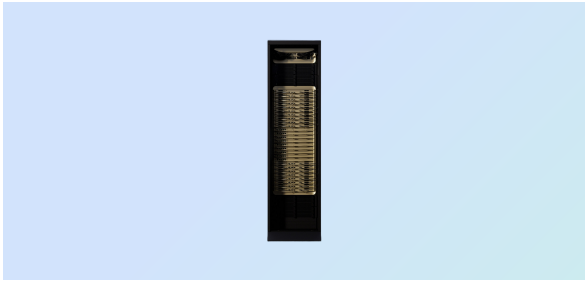
大規模な学習と推論向けに設計された AI Accelerator

A4X

- FExascale のパフォーマンス、ラックスケール
- 72x Blackwell with 5th gen NVLink
- 28.8 Tbps RoCE によるスケールアウト

A4X Max

- テスト時スケーリングと AI リーズニング
- 72x Blackwell Ultra with 5th gen NVLink
- 57.6 Tbps RoCE によるスケールアウト

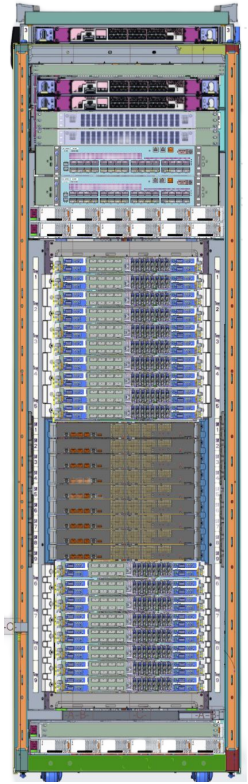


A5X

- Agentic AI のための AI Accelerator
- 72x Rubin with 6th gen NVLink
- 115.2 Tbps RoCE によるスケールアウト

Google Cloud Titanium, Cluster Director そして 1 GW 以上の先進的な液冷インフラ

A5X インスタンス
NVIDIA VR200 (Vera Rubin)



Compute とラックスケールの設計

- 高密度な ORv3 準拠のラック + トレイ設計
- 72x GPU NVLink ドメイン | 21TB GPU メモリ @ 3,600 Gbps
- GB200 と比較して 3.5 倍 の FLOPs と 3 倍 のメモリ帯域幅
- Google の第 4 世代集中冷却システム + 液冷

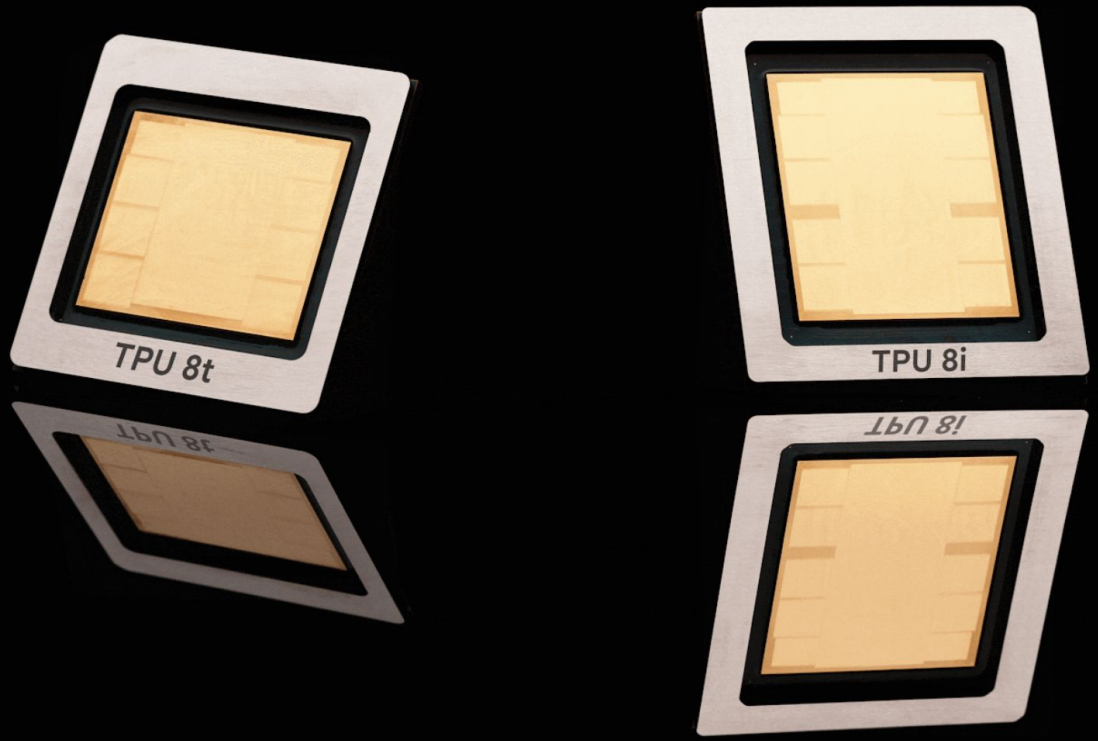
ネットワーク

- Virgo ファブリック + 数千台の GPU を備えたノンブロッキング RDMA クラスター
- RoCE v2, 4x 1600Gbps CX9 NIC

インテグレーション

- Cluster Director、Vertex AI、GKE、Cloud Run との統合
- パフォーマンス向上のためのベアメタルGCE インスタンス

New



10 年以上 に渡る Google TPU イノベーション

TPU v1
2015

社内用の
推論チップ

TPU v2
2018

256 chips
分散共有メモリ

TPU v3
2020

液体冷却

TPU v4
2022

光学的に
構成変更できる
3D トーラス
アーキテクチャ

TPU v5e
2023

中規模から大規模な
学習と推論に向けて
専用設計

TPU v5p
2023

パワフル、
スケーラブル、
フレキシブルな
AI アクセラレータ

Trillium
2024

極めて高い
パフォーマンスと
効率性

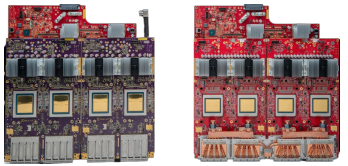
Ironwood
2025

5 倍のピーク性能
6 倍 大きい HBM
2 倍の電力効率

*Trillium 比

Available Today!

TPU 8t & TPU 8i
2026



Just Announced!!

Training: TPU 8t

121 ExaFlops / pod

FP4 compute

1 Million+

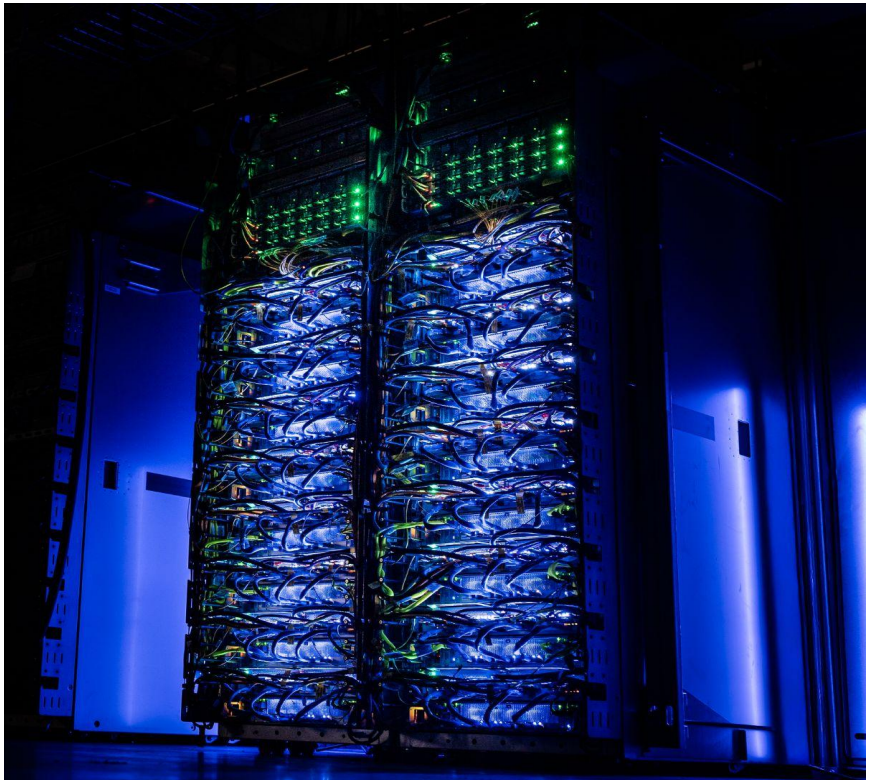
の TPU チップを一つの学習クラスタに

4 倍

の DCN 帯域幅 (TPU 7x 比)

2 倍

の電力効率



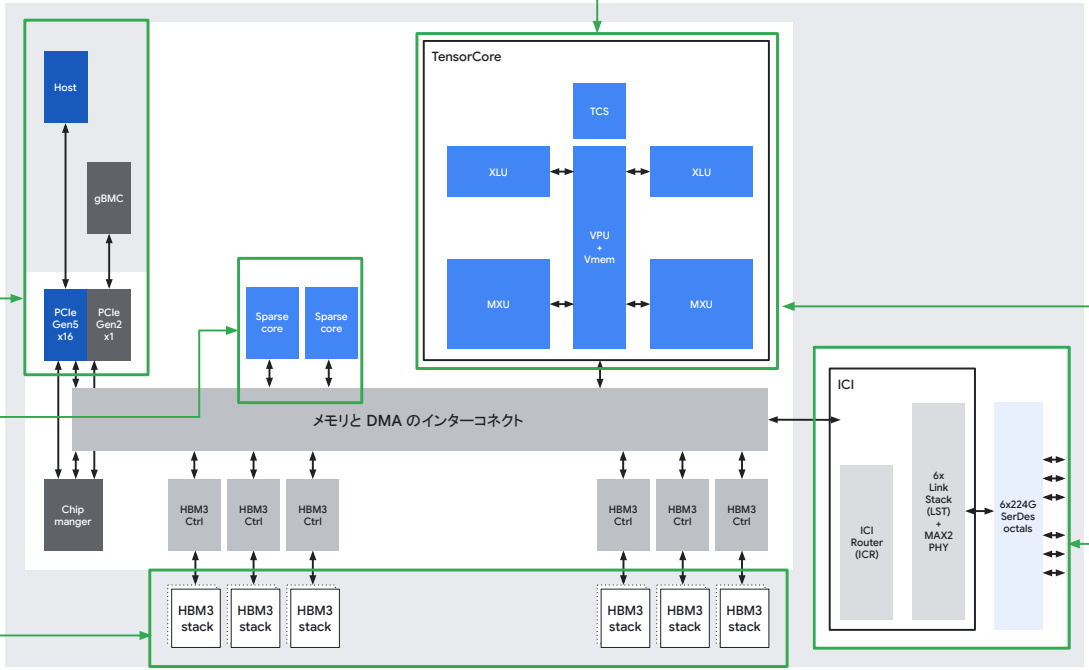
TPU 8t ASIC

3x PFLOPs FP4*
混合精度（BF16, FP8）を
サポート
高速な行列演算
（乗算、変換、転置）に最適化

高速な chip-to-computer
PCIe 接続

Sparsecore ベースの LLM
Decoder Engine

チップあたり 216GB の高速なイ
ンパッケージ HBM



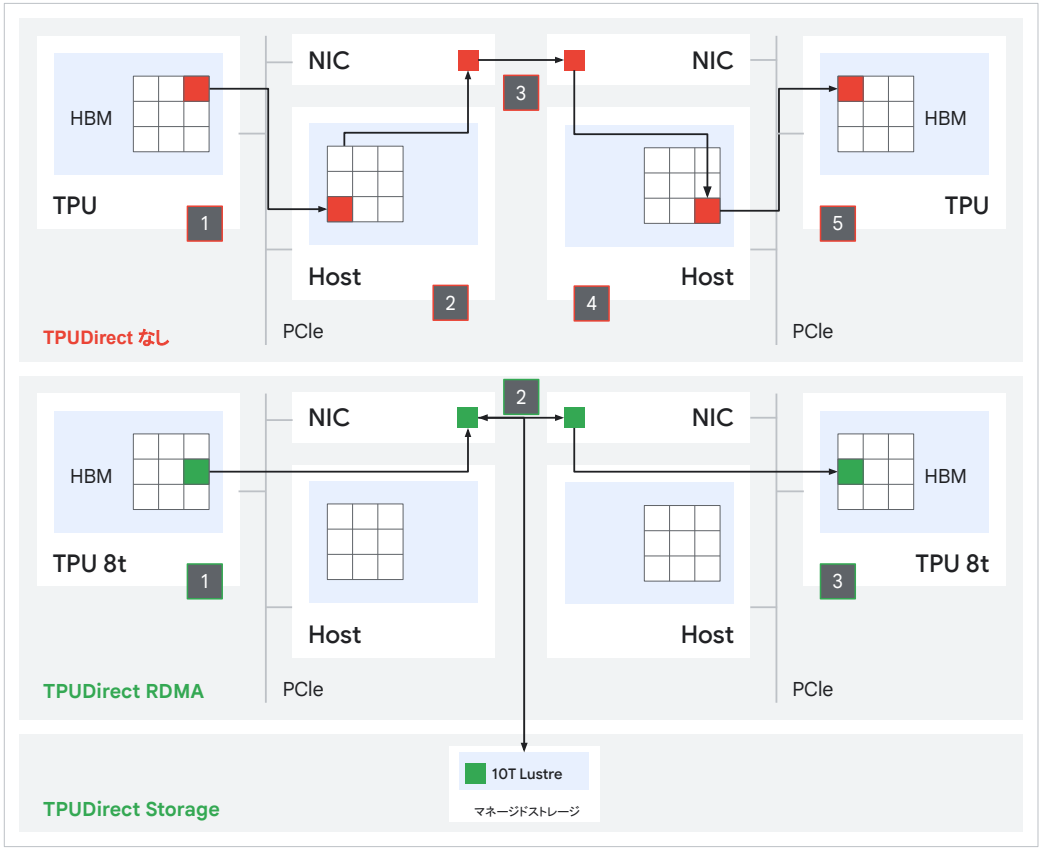
高効率な行列演算を実現する
シストリック・アレイ

2x 高速化 * したチップ間
接続帯域幅

*vs 7th gen Ironwood TPU

TPU RDMA と TPUDirect Storage

1. TPU HBM と NICs の間の直接データ転送
2. ホストメモリ経由のホップを排除し、データ転送を削減
3. DCN 全体のコレクティブ通信を高速化
4. ペタバイト級のデータを HBM に直接、10 倍 高速に読み込み

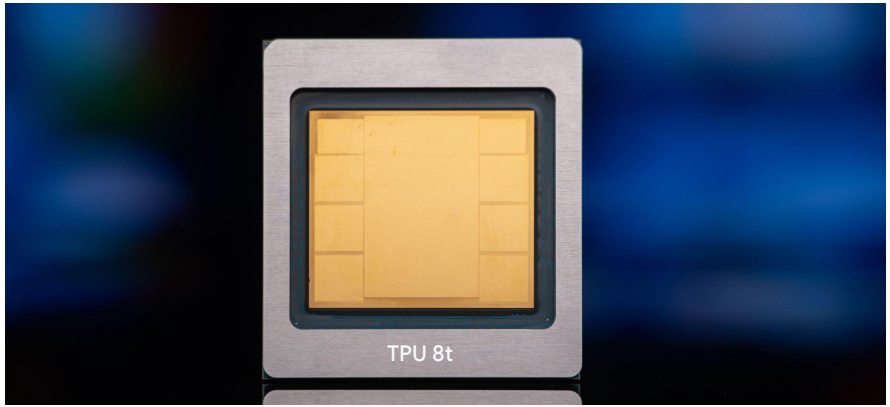


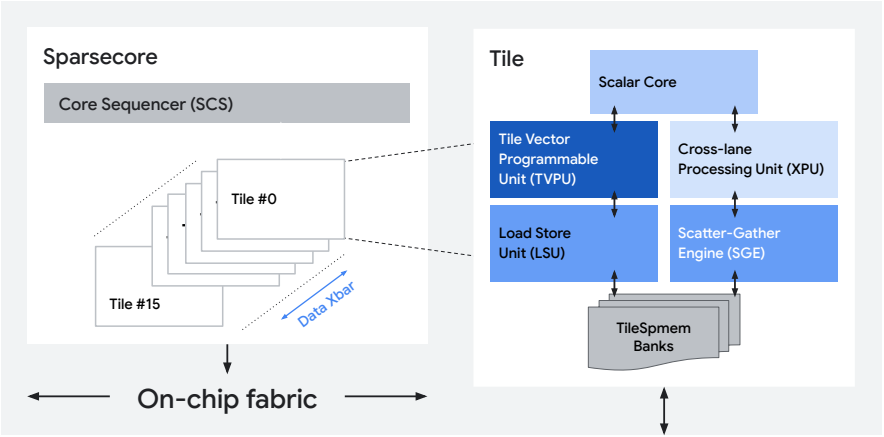
Specialized SparseCore と LDE

SparseCore は、TPU 8t の LLM Decoder Engine (LDE) とともに Google の TPU に統合された専用のコプロセッサです

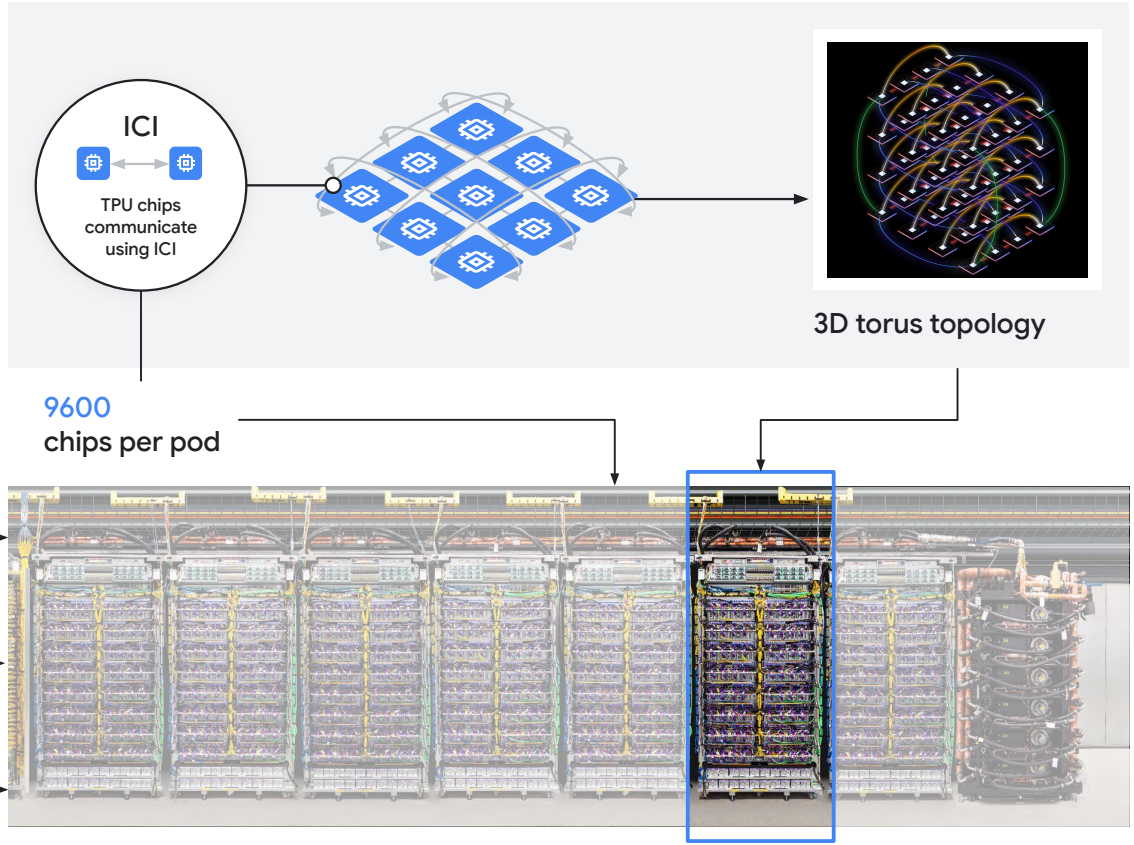
- embeddings 重視のワークロードを加速
- MoE のコレクティブ通信をオフロード
- LLM decoder engine により prefill と decode をオーバーラップ

DLRM DCN v2 などのモデルを 5x 加速
LDE の追加により Sparsecore の演算強度が最大 30x 向上





Scale-up:
3D torus ICI domain



121 exaFLOPs
FP4 Compute

2PB of
shared HBM

19.6 Tbps bidirectional
ICI bandwidth

Scale-out: TPU 8t Virgo Network

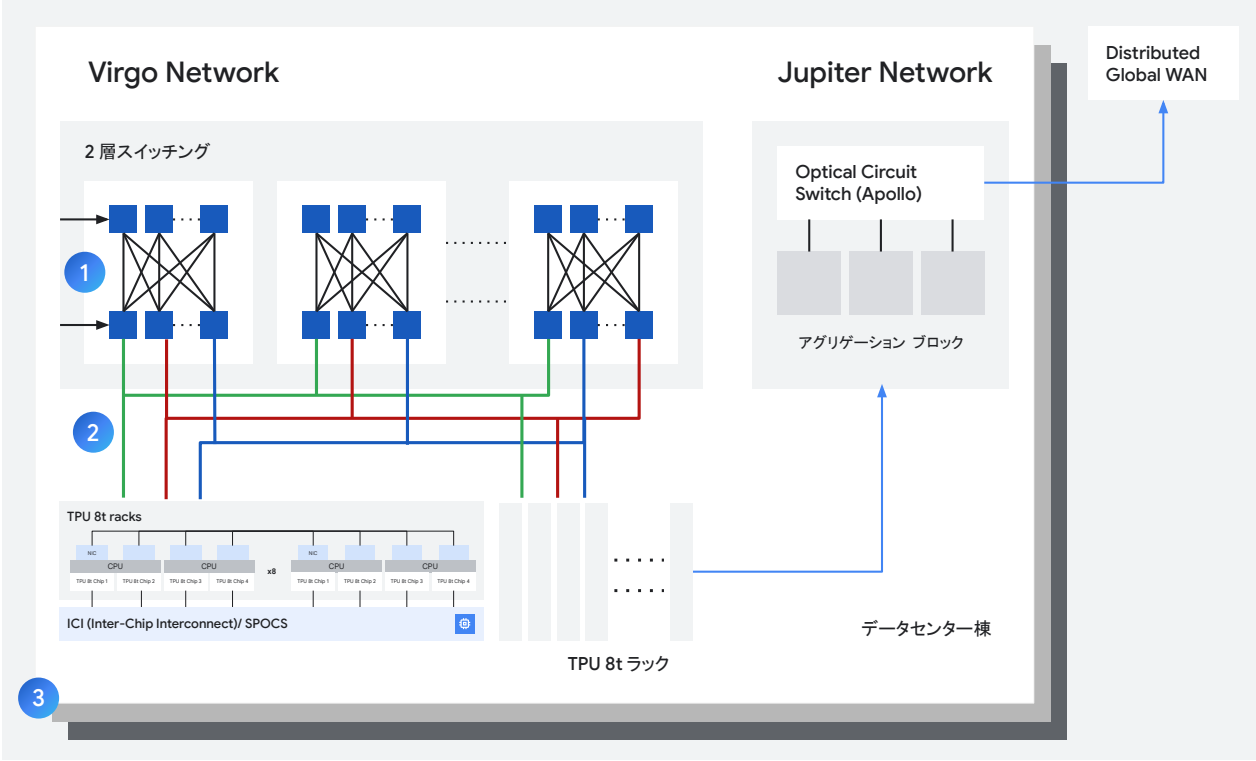
アーキテクチャの転換

1. 2 層スイッチング
   完全 non-blocking
2. 独立したプレーンを持つ
   レジリエントなファブリック
3. マルチデータ センターサイト
   への拡張性

単一ファブリックは建物規模

最大 134,000 TPU 8t Chips

レイテンシを 40% 低減



Inference: TPU 8i

最大 7 ホップ

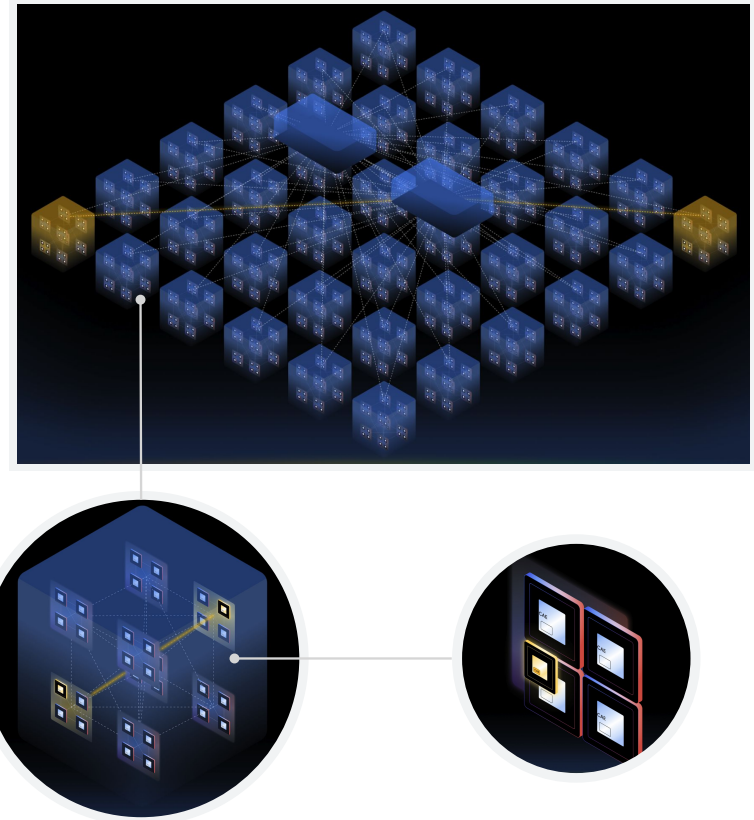
新しい ICI トポロジにより
ネットワーク距離を 58% 低減

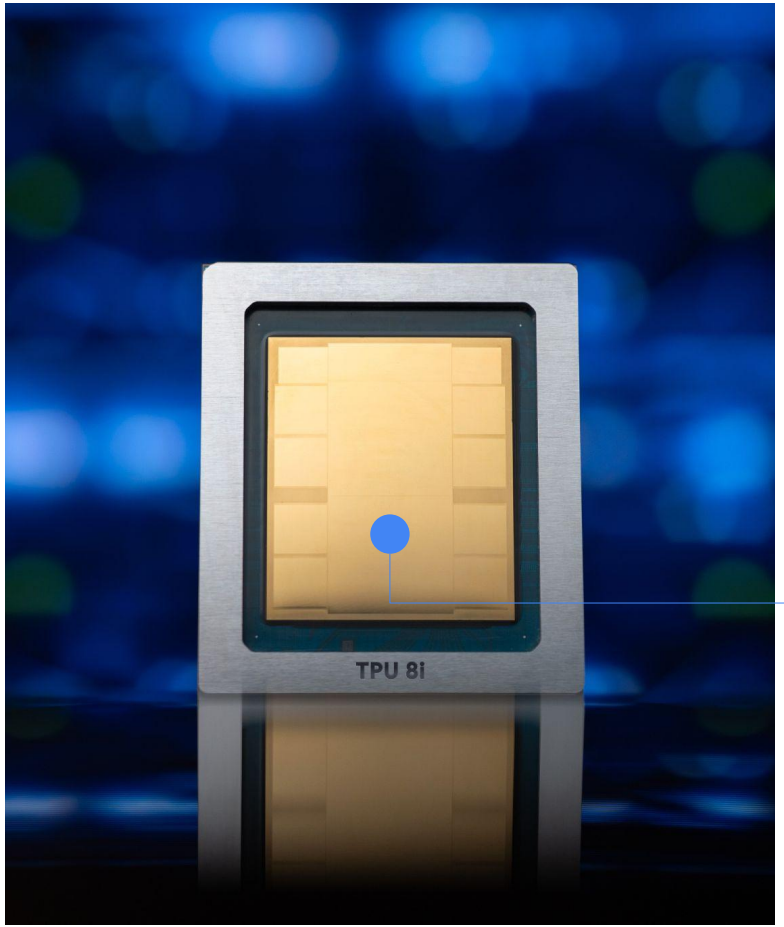
384 MiB SRAM

更に大きい KV Cache ホスティングにより
低レイテンシな推論を可能に

2x 低レイテンシ

3D Torus topology と比較した場合





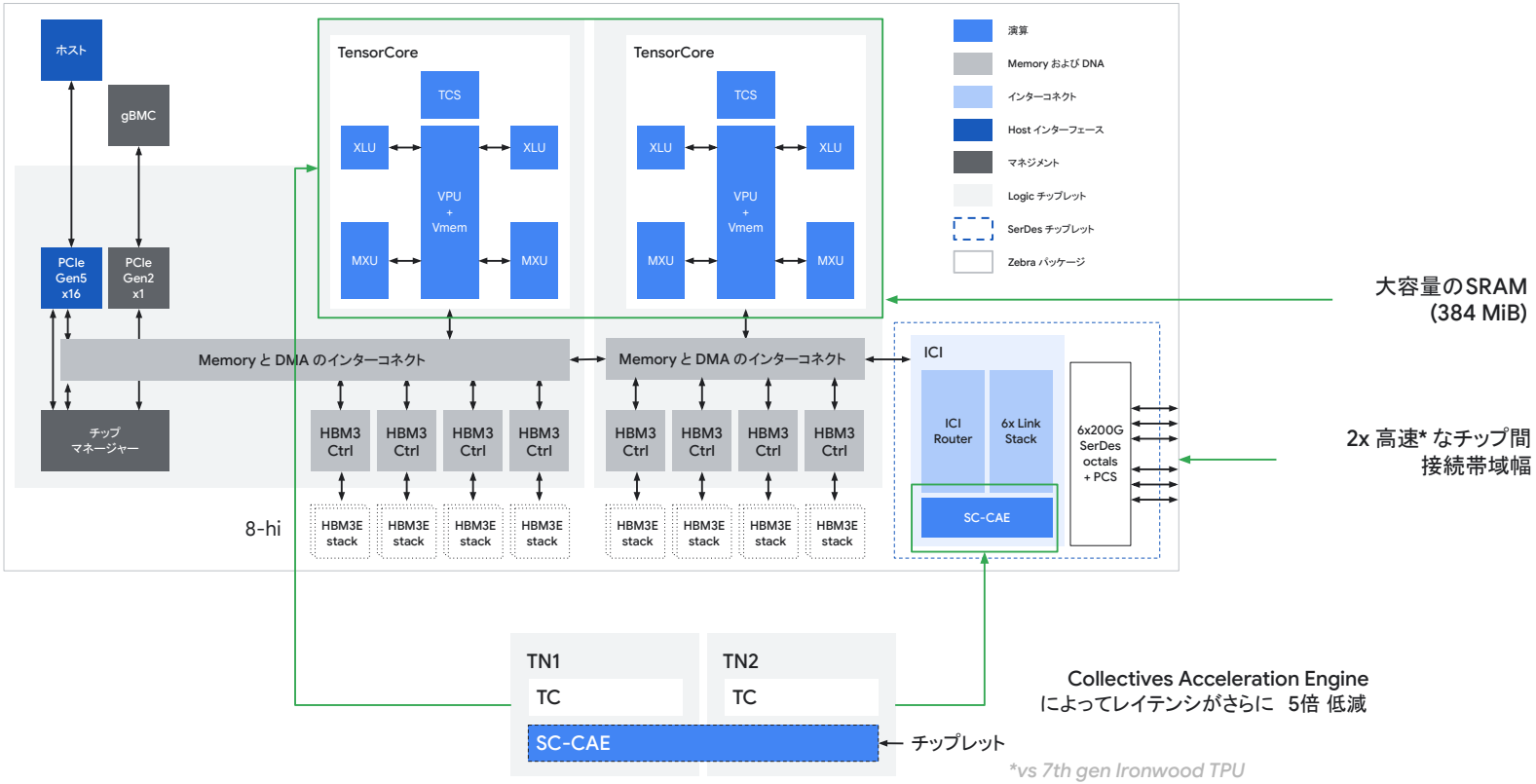
レイテンシの壁を突破

NEW

SparseCore を Collectives Acceleration Engine (SC-CAE) として統合

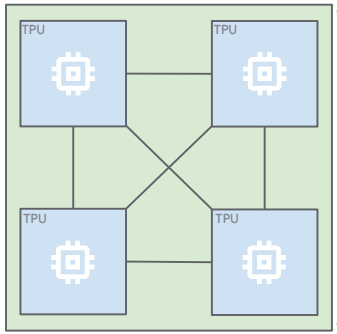
統合 on-chip SRAM を大幅に強化 (384 MiB)

TPU 8i ASIC

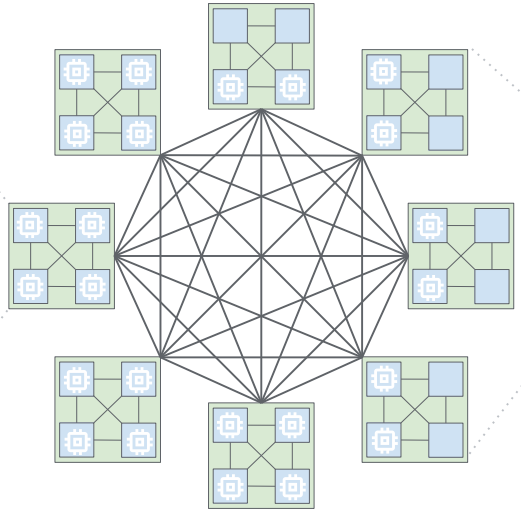

ホスト
gBMC
PCIe Gen5 x16
PCIe Gen2 x1
チップ マネージャー
TensorCore
TCS
XLU
VPU + Vmem
MXU
TensorCore
TCS
XLU
VPU + Vmem
MXU
演算
Memory および DNA
インターコネクト
Host インターフェース
マネジメント
Logic チップレット
SerDes チップレット
Zebra パッケージ
大容量のSRAM (384 MiB)
Memory と DMA のインターコネクト
Memory と DMA のインターコネクト
ICI
ICI Router
6x Link Stack
6x200G SerDes octals + PCS
SC-CAE
2x 高速* なチップ間 接続帯域幅
HBM3 Ctrl
HBM3E stack
8-hi
TN1
TC
TN2
TC
SC-CAE
チップレット
Collectives Acceleration Engine によってレイテンシがさらに 5倍 低減
*vs 7th gen Ironwood TPU

Boardfly トポロジ

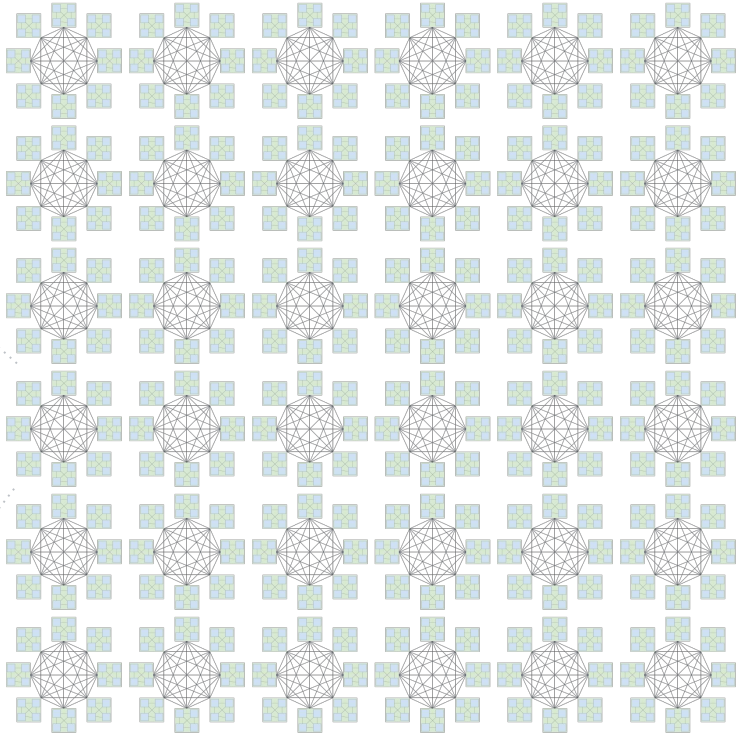
非トーラス型 Boardfly トポロジ
最大 1152 台の TPU を相互接続



ボード（マシン）
完全接続された4 台の TPU



グループ（ラック）
完全接続された8 枚のボード

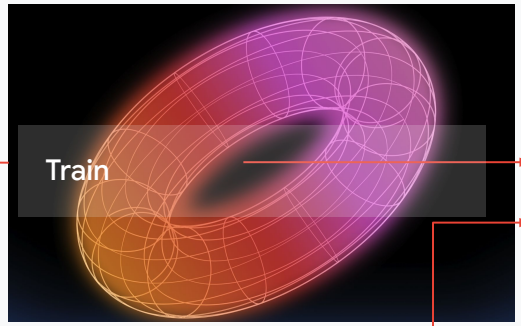


ポッド
完全接続された36 グループ
（計 1152 チップ）

TPU 8t

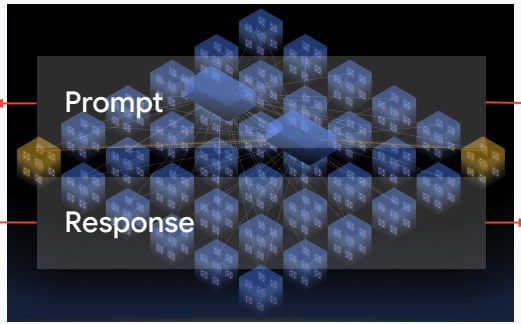
TPU 8i

Pre-training



Model

Inference



Agent

Reinforcement learning

PyTorch on TPU

ネイティブな PyTorch 体験

最小限のコード変更で動作

追加の API は必要なし

Full eager mode

torch.compile and torch.distributed

でのお馴染みのスケーリング

高い 費用対効果

TPU における効率と費用対効果を加速

学習、事後学習、推論など AI ライフサイクルの全てに最適化

X00B+ params までリニアに スケールするように設計

PyTorch コミュニティの一員に

PyTorch をバックエンドに

torchtitan, torchtune そして torchao を含め幅広いエコシステム

vLLM や PyTorch Lightning のようなよく使われるソリューションも 利用可能に

PyTorch with CUDA backend

```python
import torch
import torch.distributed as dist
from torch.distributed.fsdp import FullyShardedDataParallel as FSDP
import torch.optim as optim
from torch.utils.data import DataLoader, Dataset, DistributedSampler
from transformers import AutoModelForCausalLM

def setup_dist():
 dist.init_process_group(backend="nccl")

def simple_example():  # pylint: disable=missing-function-docstring
 # Distributed device setup
 setup_dist()
 rank = dist.get_rank()
 device = torch.device("cuda", rank)
 torch.cuda.set_device(device)

 model = AutoModelForCausalLM.from_pretrained("meta-llama/Meta-Llama-3-8B")
 model = FSDP(model, device_id=rank, auto_wrap_policy=...)
 optimizer = optim.AdamW(model.parameters(), lr=1e-4)

 # Dataset and DataLoader
 dataset = ...
 sampler = DistributedSampler(dataset, ...)
 dataloader = DataLoader(dataset, ...)

 # Training loop
 num_epochs = ...
 model.train()
 for epoch in range(num_epochs):
   sampler.set_epoch(epoch)
   for batch in dataloader:  # pylint: disable=unused-variable
     # Move batch to device
     input_ids, labels = ...  # pylint: disable=unpacking-non-sequence
     optimizer.zero_grad()
     outputs = model(input_ids=input_ids, labels=labels)
     loss = outputs.loss
     loss.backward()
     optimizer.step()
```

PyTorch with TPU backend

```python
import torch
import torch.distributed as dist
from torch.distributed.fsdp import FullyShardedDataParallel as FSDP
import torch.optim as optim
from torch.utils.data import DataLoader, Dataset, DistributedSampler
from transformers import AutoModelForCausalLM

def setup_dist():
 dist.init_process_group(backend="tpu")

def simple_example():  # pylint: disable=missing-function-docstring
 # Distributed device setup
 setup_dist()
 rank = dist.get_rank()
 device = torch.device("tpu", rank)
 torch.tpu.set_device(device)

 model = AutoModelForCausalLM.from_pretrained("meta-llama/Meta-Llama-3-8B")
 model = FSDP(model, device_id=rank, auto_wrap_policy=...)
 optimizer = optim.AdamW(model.parameters(), lr=1e-4)

 # Dataset and DataLoader
 dataset = ...
 sampler = DistributedSampler(dataset, ...)
 dataloader = DataLoader(dataset, ...)

 # Training loop
 num_epochs = ...
 model.train()
 for epoch in range(num_epochs):
   sampler.set_epoch(epoch)
   for batch in dataloader:  # pylint: disable=unused-variable
     # Move batch to device
     input_ids, labels = ...  # pylint: disable=unpacking-non-sequence
     optimizer.zero_grad()
     outputs = model(input_ids=input_ids, labels=labels)
     loss = outputs.loss
     loss.backward()
     optimizer.step()
```

Goal: Switching from GPU to TPU requires changing only 3 lines related to device initialization.

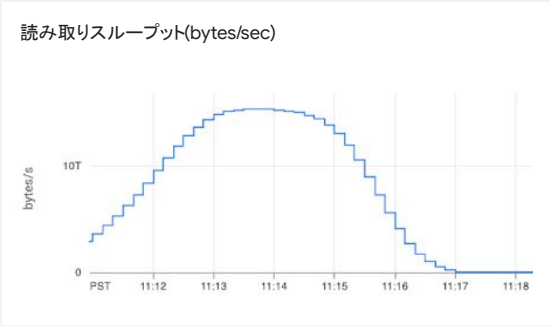
Rapid Bucket NEW

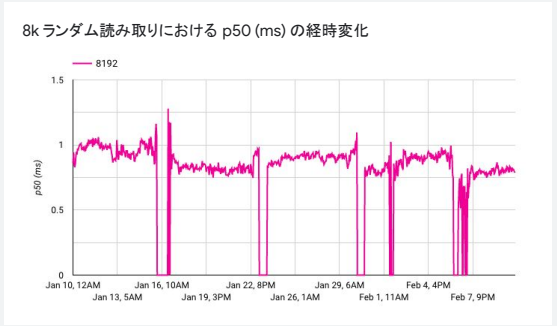
AI およびアナリティクス ワークロード向けの高性能な Zonal Bucket

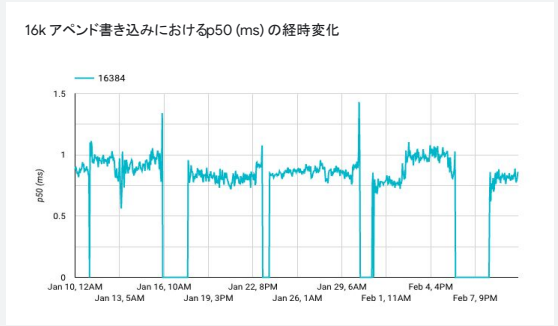
コロケーション ゾーン バケット
最大 15+ TiB/s のスループット、20M QPS、1 ms 未満のレイテンシを提供

YouTube や Gemini を支える Google 内部の非常にスケーラブルな分散ファイル システム、Colossus を活用

主なワークロード: AI/ML マルチモーダル トレーニング / チェック ポインティング、バッチ / ストリーミング アナリティクス、ロギング、データベース アーキテクチャ



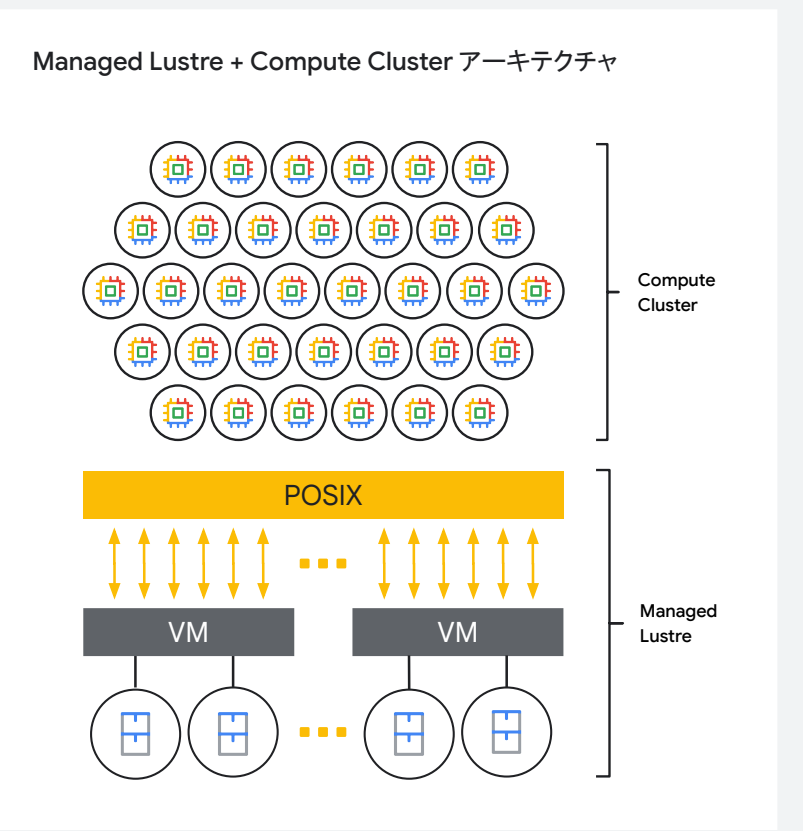




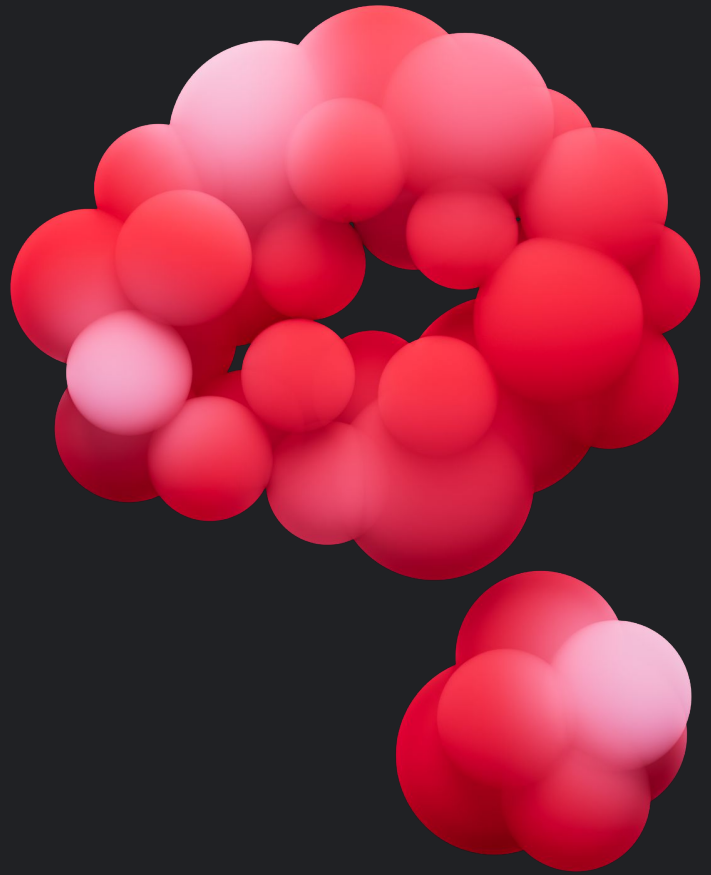
Google Cloud Managed Lustre

DDN EXAScaler ベースの フルマネージド並列ファイルシステム

- DDN EXAScaler を採用
- サブミリ秒のレイテンシ
- インスタンスあたり最大 **10 TB/s** の読み取りスループット
- マルチ NIC クライアントのサポート
- 拡張性はインスタンスあたり **9 TB → 80 PB**
- 使いやすい **Storage Transfer** API
- マネージド **GKE CSI ドライバ**
- **Vertex AI トレーニング クラスター** で広く利用
- POSIX 準拠
- **99.9% SLA**
- CMEK、VPC-SC、メトリクスなど



02. Application Platform (Cloud Run & GKE)



“

2025 年は Cloud Run を利用する開発者とアプリケーション数が**倍増**しました”

Spend caps（支出上限）

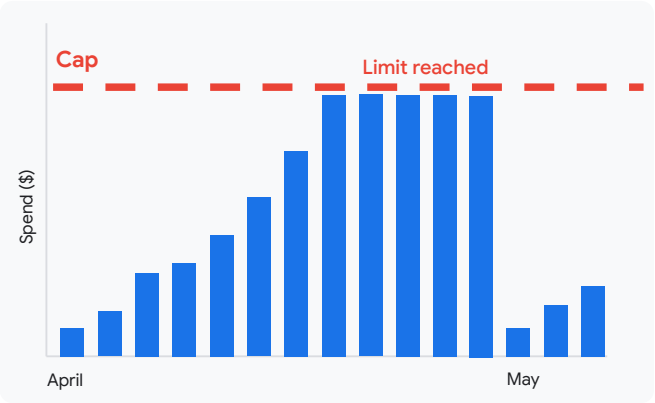
Vibe Coded Apps

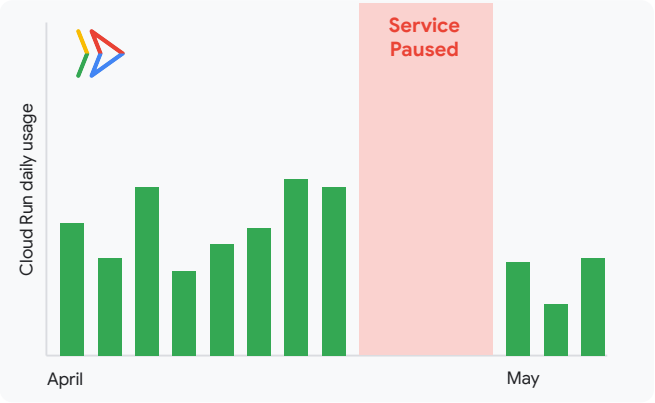
Coming Soon

予算の上限を予測可能にすること は
非常に重要なこと

Billing caps（請求の上限設定） を発表

毎月の最大支出額を設定し、上限に達すると、
Cloud Run のリソースは
自動的に一時停止 される



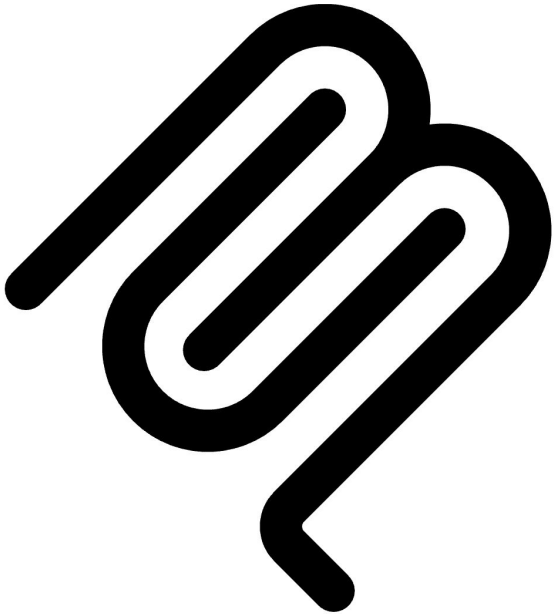


Fully Managed MCP Server

公式の Cloud Run MCP（Model Context Protocol）サーバーにより、お客様とエージェントは Cloud Run アプリのデプロイと管理をさらに容易に行うことが可能

Tools:

- コンテナ イメージのデプロイ
- ソースコード アーカイブからのデプロイ
- ファイル コンテンツからのデプロイ
- サービス一覧、単体の取得



AI Agents

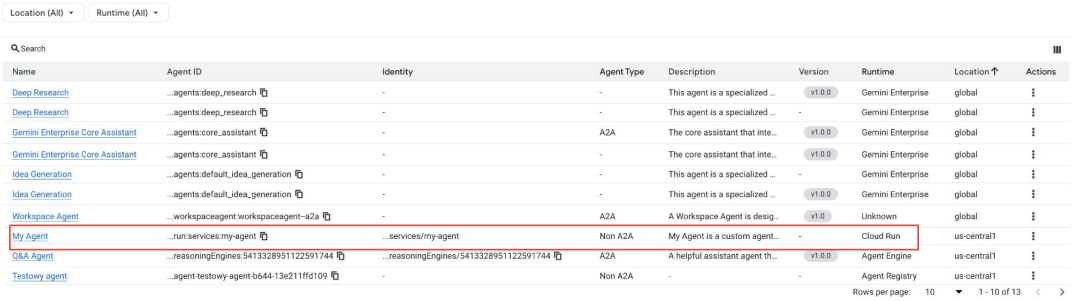
プライベート プレビュー

Gemini Enterprise Agent Platform との統合

エージェントを
Cloud Run にデプロイすると:

- Agent registry にすぐ登録される
- 専用の Agent identity を取得
- これらを中央集権的な強化と
  ガバナンスによって保護

```
$ gcloud run services update my-agent \
                    --functional-type AGENT \
                    --identity-type AGENT_IDENTITY
```



Location (All) Runtime (All)

Search

Name	Agent ID	Identity	Agent Type	Description	Version	Runtime	Location	Actions
Deep Research	...agents:deep_research	-	-	This agent is a specialized ...	v1.0.0	Gemini Enterprise	global	
Deep Research	...agents:deep_research	-	-	This agent is a specialized ...	-	Gemini Enterprise	global	
Gemini Enterprise Core Assistant	...agents:core_assistant	-	A2A	The core assistant that inte...	v1.0.0	Gemini Enterprise	global	
Gemini Enterprise Core Assistant	...agents:core_assistant	-	-	The core assistant that inte...	v1.0.0	Gemini Enterprise	global	
Idea Generation	...agents:default_idea_generation	-	-	This agent is a specialized ...	-	Gemini Enterprise	global	
Idea Generation	...agents:default_idea_generation	-	-	This agent is a specialized ...	v1.0.0	Gemini Enterprise	global	
Workspace Agent	...workspaceagent:workspaceagent-a2a	-	A2A	A Workspace Agent is desig...	v1.0	Unknown	global	
My Agent	...run:services:my-agent	...services/my-agent	Non A2A	My Agent is a custom agent...	-	Cloud Run	us-central1	
Q&A Agent	...reasoningEngines:5413328951122591744	...reasoningEngines/5413328951122591744	A2A	A helpful assistant agent th...	v1.0.0	Agent Engine	us-central1	
Testowy agent	...agent-testowy-agent-b644-13e211ffd109	-	Non A2A	-	-	Agent Registry	us-central1	

Rows per page: 10 1 - 10 of 13

```
$ gcloud run services update my-mcp \
                    --functional-type MCP_SERVER
```

Name	MCP Server ID	Description	Runtime	Location	Tools	Actions
bigquery.googleapis.com	...locations:global:bigquery	BigQuery MCP server prov...	Unknown	global	6	
My MCP	...run:services:my-mcp	My MCP Service with...	Cloud Run	us-central1	1	

Cloud Run サンドボックス

AI Agents
Coming Soon

安全な On-the-Fly（臨機応変な）実行：

Cloud Run リソース内から、一時的（ephemeral）で隔離されたサンドボックスを起動

エージェントが生成したコード、スクリプト、または Chromium を安全に実行

Execute Python on the fly

```
app.post('/execute', (req, res) => {
 const escapedCode = req.body.code.replace(/"/g, '\\"');
 exec(`sandbox do -- /usr/bin/python3 -c "${escapedCode}"`, (e, stdout, stderr) => {
   res.send({ stdout, stderr });
 });
});
```

Cloud Run インスタンス

AI Agents
プライベート プレビュー

新たな基本構成：

事前定義されたリソース タイプに縛られず、個々の Cloud Run インスタンスを直接管理でき、わずか数秒で起動可能

エージェントに最適：

隔離された専用環境での実行が求められる、非同期かつ長時間稼働のバックグラウンドプロセスに最適なオンデマンド実行環境を提供

OpenClaw

```
gcloud alpha run instances create \
  --image alpine/openclaw:latest \
  --port 18789 \
  --memory 4Gi \
  --default-url \
  --add-volume
mount-path=/home/node/.openclaw,type=cloud-storage,bucket=$
BUCKET_NAME
```

SDK

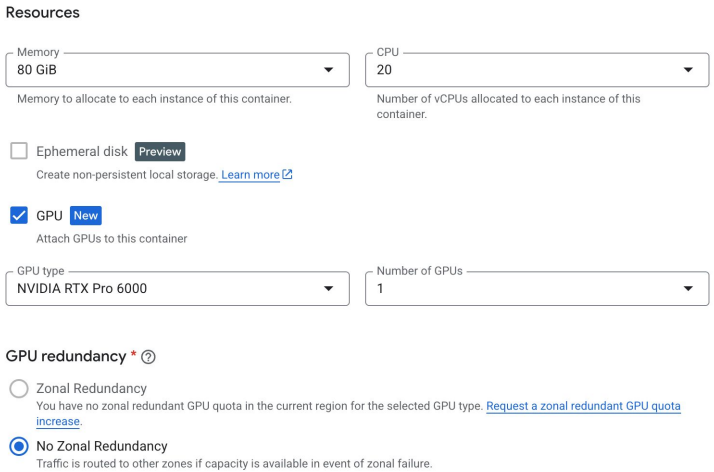
```
import { InstancesClient } from '@google-cloud/run';

new InstancesClient().createInstance({
 parent: 'projects/my-project/locations/europe-west9',
 instance: { containers: [{ image: 'steren/my-agent' }] },
});
```

AI Models

一般提供開始

NVIDIA RTX PRO 6000 Blackwell GPU

- ハイエンドな AI 推論および
  ファイン チューニングに最適
- 700 億（70B）パラメータ超の
  巨大モデルもデプロイ可能
- 圧倒的なコスト パフォーマンス
- 高速起動と「ゼロスケール」による究極の効率化



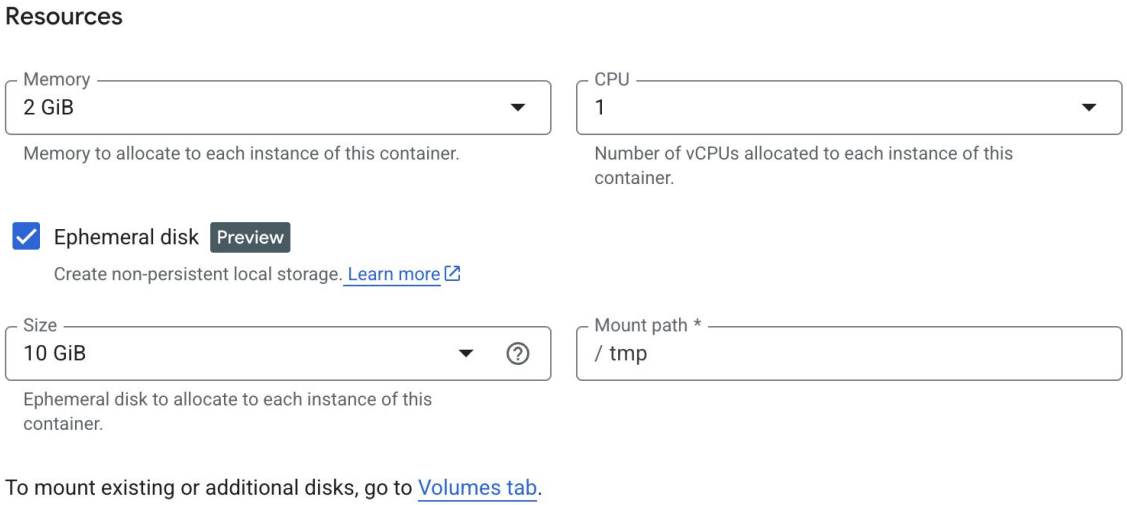

```
$ gcloud run deploy \
  --image zmlai/llmd --port=8000 \
  --args=--model=gs://bucket/gemma-3-27b-it,--max-token-count=1024,--seq-len=128000,--batch-size=2 \
  --network=my-vpc --vpc-egress=all-traffic \
  --gpu-type nvidia-rtx-pro-6000
```

Ephemeral disk

AI Models

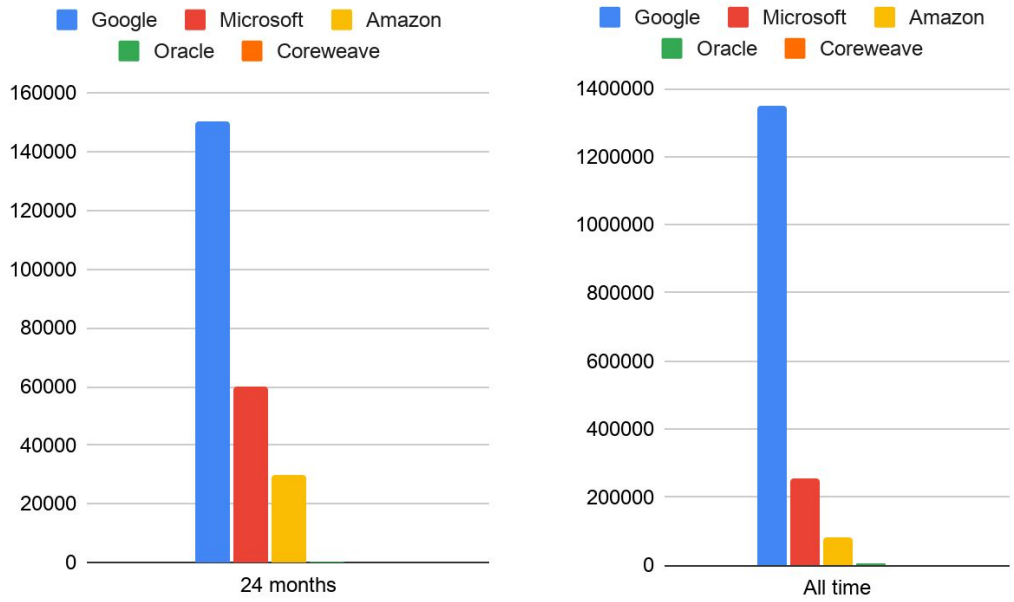
パブリック プレビュー

- インスタンスごとに割り当てられる一時的なディスク ストレージ
- インスタンスの起動時に作成され、停止時に自動削除
- 一時的な作業領域、キャッシュ、または巨大なファイルの処理に最適



Google は Kubernetes イノベーションのリーダーであり続けている

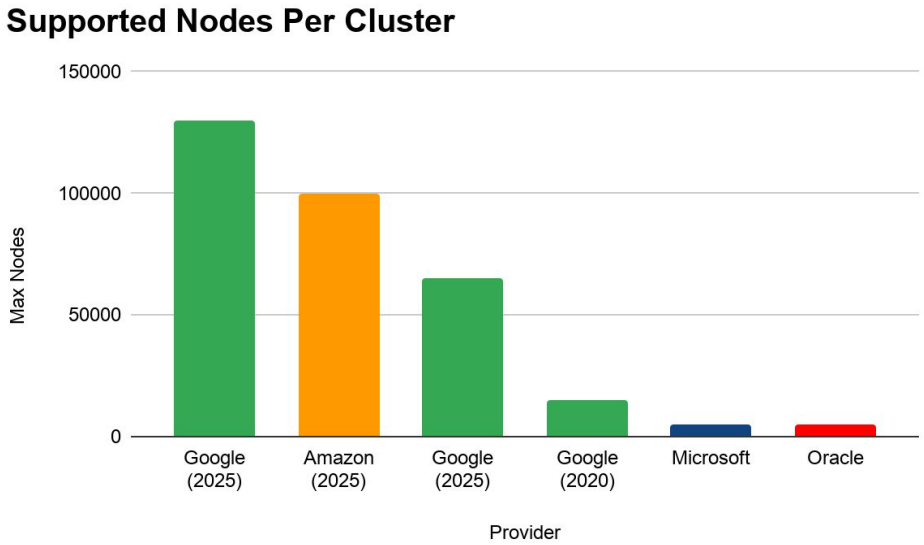
Google は開始当初から
常に最大のコントリビューター



プロバイダーごとの Kubernetes コントリビュート数

Kubernetes の スケールについてもリーダー

- GKEは **130K ノードのサポート** により、スケーラビリティの限界を押し広げ続けています
- AI とステートレス ワークロードへの対応のため、コントロール プレーンのリアーキテクト
- **100** から **1K** 、さらに **20K+ ノード** に容易に拡張できるように



プライベート プレビュー

GKE Hypercluster

- GKE Hypercluster は、
  単一 GKE コントロール プレーンで、
  複数のリージョン にまたがる **256,000 ノード** に分散された
  数百万のアクセラレーターを管理可能
- これは、プライベート AI コンピューティングを提供する、
  Titanium Intelligence Enclave
  (ソフトウェアで強化されたセキュリティ エンジン) に依存しています。
- 「**No Admin Access**」モデルにより、ハードウェアで保証された、
  Pod レベルの分離を提供

Pod スナップショット

- **これまでの課題:** 大規模モデルのデータを GPU メモリにロードするのに時間がかかる
- **解決策: GPU と CPU メモリを含む** Pod の状態の**チェック ポイントを保存**、新しいレプリカは、最初から初期化するのではなく、この**スナップショットから復元**
- あらゆるモデルとモデルサーバーに適用可能
- CRD を使用してスナップショットを作成
- 非常に大規模な機械学習モデルに最適

高速な起動

最大

89%

推論ワークロードの起動が高速に

一般提供開始

GKE Agent Sandbox

高スループット

300

1 クラスター 1 秒間に
サンドボックスを作成

低レイテンシ

<1 sec

起動時レイテンシーが
1 秒以下

コスト効率

最大

30%

高コスト パフォーマンス

Axion を利用した場合、他の
ハイパースケーラーとの比較

プレビュー

Cold Standby Nodes

これまでの課題:

予測不可能な需要に対し、迅速なサンドボックスの起動を提供するには、
高額なノードの過剰プロビジョニングが必要

解決策:

GKE Capacity Buffers API と Cold Standby Nodes を使用し、
予備容量を事前にプロビジョニングすることで、
サンドボックスの起動レイテンシを削減

バッファ用に固定数の Pod、スケーラブルなパーセント バッファ、
またはバッファに利用可能な CPU/ メモリの総量を設定

一般提供開始

GKE Inference Gateway

これまでの課題:

LLM 推論応答のばらつきが大きいため、ラウンドロビンなどの従来の負荷分散では
LLM リクエストのルーティングが最適化されず、GPU/TPU の利用率が低下してしまう

解決策:

GKE Inference Gateway は、LLM モデルサーバーのメトリック
(例:KVCache の使用率、ペンディング キューの長さ)と
プロンプトのプレフィックスに基づいてリクエストをルーティング

レイテンシの短縮、コストの削減、スループットの向上

一般提供開始

Predictive Latency Boost

これまでの課題:

「traffic mix」が変化するため、最適な Inference Gateway リクエスト ルーティング ヒューリスティックを定義するのが困難

解決策:

手動のリクエスト ルーティング ヒューリスティックを、最新のリクエストの実際のレイテンシ データを使用してトレーニングされた ML モデルによるルーティング決定に置き換え

応答レイテンシを短縮しながら、アクセラレータの利用率を最大化

Lower Latency

最大

70%

Predictive Latency Boost を使用しない場合よりも time-to-first-token (TTFT) レイテンシを短縮

一般提供開始

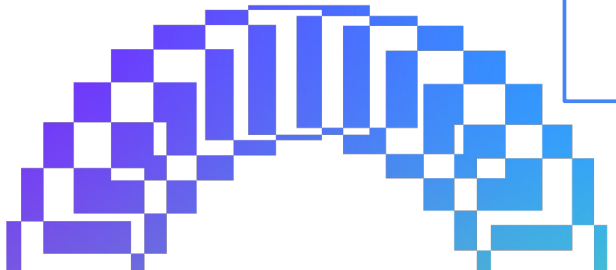
KV Cache Offloading

これまでの課題:

KV キャッシュのサイズは GPU / TPU メモリを
すぐに超える可能性があり、
特に長いコンテキスト ウィンドウを持つ
エージェント型のユースケースで顕著

解決策:

KV キャッシュが GPU / TPU メモリから CPU メモリ、
ローカル ストレージ、そして最終的には
リモート ストレージへと
「オーバーフロー」できるように



Lower Latency

最大

79%

**KV Cache Offloadingを
使用しない場合よりも
time-to-first-token (TTFT)
レイテンシを短縮**

Shared KV Cache

一般提供開始

これまでの課題:

モデルへの多数の同時リクエストによる高負荷は、
利用可能な KV キャッシュ ストレージ スペースを
使い果たすことがある

解決策:

KV キャッシュをすべての vLLM レプリカ間で共有される
ファイル システムにオフロードし、レプリカ間での
KV キャッシュの再利用を可能にするとともに、新しい
レプリカがキャッシュを迅速に「ハイドレート」できるように

Higher Throughput

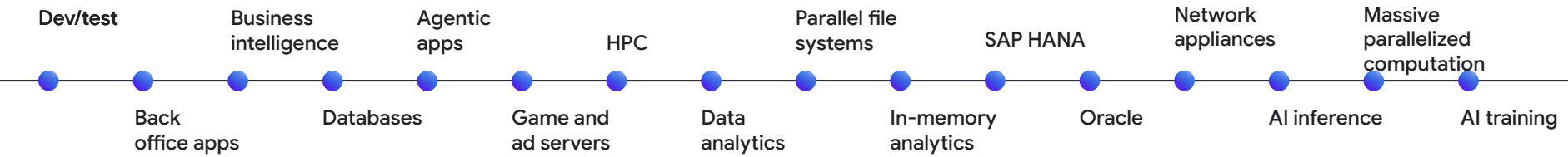
最大

13x

**高負荷時のスループット
（tokens per second）を向上**

03. Infrastructure

ワークロードに最適化 されたコンピューティング



汎用

Balanced price-perf	Performance and enterprise
N-Series	C-Series
E-Series	New shapes!

用途特化

Compute-optimized	Storage-optimized	Memory-optimized	Network- and Hd-optimized (New family!)	Accelerator-optimized	
H-Series	Z-Series	M-Series	C*N-Series	A-Series	TPUs
	New shapes!	X-Series	M*N-Series	G-Series	

Hyperdisk

Balanced | Hyperdisk Throughput | Hyperdisk Extreme | Hyperdisk ML

General purpose
C-series products

New shapes!

C4

最新の Intel Xeon “Granite Rapids”

C3 と比較して最大 36% のパフォーマンス向上：
価格対性能比で業界トップクラス

4.2 GHz CPU周波数で、あらゆる VM の中で
最高レベルを実現

最新世代の第 6 世代 Intel プロセッサー
（Granite Rapids）に対応し、
Intel AMX およびネイティブFP16 をサポート

最大 18 TB の Titanium LSSD に対応

Standard / highmem それぞれの
標準インスタンス、ローカル SSD 搭載マシンタイプで 4
つ のベアメタル インスタンスをリリース

C4D

最新の AMD Epyc “Turin”

C3D と比較して最大 30% 優れた
コスト パフォーマンスを実現

業界トップクラスのコスト パフォーマンス、
ワークロード パフォーマンスの向上

Web：NGINX +80%

データベース：MySQL +55%、Redis +35%

最大 12 TB の Titanium LSSD

初の AMD ベースメタル サーバー、
3種類のインスタンスをリリース

C4A

Google Axion

Google 独自の Arm ベースプロセッサ
「Axion」を搭載

競合製品（例：Graviton4）と比較して、
最大 10% 高いパフォーマンスと
コスト パフォーマンスを実現

同等の x86 ベース VM と比較して、
最大 65% 優れたコスト パフォーマンス

最大 60% 優れたエネルギー効率

最大 6 TB の Titanium ローカル SSD を搭載

General purpose
N-series products

N4

Intel Xeon “Emerald Rapids”

N2 と比較して最大 20% 高い
コストパフォーマンス

2～80 vCPU の標準 VM シェイプ（high-cpu、Standard、High-mem シェイプを含む）

vCPU とメモリを最適化できる
カスタムマシン タイプ

42 のGoogle Cloud リージョンで利用可能で、広範囲なリージョン カバレッジを実現

New!

N4D

Latest AMD Epyc “Turin”

N2D と比較して最大 50% のパフォーマンス向上

ワークロード パフォーマンスの大幅な向上

Nginx：最大+250%

SpecJBB：最大+70%

vCPU とメモリを最適化できる
カスタムマシン タイプ

New！最大 768GB のメモリを搭載した
96 個の vCPU シェイプ

New!

N4A

Google Axion

現行世代の x86 と比較して、最大 2 倍のコスト パフォーマンスと 80% 優れたエネルギー効率を実現

SIR17：最大+105%

Nginx：最大+90%

SpecJBB：最大+85%

vCPU とメモリを最適化できるカスタムマシン タイプ

1～64 個の vCPU、最大 512GB のメモリを搭載した VM 構成

標準規格に準拠した幅広いエコシステム

各最適化マシンは
特定のワークロードに合わせて設計

New family!

C4N | M4N | C4NX

Network- and Hyperdisk- Optimized

高性能ネットワーク
最大 400 Gbps、95M PPS
I/O
25 GB/s、1M IOPS

対象ワークロード:
データベース、データ分析、
ネットワーク、通信

New shapes!

Z3 | Z4D | Z4M

Local Storage Optimized

高密度ローカル SSD と
低遅延 I/O
最大 168 TB ローカル SSD

対象ワークロード:
分散データベース
（NoSQL、ScyllaDB）
データウェアハウス、ログ分析、
AI/ML ワークロード、
分散並列ファイルシステム

H4D

High-Performance
Compute Optimized

最大 192 コア、
200 Gbps RDMA 対応、
Titanium SSD、
クラスタ管理機能を搭載

対象ワークロード:
HCLS、気象、製造、
EDAにおけるHPC

M4 | X4

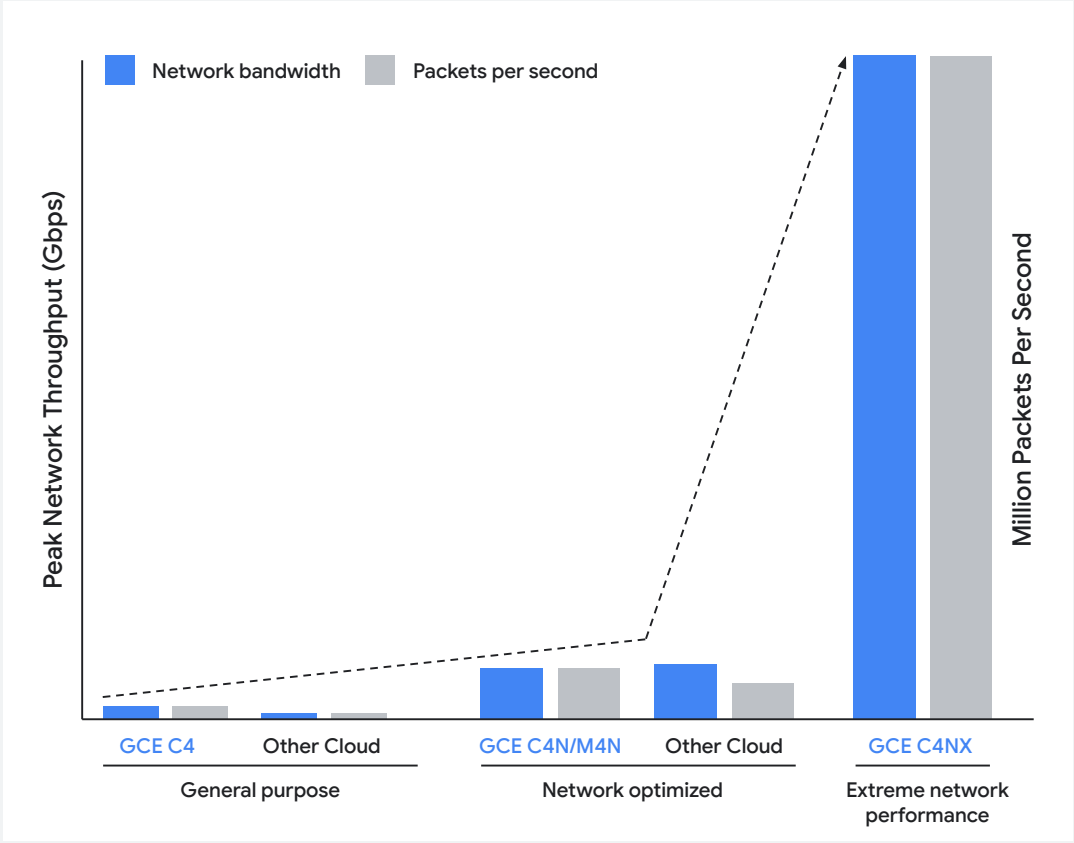
High Memory Optimized

最大 32TB の DDR5 RAMと
1920 個の vCPU を搭載した大規模
データベース向けに ミッション
クリティカルな信頼性を実現

対象ワークロード:
SAP HANA、
SQL Server 、
OLTP/OLAP

Google ネットワーク最適化マシン

Titanium 搭載
アーキテクチャへの
投資により
ネットワーク最適化に
おいて飛躍的向上を
実現します



Z4M

Coming Soon

AI/ML ワークロード向けの大容量のローカル　ストレージ

最大のストレージ 容量

- 最大 168TiB のローカル SSD
  vCPU/LSSD比 1:875
- 最大 5 つの VM インスタンスと
  最大 192 vCPU、1536 GiB メモリを
  備えた Intel EMR を使用した
  ベアメタル オプション

ネットワーク機能の強化

- 最大 Z4M 1台あたり、最大 400 Gbps の
  安定したネットワーク接続
- Z4M ノード間および Z4M と
  GPU/TPU 間のネットワーク遅延を
  低減するための RDMA のサポート

Cluster Director による 可用性の向上

- ネットワーク遅延を低減するために
  GPU/TPU クラスタとの
  コロケーション を実施
- コンピューティング、ネットワーク、スト
  レージ全体にわたるクラスタ管理の簡素
  化

AI/ML 向けおよび Vast Data やSycomp などの並列ファイル システム
向けの非常に大容量のローカル ストレージとスループット

データ検索、ストリーミング、大規模データセットを扱う
ワークロード向けに高い vCPU:LSSD 比率を実現

What’s new in Hyperdisk for AI

Exapools

一般提供開始
既に 10 TiB/s の Lustre を稼働

エクサバイト級の容量と TiB/s を備えたプールで TPU/GPU クラスターを強力にサポート

TPU と GPU にデータを転送する

Hyperdisk ML(HdML)

単一ディスクから
最大 2 TiB/s のスピードを実現

最新世代の TPU および GPU に対応

ポッドの起動を高速化

What’s new in Hyperdisk for Enterprise

Hyperdisk Balanced High Availability

接続されているすべてのインスタンスのパフォーマンスを動的に活用

最大 2.4 GiB/s

Hyperdisk Extreme

最大 25 GiB/s、1M IOPS

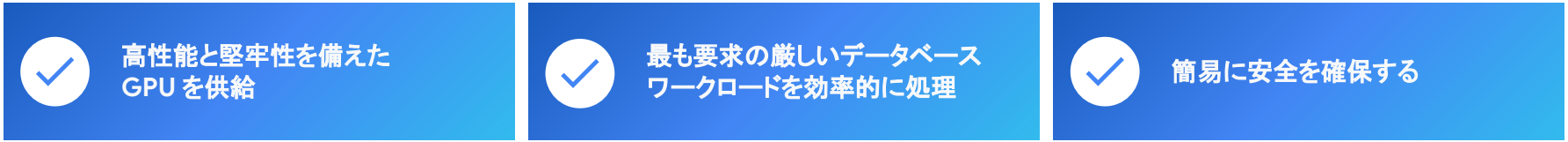
C4N および M4N 対応

スループット向上、vCPU 削減

暗号化

ダウンタイム ゼロの鍵ローテーション

オンラインで鍵を更新

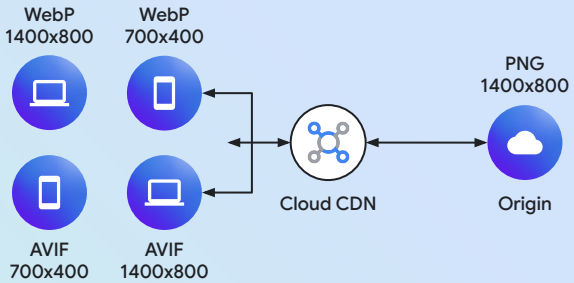


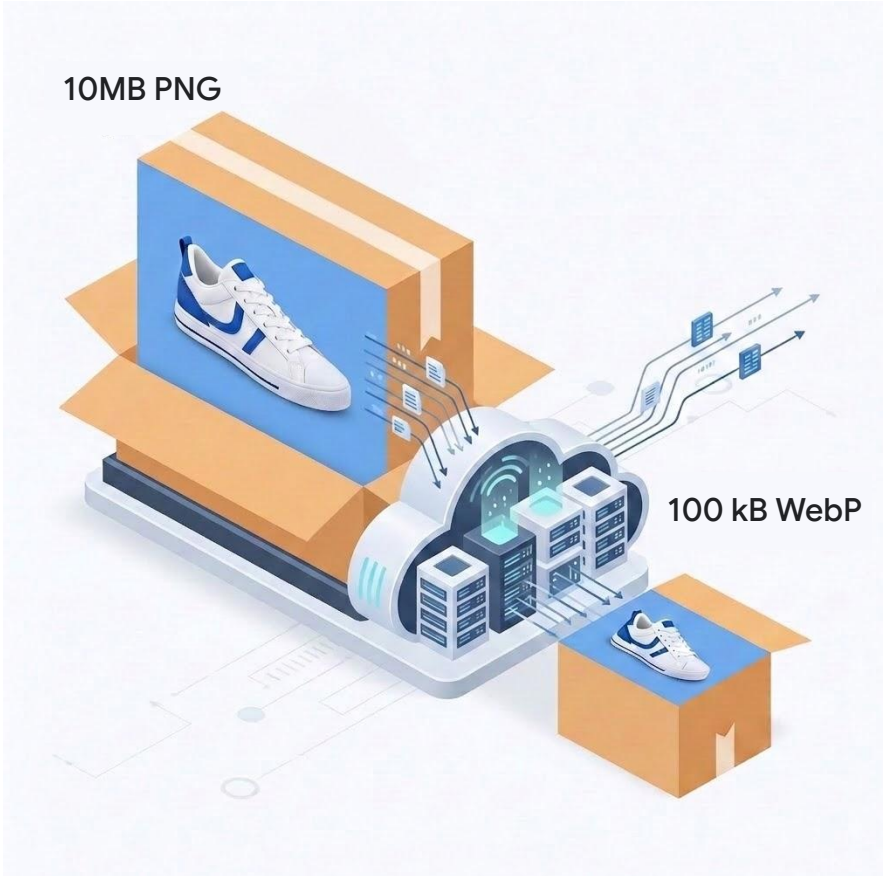
Cloud CDN での画像データ最適化

Image Optimization Preview

オリジン サーバーに負荷がかかる重い画像処理を、
Cloud CDN のエッジにオフロード。
Web アプリケーションのパフォーマンスの向上、SEO の最適化、
そしてデータ転送コストの削減を実現。

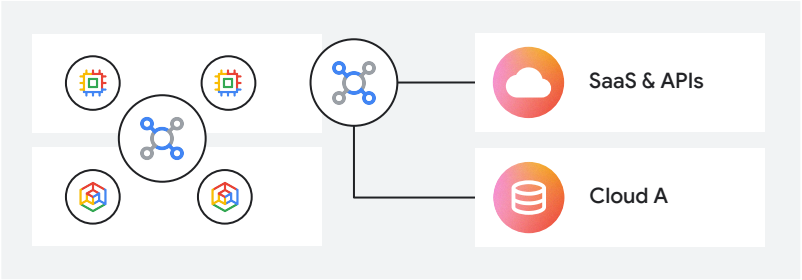
WebP, AVIF, GIF, PNG, JPEG
画像のフォーマット変換、リサイズ、クロップ、ぼかし等の処理が可能





Cloud NGFW:
次世代のファイアーウォール

Google Cloud のネットワーク ファブリックにおける機能の 1 つとして、ワークロードを全方向で保護



ワークロードに対する堅牢なセキュリティを、運用の複雑さを排除して実現

動的なワークロードにも追従する網羅性と、優れた拡張性・シンプルな管理を分散型のファイアーウォールで提供

階層型のポリシー、セキュアタグ、そして Firewall Insights を駆使し、ネットワーク全体のセキュリティを可視化・統制

Palo Alto Networks の防御技術と Google Threat Intelligence でクラウドを牽引する強力な脅威保護を提供

New @ Next

Advanced Malware Sandbox

Palo Alto Networks の技術で 7 万社以上の顧客データで学習した AI モデルを駆使し、既知および未知のマルウェアを 99% ブロック

プレビュー

Internal Load Balancer での NGFW サポート

GKE や Cloud Run、PSC といったサービスに対しても一貫したセキュリティを適用することが可能に

New @ Next

プロジェクト階層のリソース

Cloud NGFW エンドポイント、セキュリティ プロファイル、およびプロファイル グループの作成・管理をプロジェクト レベルで定義可能にし、管理者の運用負荷を大幅に軽減

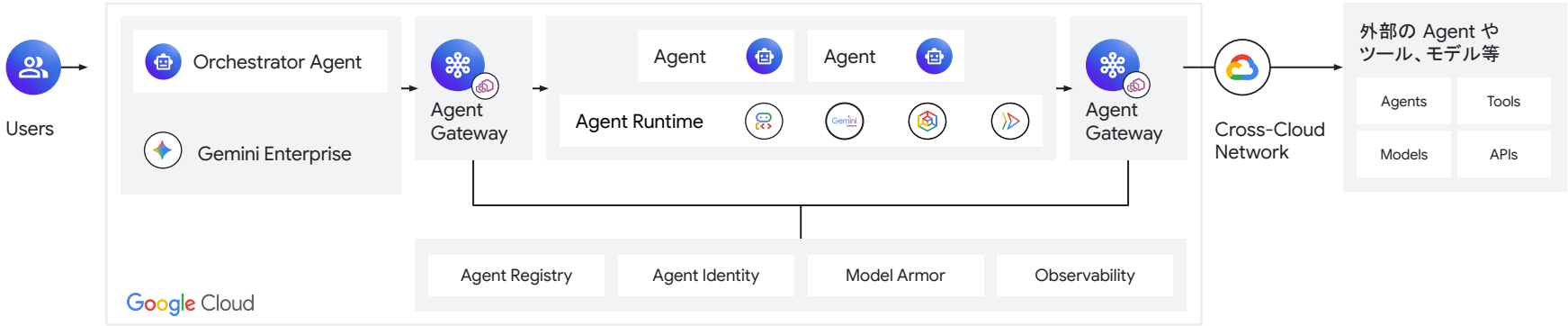
リリース済

ワイルドカードのドメイン フィルタリング

Cloud NGFW Enterprise において、ワイルドカードを用いた L7 ドメイン / SNI フィルタリングをサポート

Agent を脅威から守る：Agent Gateway

Agent の通信に対して統一されたセキュリティとデータのポリシーを適用



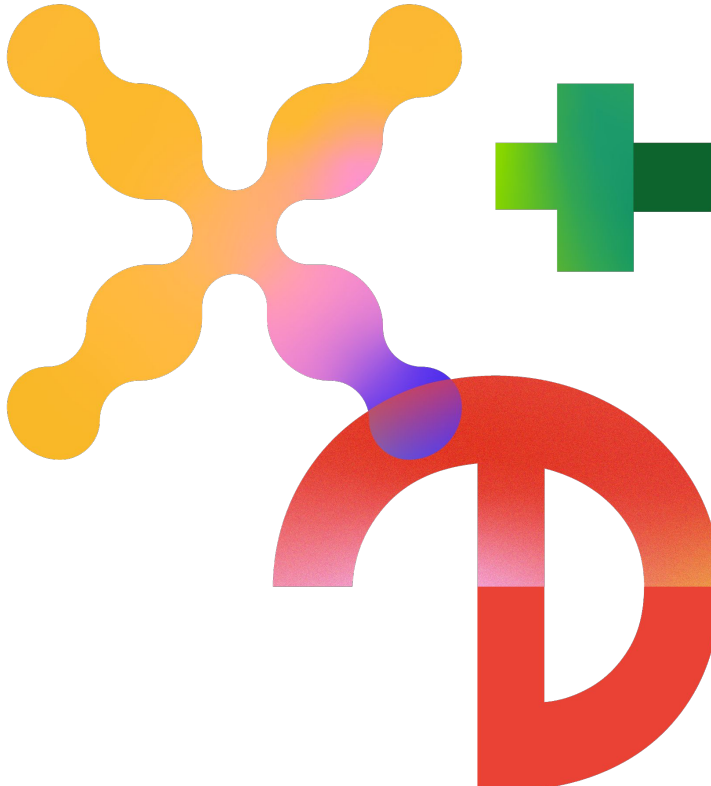
Agent Gateway は 既に実績のある機能郡で実現

- Application Load Balancer
- Secure Web Proxy
- GKE Inference Gatway

Google Cloud Next 26

Agentic Data Cloud

(Data Analytics)



アジェンダ

Agentic Data Cloud

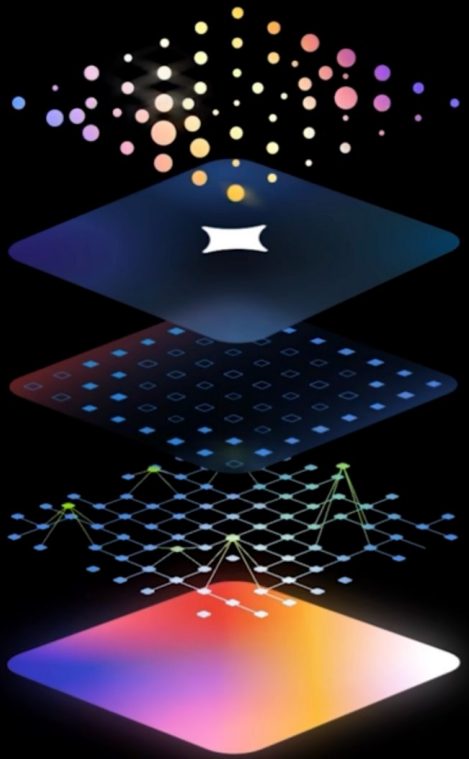
Data Analytics 製品の
主要なアップデート

Agentic AI is reshaping every

task industry job task industry

Only Google

最もセキュアなクラウド上に構築された、完全なデータ & AI スタックを提供します



Agentic Taskforce

Agentic Platform and Models

Agentic Defense

Agentic Data Cloud

AI Hypercomputer

System of intelligence から System of actionへの進化

01

「人間の規模」から**「エージェント (Agent)」の規模へ**

02

受動的なインテリジェンス から **自律的な実行（action)へ**

03

データから **意味を持つ ナレッジへ**

ポートフォリオ全体で製品名を刷新

従来製品名	新名称
BigLake	Lakehouse
Dataplex	Knowledge Catalog
Dataproc	Managed Service for Apache Spark
Composer	Managed Service for Apache Airflow
Looker Studio	Data Studio

お客様への影響

- API、料金体系（SKU）、既存の機能への影響はありません。
- 今回の名称変更に伴う、製品や機能の廃止（デプロイケーション）はございません。

Next 26: 主要な Data Cloud の Update

Universal Context

- **あらゆる構造化・非構造化データを エンタープライズデータソースに**
  - Smart Storage & Context[Preview]
  - Automated curation[Preview]
  - Multimodal extraction[Preview]
  - Zero-copy SaaS[Preview]
- **ガードレールを備えた 大規模なエージェントの活用**
  - Data Products[GA]
  - Golden queries and semantic guardrails[GA]
  - Deep research agent, powered by the Knowledge Catalog[Preview]

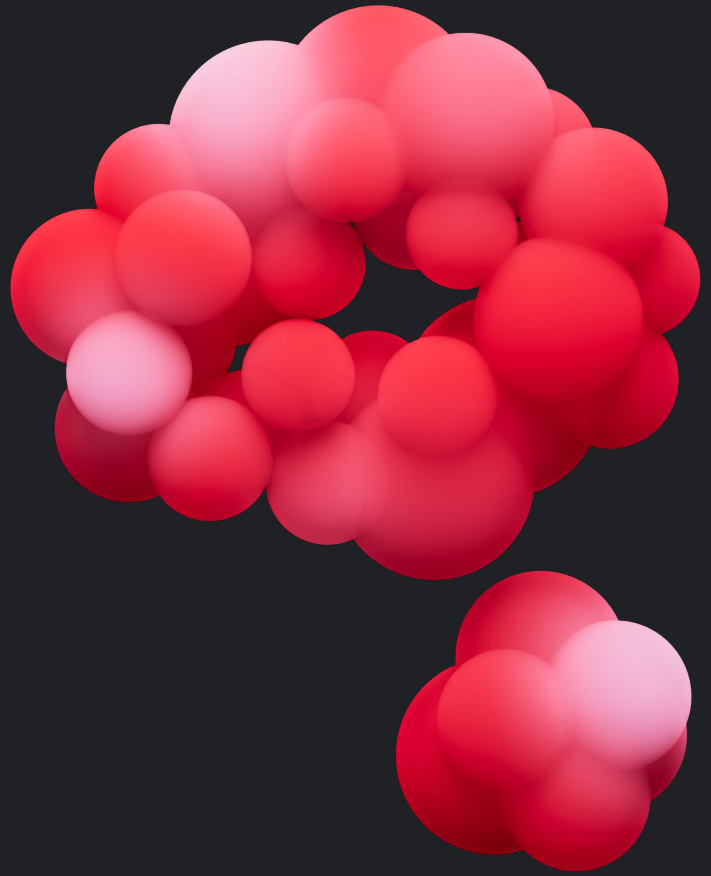
AI エージェント ファーストの ユーザー体験

- Google Cloud Data Agent Kit[Preview]
- Conversational analytics across Data Cloud
- Agentic Workflows [Experimental]
- LookML Modeling Agent[Preview]
- Data Engineering Agent[GA]
- Data Science Agent[GA]
- Database Observability Agent[Preview]

AI-Nativeな Cross-Cloud Lakehouse

- Lakehouse cross-cloud interconnect and cross-cloud caching[Preview]
- Lakehouse catalog federation[Preview]
- Lakehouse Governance[GA]
- Spanner Omni[Preview]
- Lakehouse federation and integrations for AlloyDB[Preview]

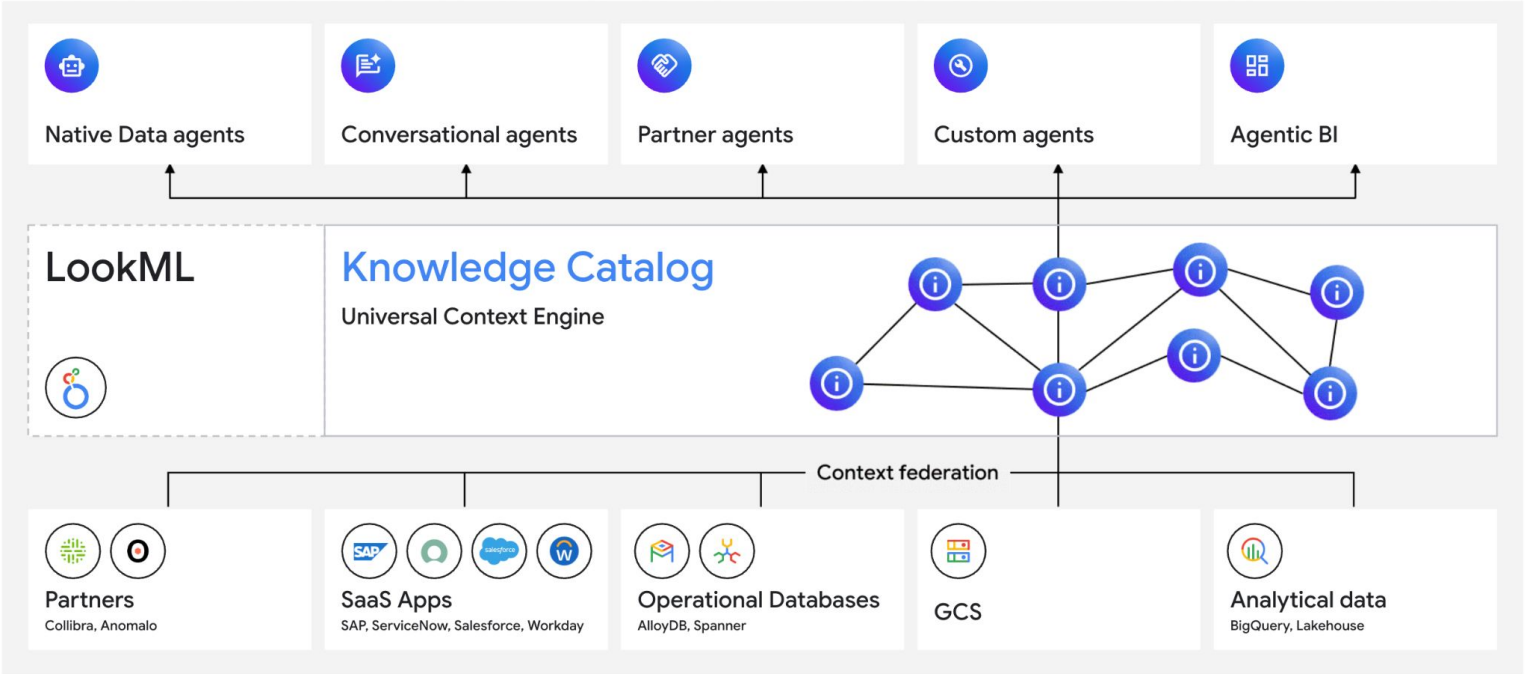
Universal Context



ナレッジ カタログ (Knowledge Catalog)

途切れることのない universal context engine ～全てのデータを知識の源泉へ～

Universal Context



エンタープライズデータを徹底活用した Deep Research

Universal Context

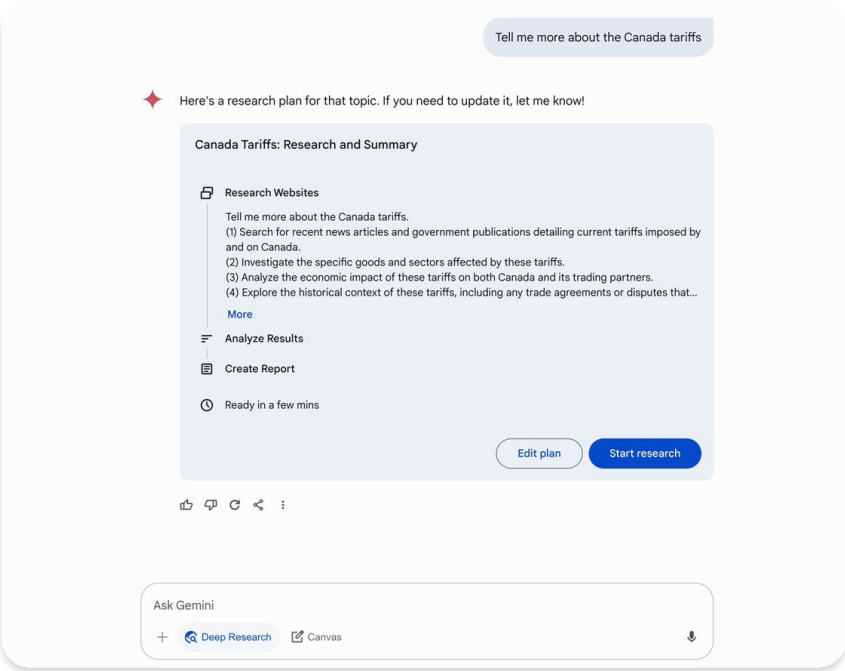
New プレビュー

Deep Research Agent

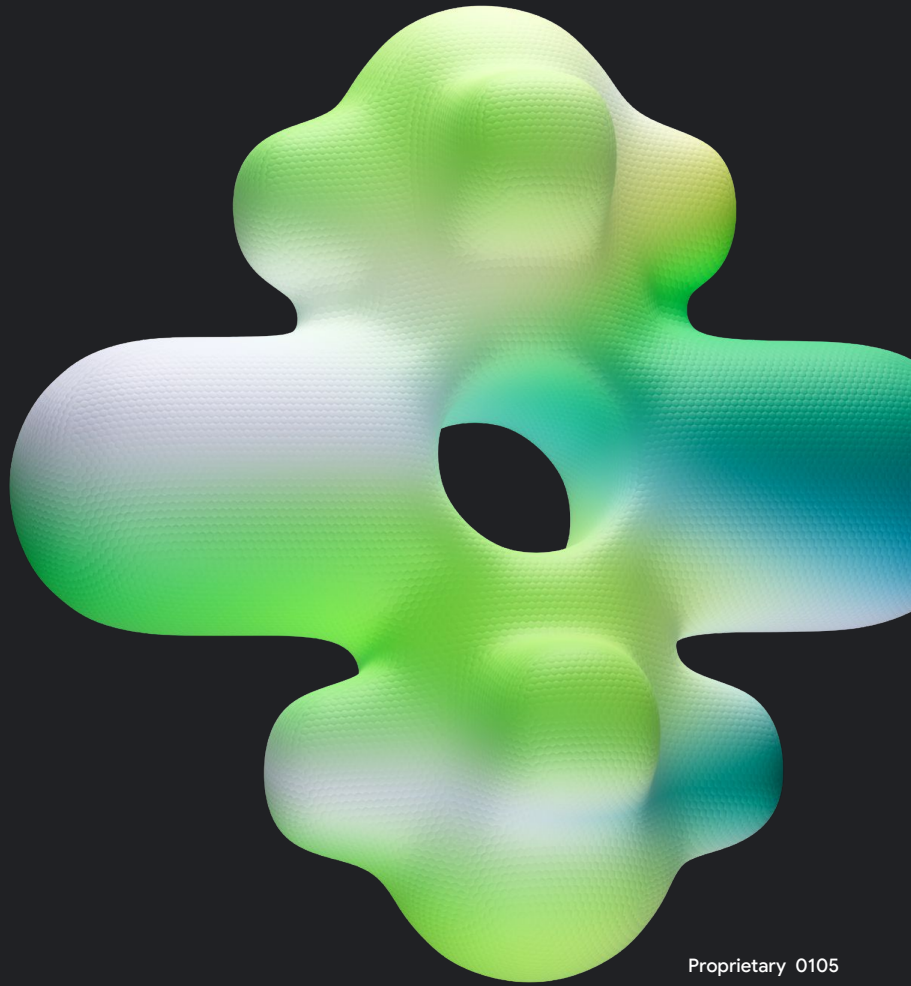
Knowledge Catalog の
活用により高度な調査、分析へ

**社内ドキュメントから BigQuery、
BI 層までを統合 安全かつプロアクティブに、
多角的な調査を自律実行します**

Gemini Enterprise



AI エージェント ファーストの ユーザー体験



New プレビュー

AI エージェント ファーストの
ユーザー体験

Google Cloud Data Agent Kit

データ開発者の体験を再定義する

開発体験の劇的な効率化

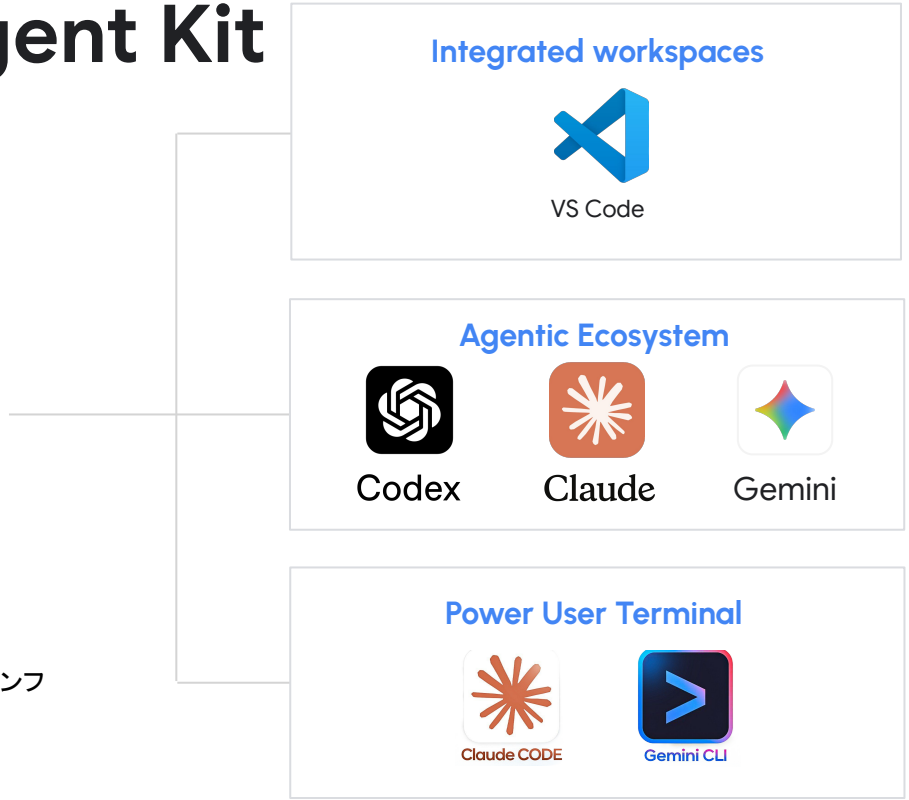
- 自律型エージェントのオーケストレーション：
  データの探索から開発、デプロイまで、
  データライフサイクルの全工程をシンプルに
- 運用システム（Operational）から分析システム
  （Analytical）まで、すべてを一箇所で完結させます。

データから AI へのワークフロー自動化

- 自然言語や直感的な指示(Vibe) によるエージェント、
  アプリ、パイプライン、ワークフローを構築します
- 企業の膨大なデータを即座にAI ワークフローへと
  組み込みます

エンタープライズ基準のスケールとパフォーマンス

- Google Cloud の強力なコンピューティングとデータエンジンなどインフラを活用し、手間なくスケールアップが可能し、
  ミッション クリティカルな業務にも対応する、
  高信頼なパフォーマンスを提供します。

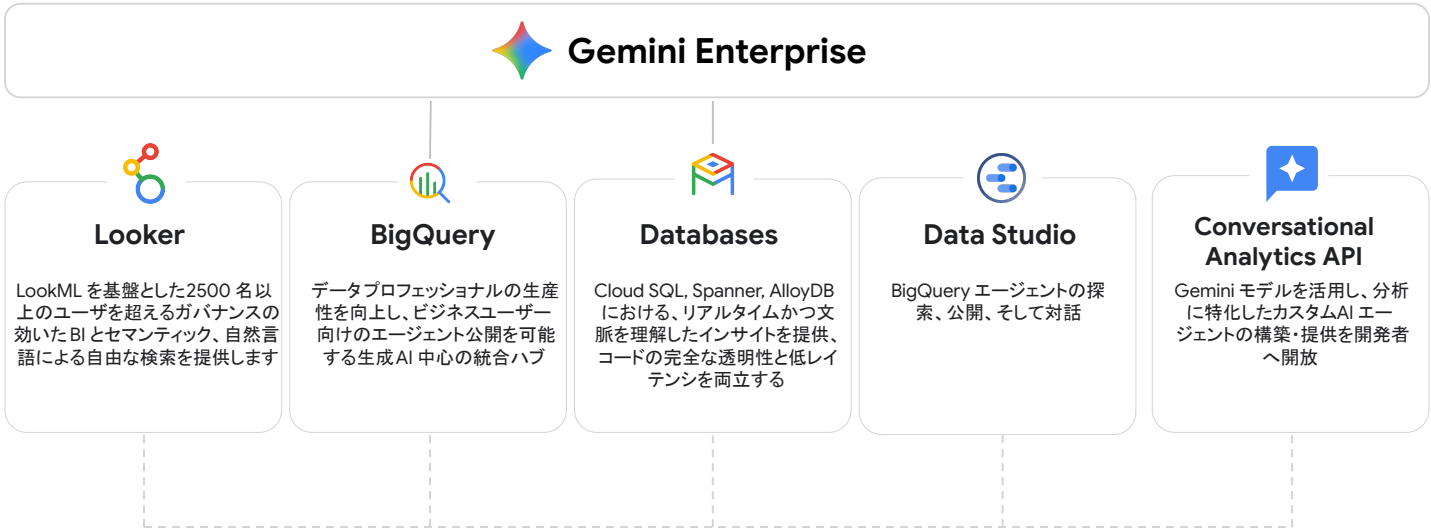


Conversational Agents via Gemini Enterprise

AI エージェント ファーストの
ユーザー体験

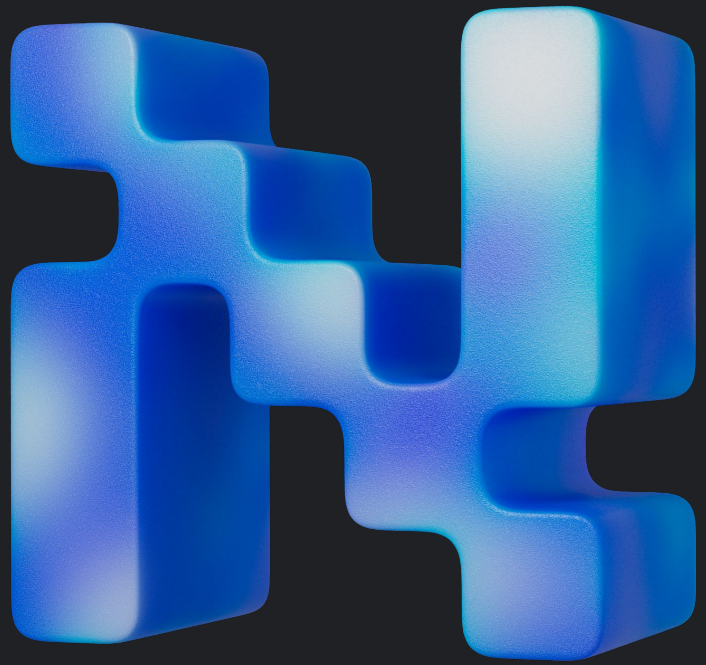
プレビュー

Looker, BigQuery, Database, Data Studio - あらゆるビジネス プロセスを AI との対話で進化させる



Conversational Analytics 利用数 の 対月比 139% 増

AI-native な Cross-Cloud Lakehouse



New プレビュー

Cross-Cloud Lakehouse

クラウドの壁を超え、 BigQuery の分析能力を企業内のあらゆるデータへ

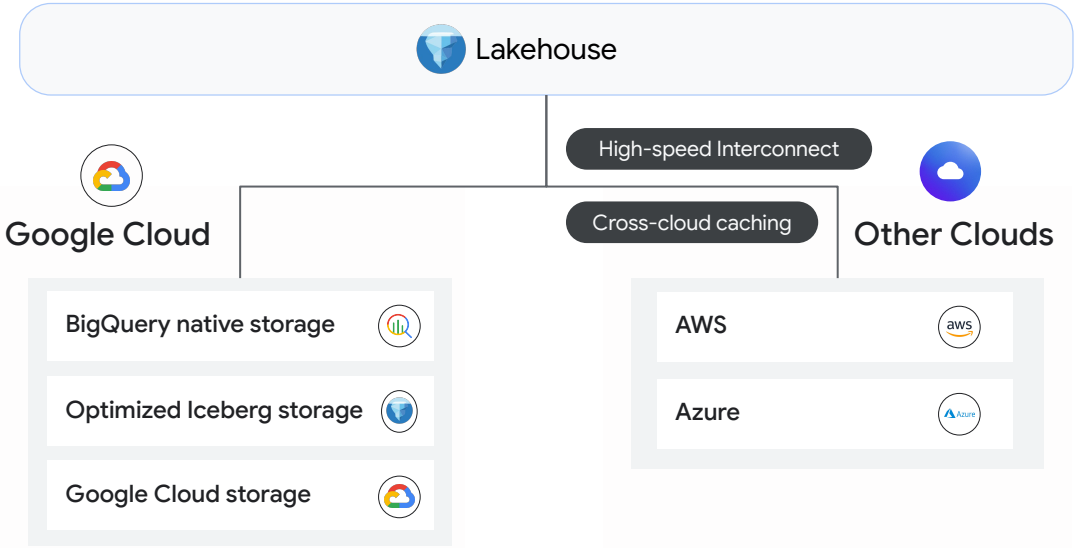
Cross-cloud 相互接続 (interconnect) とマネージド キャッシュ

他社ハイパースケーラーやオンプレミスのデータセンターに対し、高スループット、低遅延、高セキュリティな専用パスとなる Lakehouse マネージドなcross cloud interconnects を提供します

Preview Q3'26

ユーザー管理型 Cross Cloud キャッシュ

インテリジェントなキャッシュ機能により、初回の読み取り時にデータを一時保存。2 回目以降のクエリ（検索）を劇的に高速化します



Data Analytics Announcements

BigQuery

26 Tier 1 launches

Summary of BigQuery launches

General Availability

- Fluid Scaling
- Price-Performance Optimizations
- Vector Search
- Zero-Shot Tabular FM
- TimesFM
- Python UDFs
- Conversational Analytics in BigQuery
- Data Engineering Agent
- Data Science Agent
- BigQuery Tools for Custom Agents
- BigQuery Agent Analytics Plugins
- Data Products
- Golden Queries & Semantic Guardrails
- Multimodal Data (ObjectRef)
- Managed AI functions

Preview

- AI-First Observability
- BigQuery Assistant
- BigQuery Graph Analytics
- BigQuery Measures
- Hybrid Search
- Document Parsing (ai.parse_document)
- Continuous Queries Stateful Processing
- Continuous Queries (Pub/Sub Sinks)
- Data Apps
- SAP BDC for BigQuery
- Agentic Workflows (Private Preview)

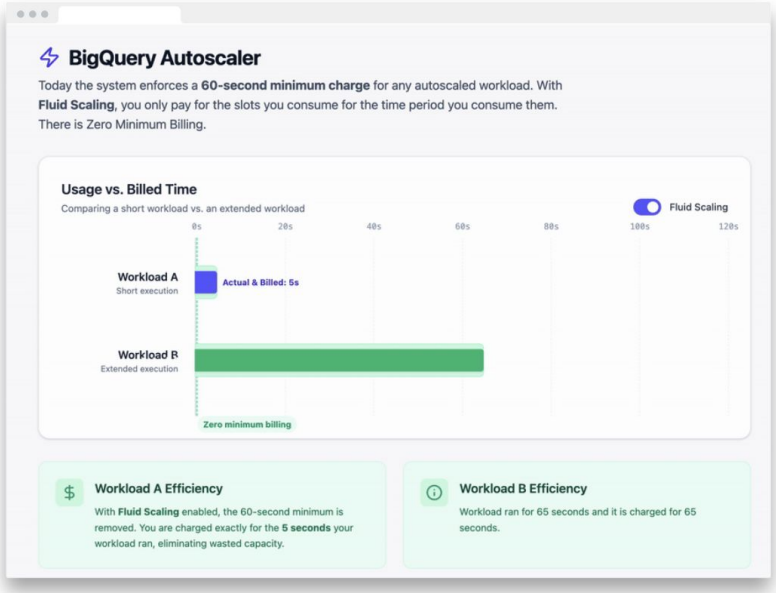
New プレビュー

BigQuery Fluid Scaling

変動の激しいワークロードで最高のコスト　パフォーマンスを実現

実際に使用した計算リソースに対して、秒単位で課金。AI
エージェントによる処理のように、スパイクする
処理に最適で、事前の設定不要で瞬時かつ柔軟に
自動拡張（スケーリング）

- **最低課金時間ゼロ**
- **予測可能なコスト管理**
- **容易なワークロード分離**
- **コスト削減（+34%）**



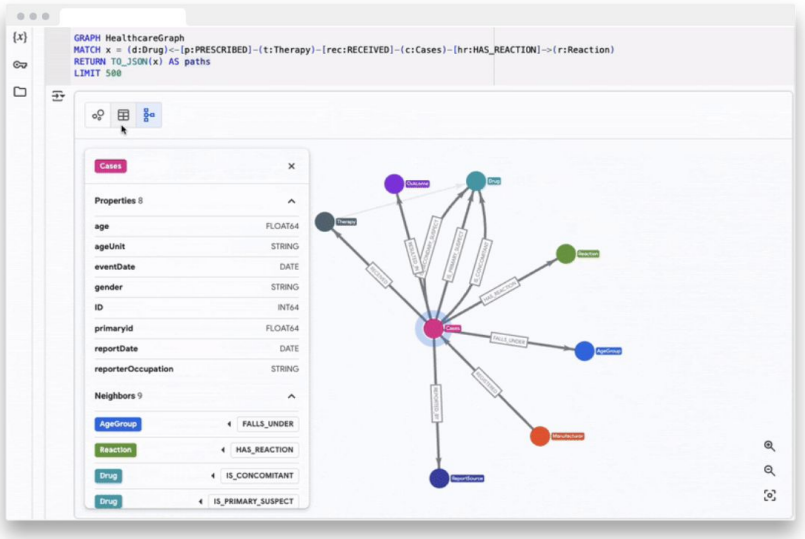
New プレビュー

BigQuery Graph Analytics

散在するデータ間の相関をグラフで構造化し、エージェントに “点と点をつなぐ” 高度な洞察力を与える

Graph over your existing data

```
CREATE PROPERTY GRAPH FinGraph
(
  NODE TABLES(Person,Account)
  EDGE TABLES (
    PersonOwnsAccount
      SOURCE KEY (id) REFERENCES Person (id)
      DESTINATION KEY (account_id) REFERENCES Account (id)
    )
);
```



Walmart
店舗におけるトラフィック（人流・行動）パターンの分析

slb
財務・資産データを横断的に参照して回答するチャットボット

Virgin media O2
顧客のネットワークを通じた不正検知（4,210万件という膨大な顧客ノードを解析）

New プレビュー

BigQuery Measures

ビジネスの重要指標 （メジャー）を グラフ（関係性）と組み合わせ、エージェントの思考を強化する

エンジンレベルのロジック組み込み

標準的な SQL を利用して、ビジネス上の重要指標（メジャー）の計算ロジックを BigQuery の実行レイヤーへ直接埋め込む

推測に頼らない確実な AI推論

AI が「生のテーブルデータから構造を推測する」ような不確実なアプローチを排除し、データの関係性を明示する構造化グラフへと置き換える。 AI エージェントに正確な推論を導き出すための明確な道筋を与え、信頼性の高い回答を実現

複数ステップをまたぐコンテキスト

現場のオペレーション・シグナル （例：カスタマー サポートにおける顧客の感情など）から、最終的なビジネス成果（例：収益など）に至るまでの遠く離れた因果関係を、瞬時かつ正確な経路として直接結びつけて追跡・把握

Agents

New 一般提供開始

Conversational Analytics for BigQuery

Best of Google で実現する “データと会話 ” する体験

Chat with your BigQuery data

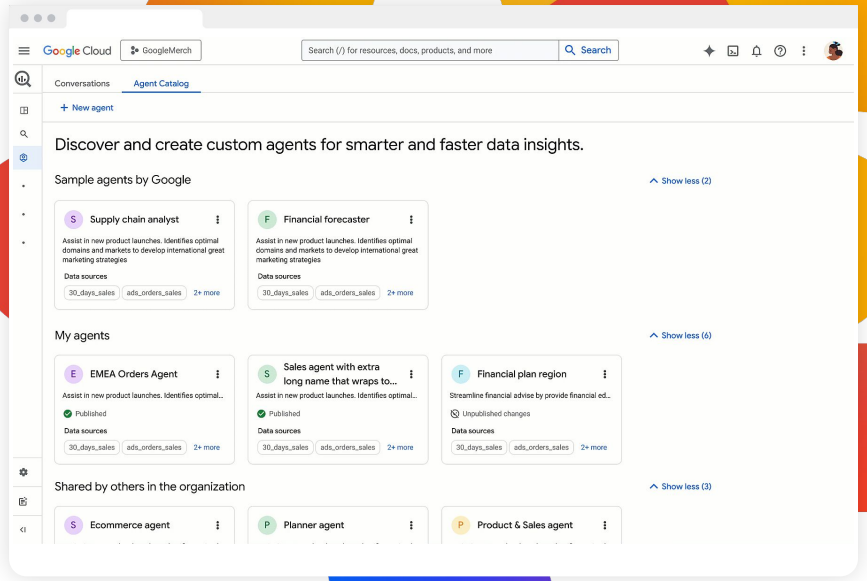
自然言語を用いてBigQuery データから迅速にインサイトを取得する。また、ビジネスユーザー向けにカスタムエージェントを作成可能に

BigQuery governance and security

AI エージェントに対するエンドツーエンドのガバナンスを通じ、透明性、セキュリティ、コンプライアンスを確保する。
企業データへのAI アクセスを管理するポリシーにも対応

Collaboration and control

エージェントをチーム間で共有し、統制された安全なデータアクセスを維持しながら、組織全体でのインサイト活用を促進



New 一般提供開始

BigQuery Agent Analytics Plugins

エージェントの行動分析から新しいインサイトを

シームレスな統合 :

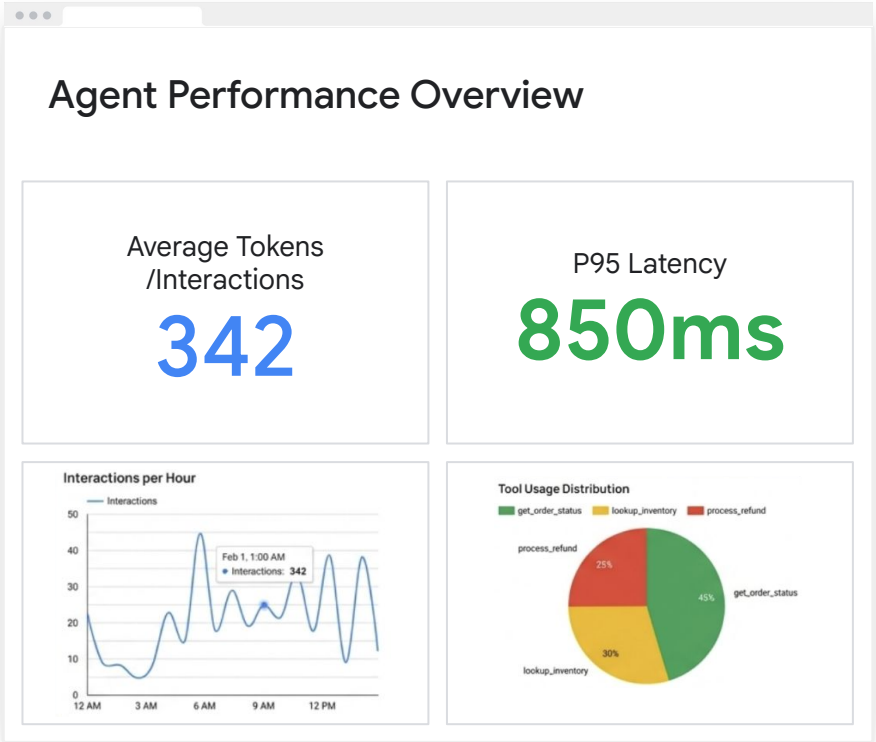
たった1行のコードを追加するだけで、ADK や LangGraph といったフレームワークと連携し、エージェントとの詳細なやり取りをログとして簡単に記録

スケーラブルなインテリジェンス :

BigQuery の強力な基盤を活用することで、インフラ管理の手間（オーバーヘッド）を一切かけることなく、エージェントが生成する膨大なデータの収集と分析を実現

アクションにつながる分析 :

ユーザーの感情（センチメント）、頻繁に寄せられる質問、エージェントの長期記憶の状況などを継続的に分析・モニタリング（監査）し、エージェントのパフォーマンスを最適化するための具体的な改善につなげる



Lakehouse

New プレビュー

Apache Iceberg × BigQuery の双方向連携

Iceberg のオープンな相互運用性 （Read/Write）と BigQuery の高度な分析能力を両立する

多様なエンジンから同じデータにアクセス

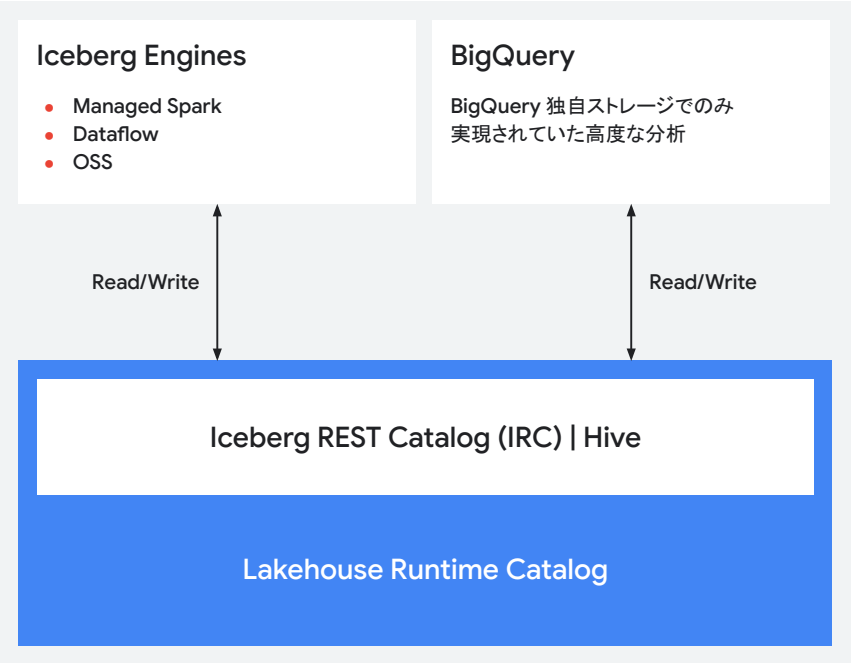
BigQuery と、Spark、Flink、Trino といった OSS（オープンソース）エンジンとの間で、シームレスにデータの読み書きが可能

複数のエンジンからの書き込みに対応

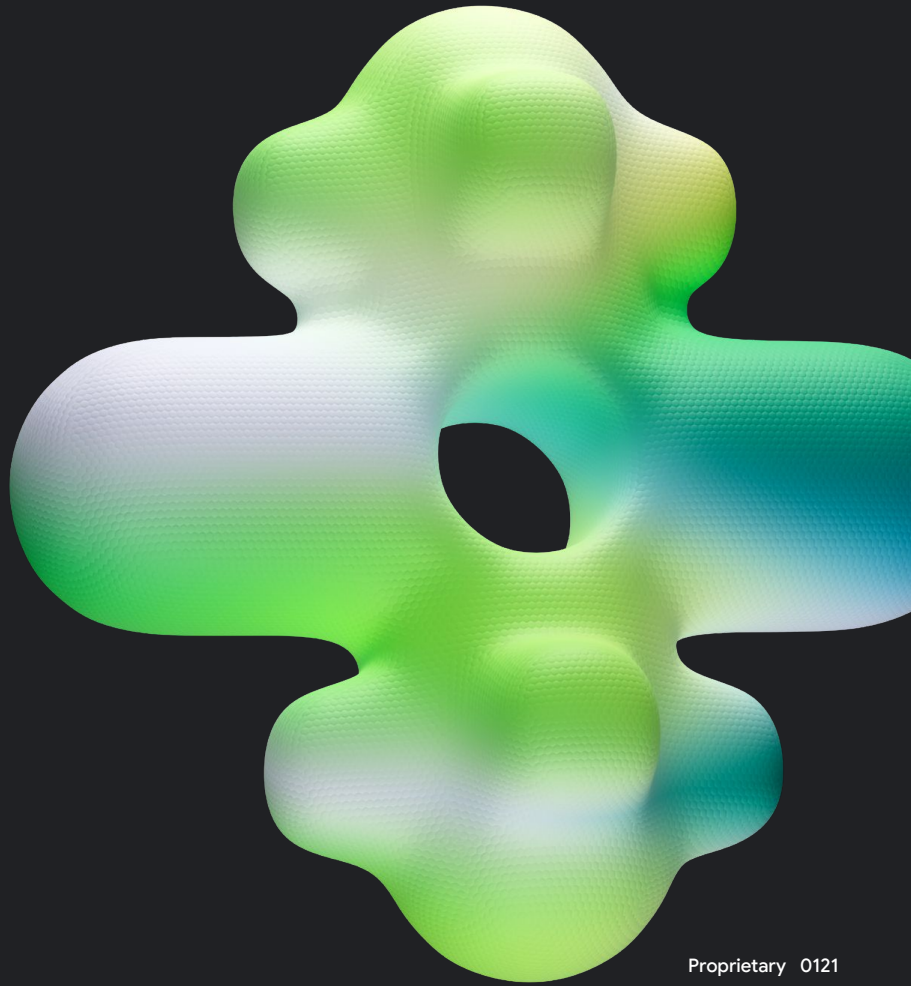
複数の Iceberg 互換エンジンから同時にデータを書き込む処理へのサポートが強化。複数のシステムから書き込みを行っても、メタデータに矛盾が生じず、常に一貫性のある正しい状態に保たれる

オープンなまま “BigQueryの強み” を引き出す

オープン標準規格としての自由な読み書き（相互運用性）を完全に維持したまま、Iceberg フォーマットのデータに対しても、BigQuery ならではの高度なテーブル機能を利用できるように



BI NEXT Announcements



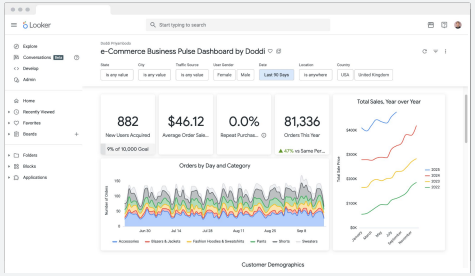
Looker は Agentic Data Cloud の Experience レイヤーへ

人とデータをつなぐ最適化された体験

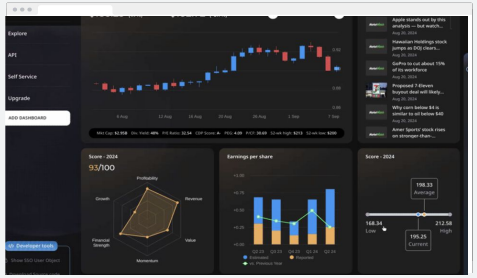
Looker は、AI が自律的にデータを活用する “Agentic Data Cloud” において、**人とデータをつなぐ最適化されたインターフェース（Experienceレイヤー）**へ進化

厳格なバージョン管理、強固なセキュリティ、そして柔軟な API 連携を基盤に、単なる「生の数値」を、**信頼性と知性を兼ね備えた「ビジネスの司令塔（ハブ）」**へと昇華させる

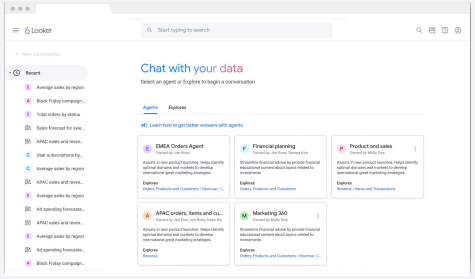
直感的なセルフサービス分析



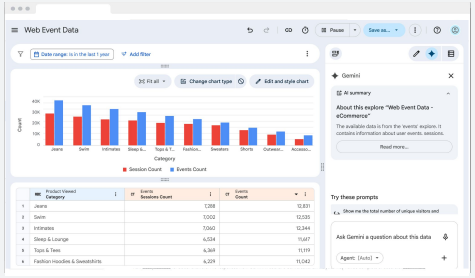
あらゆるアプリに分析機能を実装 (Embed)



AI エージェントとの対話型分析

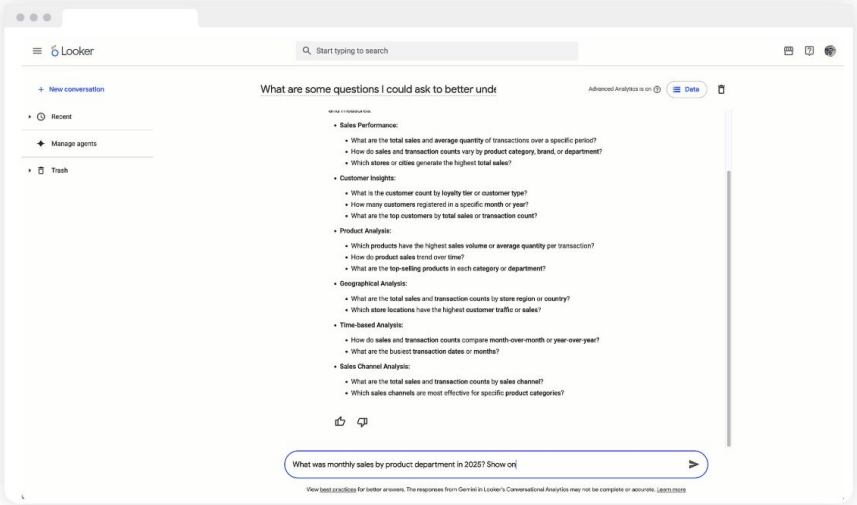


データ探索と高度なモデリング



Conversational Analytics in Looker
対話型エージェントの “コア” として進化

単なる N2SQL から、より深く・速く・信頼できる「インテリジェントな分析パートナー」へと進化
最新の Gemini モデル搭載による推論精度と回答品質の大幅な向上



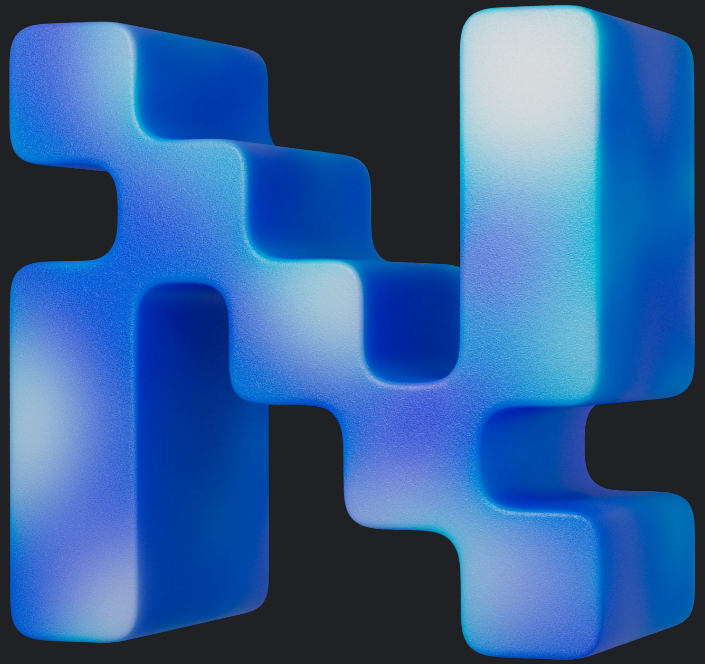
- **精度と推論能力の向上**
- **人間のような深い思考（Think Mode）**
- **エージェント主導の「曖昧さ」解消と信頼性の確立**

Google Cloud Next 26

Agentic Data Cloud

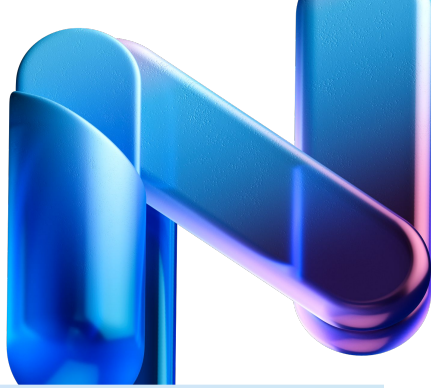
(DB)

AI-native databases and agents



AI Agent 時代のデータベース

すべてのデータベース製品にて、
マネージド Model Context Protocol (MCP) サーバが利用可能



In-memory	Relational			Key value		Document
Memorystore	Oracle AI DB@ Google Cloud	Cloud SQL	AlloyDB	Spanner	Bigtable	Firestore

最先端の AI 機能および Remote MCP サーバ

Gemini CLI/MCP Toolbox　　ベクトル検索　　SQL の AI モデル

DBの構築、運用を支援する Database Agents

Database Agents

Private preview

Onboarding agent

ワークロードに応じたデータベースの選定と型定義を自動化

Private preview

Observability agent

エンドツーエンドの可視化と、フリート全体における障害の予防および診断の実現

Coming soon

Code assist agent

SQL の即時生成、補完、最適化、およびその解説

Coming soon

Testing agent

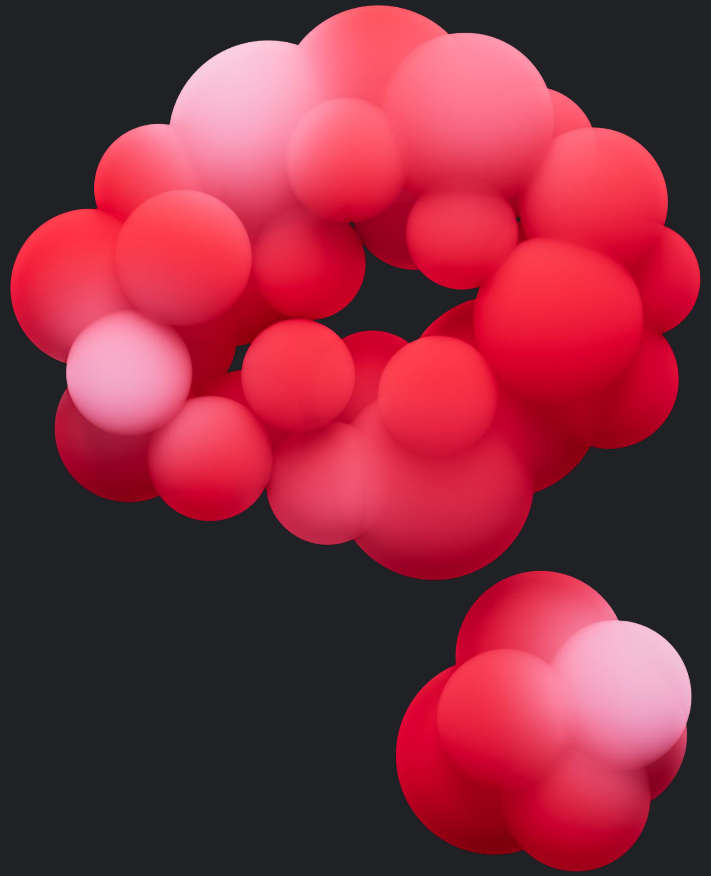
データベース変更の検証用ワークロード シミュレーションを自動化

Coming soon

Migration agent

データベース マイグレーションのライフサイクルを自動化し、リスクを低減

AlloyDB



Hot Standby

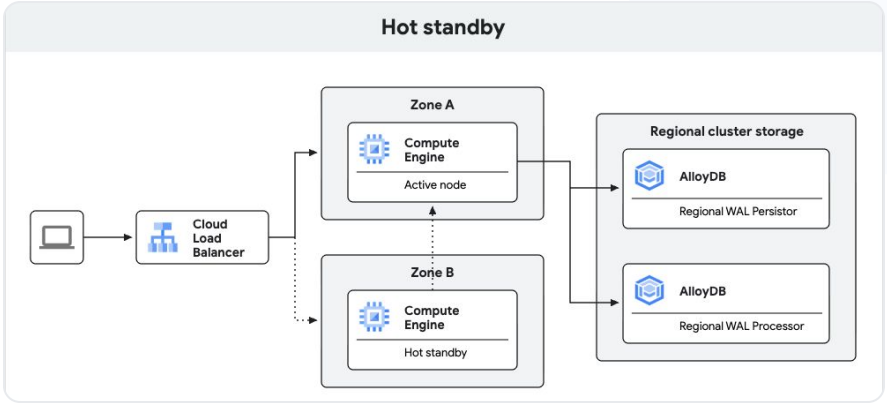
一般提供開始

フェイルオーバー時間の短縮やフェイルオーバー後の高い性能を提供

新機能の ホット スタンバイ により、HA インスタンスのスタンバイノードでアクティブな PostgreSQL インスタンスが稼働し、プライマリからの WAL レコードを常に反映する専用リードリプリカとして機能

フェイルオーバー時間の短縮

起動フェーズの排除と、障害時のプライマリ昇格の高速化によるRTO（目標復旧時間）の改善



一貫した性能

- アクティブなログの適用による、メモリ キャッシュの「暖機状態の維持」
- フェイルオーバー後の一時的な性能低下(”brownout”)を防ぐ「安定したアプリケーション運用」

Transparent Query Forwarding (TQF)

Coming soon

Transparent Query Forwarding (TQF) は、AlloyDB クラスタ内の**ワークロードをスケーリングし、リソースの利用効率を高める** ために設計された機能です。アプリケーション側のコードを変更することなく**参照クエリ(SELECT)をプライマリノードから他のノード(Read Pool)へオフロード**することで、システム全体のスループットを向上

Key Benefits

一貫性の保証
スナップショット シリアライゼーションと、LSN（ログシーケンス番号）監視に基づくインテリジェントなルーティング

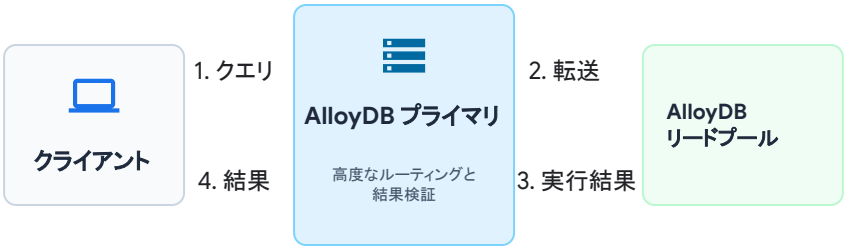
パフォーマンスの向上
ステートメントの検証による、コストや負荷に応じた重いクエリのオフロード

ROI の最大化
既存のキャパシティを有効活用した、
読み取り専用ワークロード（分析クエリなど）のオフロード

運用の簡素化
アプリケーションのコード変更を伴わない、水平スケーリングの実現

セキュアな接続
信頼されたトークン認証による、ノード間の安全な認証

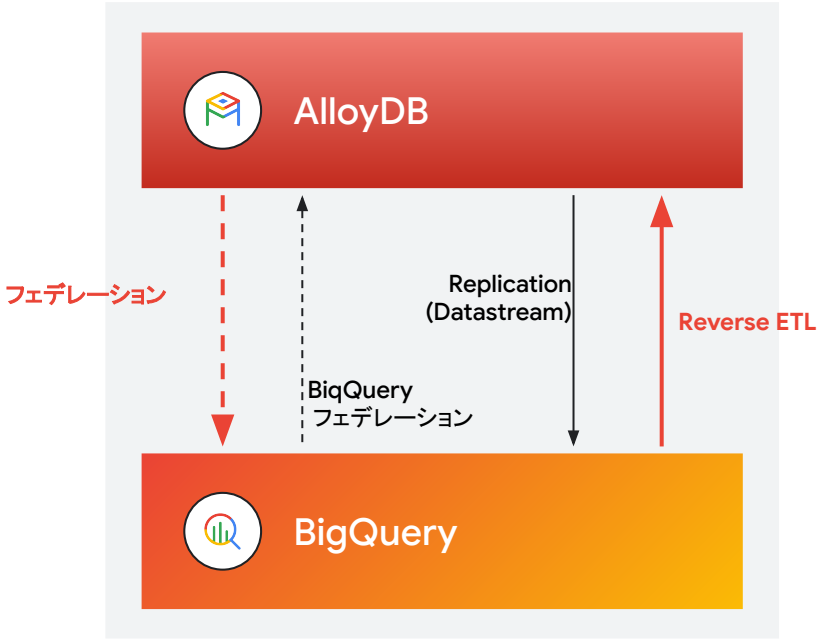
Query Forwarding Workflow



Lakehouse federation for AlloyDB

AlloyDB のデータプレーンから、BigQueryや Iceberg のデータにシームレスにアクセス

- AlloyDB for PostgreSQL から、ネイティブな BigQueryテーブルや BigLake テーブルといった**参照データにアクセス** 可能
- **フィルタリング** や集計処理の**自動プッシュダウン** により、BigQuery 上で柔軟なクエリをフルに活用
- クエリのパフォーマンス向上や、**低レイテンシかつ高同時実行環境でのインサイト提供** を実現するため、BigQuery からのデータインポートや BigQuery/AlloyDB をまたがるマテリアライズビュー ライクなデータ保持を簡単に実現可能に



Industry leading vector search

高速なパフォーマンス、高速なインデックス構築、効率的なメモリ利用

ScaNN

最大 6 倍、10 倍 高速なベクトル検索、フィルタリングされたベクトル検索

最大 16 倍 高速なインデックス作成

New **100 億 以上のベクトル サポート**

通常 1/4 程度 少ないメモリ使用

HNSW

New **最大 4x 高速 なベクトル検索クエリ**

ScaNN index: Googleの 14 年間の研究に基づく Google 検索と同じ検索アルゴリズム。標準の PostgreSQL HNSW 検索と比較して優れたパフォーマンス。並列および分散インデックス構築、インデックスの自動メンテナンス、およびオブザーバビリティをサポート

強化された HNSW index: 標準の PostgreSQL よりも優れたパフォーマンスを発揮。構文の変更もなし。

AlloyDB: 究極のハイブリッド検索エンジン

統合されたバックエンド

ベクトル、SQL 、ドキュメント、全文検索を組み合わせた「単一のバックエンド」

New

全文検索における最先端のパフォーマンス

高品質な結果をもたらす**RUM 拡張機能**とネイティブなBM25 サポート（近日提供予定）

New

シンプルなハイブリッド検索　UDF

Reciprocal Rank Fusion による、カスタムアプリ ロジックの排除

New

容易なインテグレーション

Elasticsearch バックエンドとのシームレスな接続

native BM25 integration coming soon

Use the new hybrid search UDF to combine results from vector and full text search

```
SELECT *
FROM ai.hybrid_search(                                  <-- Hybrid search function
 search_inputs => ARRAY[
     '{
       "data_type": "vector",                             <-- Vector search
       "weight": 0.5,
       "table_name": "documents",
       "key_column": "doc_id",
       "vec_column": "text_embedding",
       "distance_operator": "public.<=>",
       "limit": 5,
       "query_vector": "ai.embedding(''gemini-embedding-001'', ''managed
database'')::vector"
     }'::JSONB,                                           <-- Full text search
     '{
       "data_type": "text",
       "weight": 0.5,
       "table_name": "documents",
       "key_column": "doc_id",
       "text_column": "text_tsv",
       "limit": 5,
       "ranking_function": "ts_rank",
       "query_text_input": "database"
     }'::JSONB
 ],
 include_json_output => false);
```

AI 関数によりカスタム ロジックをシンプルに

人間のニュアンスや文脈を理解するための「常識を注入」

01. サポート対象の AI 関数 GA

ai.if, ai.rank, ai.generate, ai.forecast

02. 新たな AI 関数 New

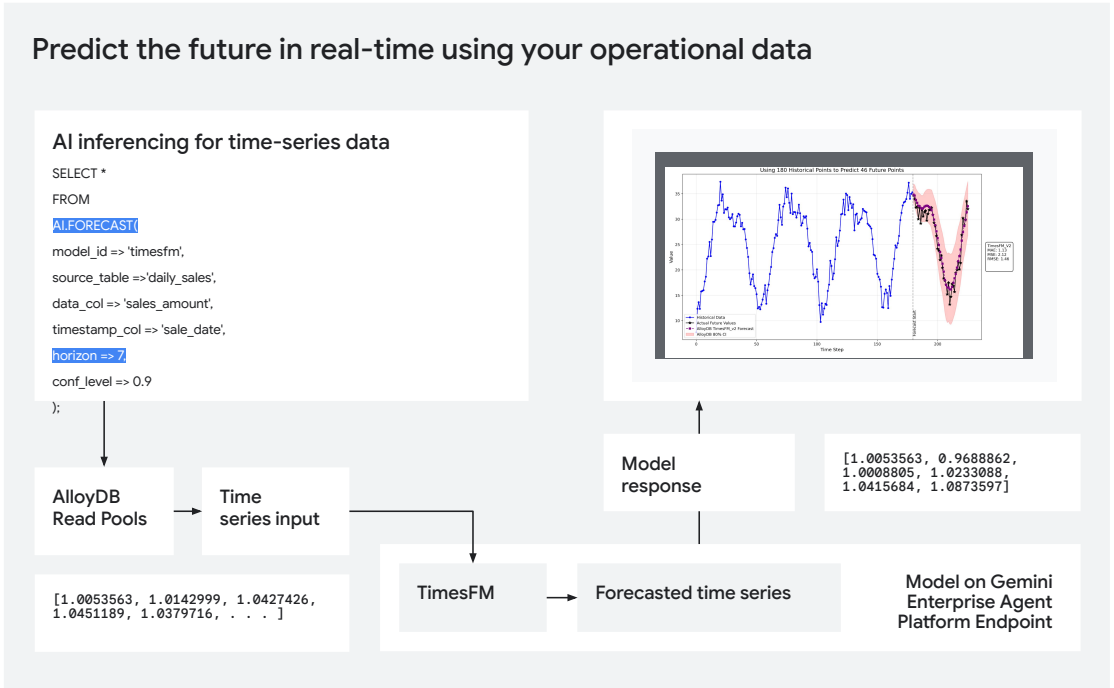
ai.analyze_sentiment and ai.summarize

03. 予測関数 GA

TimesFM を基盤とした、時系列予測による業務効率の向上

04. コスト最適化された ai.if New

Provides superior performance and cost boosts



AlloyDB Omni Next 2026

Enterprise Readiness

重要システム構築に対応する
エンタープライズレディな機能追加

- **Hybrid Cloud 対応強化**
  - **物理・論理レプリケーションを介した Omni とクラウドのデータ連携**
- PostgreSQL 18
- TDE サポート
- AD Group サポート
- FIPS compliance
- GDC software only (セルフ マネージド)
- Standalone UI
- AI parity with Cloud

Support for RPM in Preview

VM（仮想マシン）やベアメタル上でAlloyDB Omni を直接実行できるように。さらに、新たなオーケストレータのサポート。

- PostgreSQL 18 RPM
- RHEL 9, Rocket Linux
- Orchestrators
  - Command Line
  - Ansible
- Automated HA / Backup

Developer Experience

開発者が簡単にデプロイできるようにし、組織へのスムーズな導入を後押し

- Gemini CLI (MCP Toolbox) for Omni support

Spanner

Google で使われている Spanner その実績を継承

- ゼロ メンテナンス
- 常に 99.999% の可用性
- RTO = RPO = 0
- 無制限の読み書きスケーリング
- 柔軟なプロビジョニング
- グローバルでの整合性
- 柔軟なデプロイ
- 予測可能なコスト

Faster is better

パフォーマンス調整における新たな柔軟性

Network

Direct Access NEW

セキュリティや可用性を損なうことなく、ネットワーク ルーティングを簡素化

Isolation

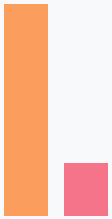
Repeatable Read NEW

読み取り中心のワークロードにおけるロックの取得を削減し、遅延の解消とデッドロックを防止

Multi-region

Read Leases NEW

マルチリージョン構成におけるストロングリード（強整合性読み取り）のレイテンシの削減



4X

レイテンシの向上

TPC-C（1,000 warehouse）ベンチマークにおける、ダイレクト アクセスおよびリピータブルリードを用いた際の p99 レイテンシ

開発者の利便性を追求する

開発者が使い慣れたスキルやエコシステムを活用しながら、Spanner のパワーと信頼性を享受する。

Familiar

SQL

共通の型と関数:UUID 型、INTERVAL 型、ON UPDATE、ON CONFLICT、ユーザー定義関数（UDF）

Integrated

コネクティビティおよびクライアント

ドライバおよびORM の最適化。セッション管理の改善と非同期処理（Async）への対応強化

Portable

PostgreSQL

互換性の拡大、マイグレーション、エコシステムサポートの強化

Innovative

Beyond SQL

柔軟な合成（コンポーザビリティ）を実現するPipes SQL 構文。モデルに直接アクセス可能なAI 関数を内蔵

Pipes: SQL but simpler, more concise, and flexible

```
FROM Produce
|> WHERE
    category IN ('fruit', 'nut')
|> AGGREGATE
    COUNT(*) AS count,
    SUM(cost) AS total_cost,
    STRING_AGG(nutrients) AS
nutrients
  GROUP BY item
|> ORDER BY
    AI.SCORE( --New
      'Rate the healthiness
of…',
      nutrients
    ) DESC;
```

Spanner queues の登場

データベース トランザクションによって安全を確保し、シームレスにタスクの スケジュール と実行 が可能。
もちろんあらゆる規模でスケーラブルに。

非同期処理

ユーザー リクエストから負荷の
高い処理を分離

アトミックな更新

データ更新と同トランザクション内で
タスクをキューに追加するが可能に

スケジュール

タスクのスケジュール機能により、
現在及び将来のイベントの両方を管理

Spanner Omni

Spanner の主要な機能をそのままに、どこへでもデプロイ可能な柔軟性を実現

Spanner の基本性能

- 事実上無制限のスケーラビリティ
- 高い可用性と耐障害性
- 強力なグローバル整合性
- 相互運用可能なマルチモデル機能
- PostgreSQL および Cassandra インターフェースへの対応
- Google Cloud によるエンタープライズレベルのセキュリティとサポート

＋

どこにでも自由にデプロイ

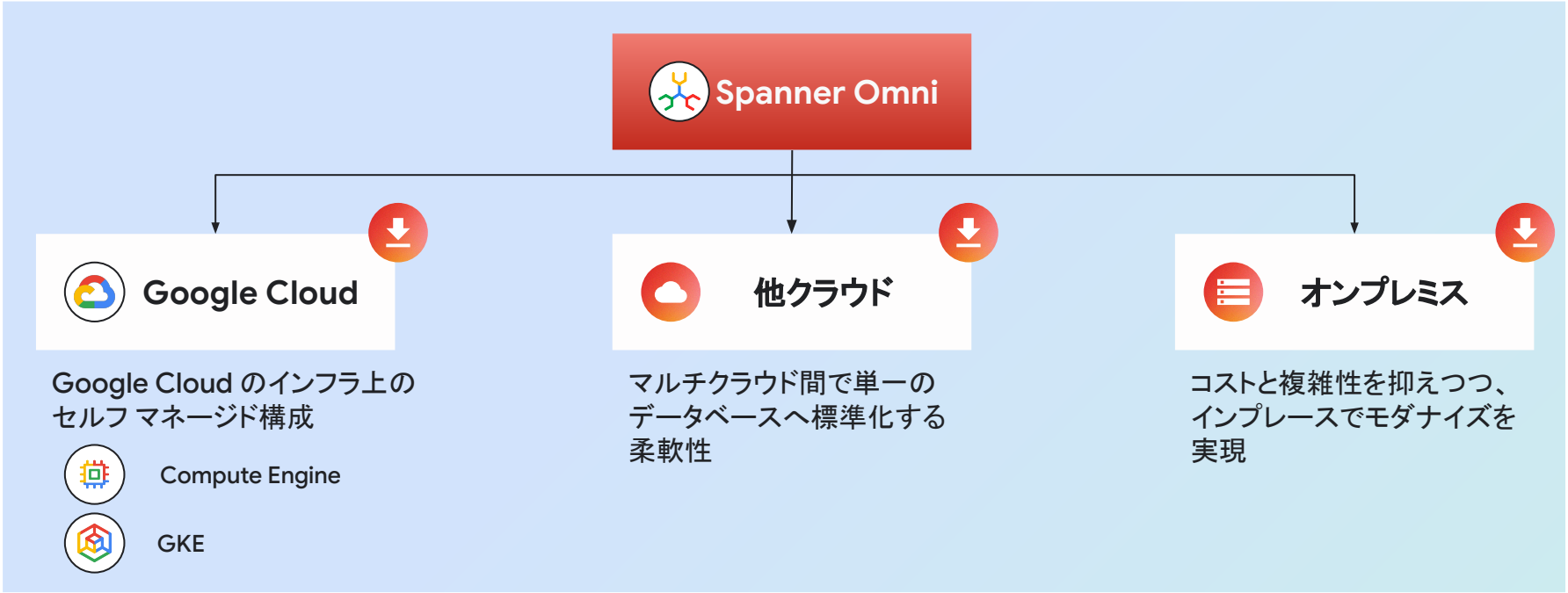
環境
PC、オンプレ環境、他クラウド

場所
Google Cloud に限らずどこでもデプロイ可能

スタック
VM、ベアメタル、Kubernetes、Linux コンテナ

ネットワーク環境
オンラインとオフライン環境両方

Spanner Omni のデプロイ先



Spanner vs Spanner Omni

	Spanner	Spanner Omni
責任	フルマネージ	セルフマネージ
運用管理	Google	カスタマー
SLA	リージョン利用時：99.99% マルチ リージョン利用時：99.999%	顧客管理のため SLA 提供なし (実際の Spanner に似た構成であれば同等の可用性は可能
環境	Google Cloud	ノート PC オンプレミス サードパーティ クラウド
構成	リージョン（単一リージョン） マルチ リージョン	シングル サーバー シングルゾーン リージョン間 / マルチゾーン マルチ リージョン / マルチ クラスター
機能	Full feature set	Google の環境に依存した機能は除外

Spanner Omni のスケーラビリティ

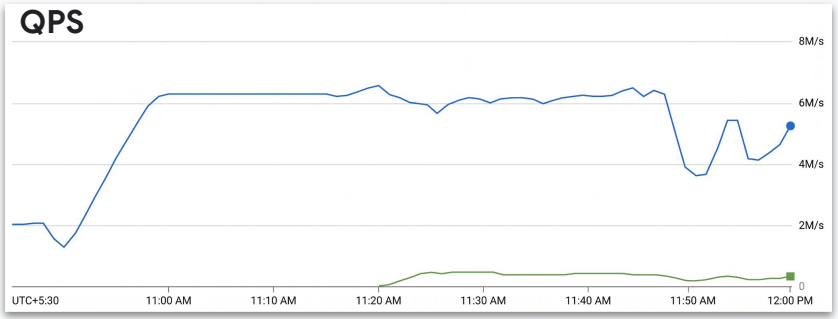
Spanner Omni は単一の環境で

ペタバイト級
のデータ

に対して

数百万
QPS

ものスループットで処理可能



Spanner Omni をお試しください！

Developer Edition:

非本番環境専用に設計された、機能を限定したバージョン

- 事前ロード済みデータセット
- サンプルクエリ
- 90 日間の有効期限（制限）
- エンタープライズ レベルのセキュリティ機能および運用機能の除外
- コミュニティ サポートのみ提供

マルチモデル をさらに向上

Spanner グラフ

SQL 対応

スキーマとスキーマレスの両立

組み込みのグラフ可視化機能

統合されたグラフ アルゴリズム New

グラフで VIEW に対応 New

UI を使ってグラフを構築 New

Spanner の全文検索

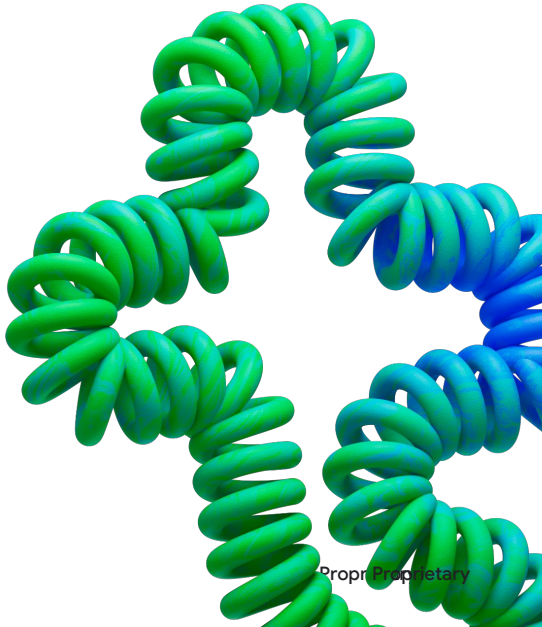
KNN & ANN によるベクトル検索

Google 検索由来の専門性

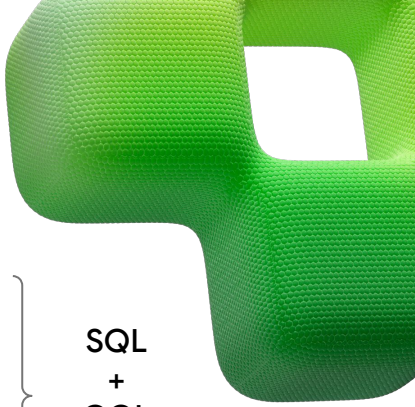
構造化データおよび
非構造化データ（JSON）のサポート

カスタム辞書 New

インライン フィルタリングと
パフォーマンスの高速化 New



AI アプリケーション向け インテリジェント検索の実現



```
GRAPH SpanMartProductPurchaseGraph
MATCH
  (u1:User {id: 1})-[:IS_FRIEND]->
  (u2:User)-[:HAS_PURCHASED]->(p:Product)
WHERE p.id IN
  (SELECT id FROM Products
   ORDER BY APPROX_COSINE_DISTANCE(
      embedding,
      ML.PREDICT(EmbeddingModel, "Brown hiking boots")
   LIMIT 200)
UNION ALL
  (SELECT id FROM Products
   WHERE SEARCH (description_tokens, "Brown hiking boots"))
   LIMIT 200)
RETURN p.id, p.price;
```

- グラフ
- SQL
- ベクトル
- ベクトル
- 全文検索
- SQL

SQL + GQL

Google Searchの知見を反映するSpanner 全文検索

- Google 検索由来の専門性
- カスタム辞書 New

任意のシノニムを登録できうるカスタム辞書対応により、拡張されるキーワードを制御することが可能に

ユーザーによる検索キーワード

"hair dye" で検索

クエリする「検索キーワード」を拡張するオプション

```
SELECT * FROM Products
WHERE SEARCH(Description_Tokens,
"hair dye", enhance_query=>true)
```

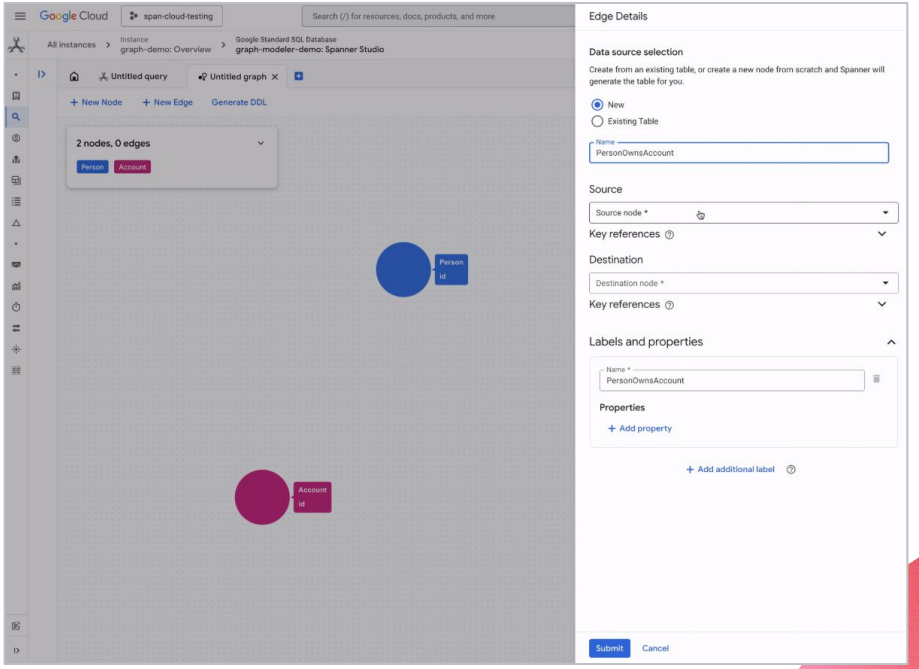
クエリを自動拡張

**“(hair OR hairdye) AND
(dye OR color OR coloring OR colour OR colouring OR dyed OR dyeing OR dyes OR dying OR hairdye)”**

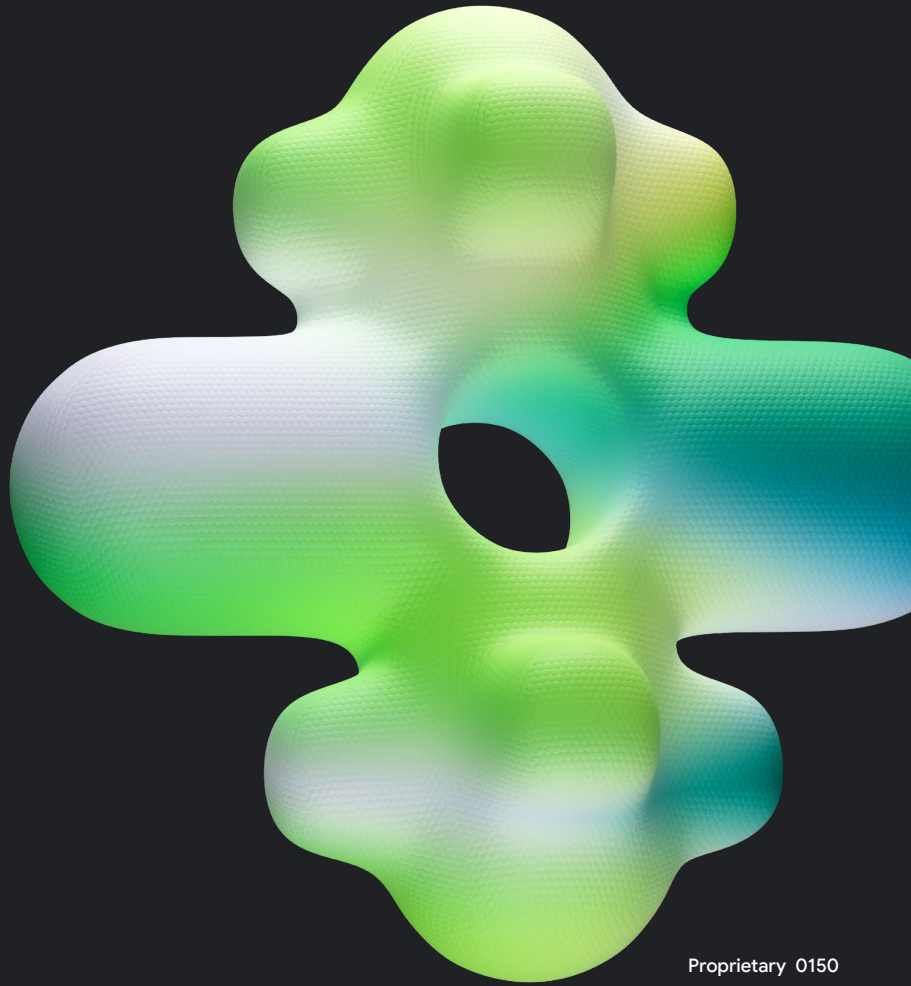
ビジュアル スキーマ モデリングの導入

プレビュー

- グラフスキーマの視覚的な定義
- GUI を用いた自動グラフ生成
  （DDL の手動作成は不要）
- グリーン フィールド（新規）およびブラウン フィールド（既存）の
  両開発環境に対応
- スキーマ設計のベスト プラクティスを組み込み済み



- Cloud SQL
- Oracle
- Bigtable
- Firestore
- Memorystore

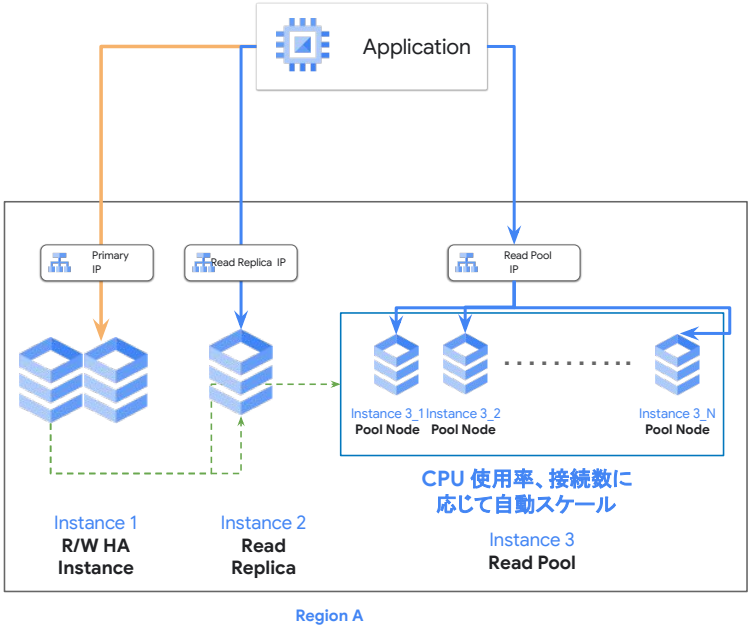


Cloud SQL の新機能

AI & Agent ready：**Data Agent** を利用することで最適なデータベース選択やパフォーマンスとシステムの健全性のプロアクティブな監視をサポート。**Remote MCP サーバ**の GA により、データベースとの自然言語でのやり取り、AI によるアプリ開発の支援、クエリパフォーマンスの最適化、エージェントによるデータベースのトラブルシューティングを行うことが容易に。

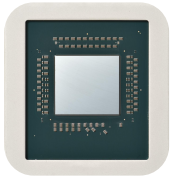
大規模ワークロードにおけるスケーラビリティ管理を容易に
：Auto scaling read pool により、CPU 使用率や接続数に応じて読み取りプールを動的に増減可能に(MySQL/PostgreSQL)。
SQL Server においても Read Pool をサポート。

開発者にとって使いやすく、多様な選択肢　：**Data API** のサポートにより、Cloud Run Functions などから API 経由で SQL を実行可能に。PostgreSQL 18, SQL Server 2025 のサポートなどの新しいメジャー バージョンへも追従。



Auto scaling read Pool の構成イメージ

Cloud SQL Enterprise Plus C4A in asia-northeast1

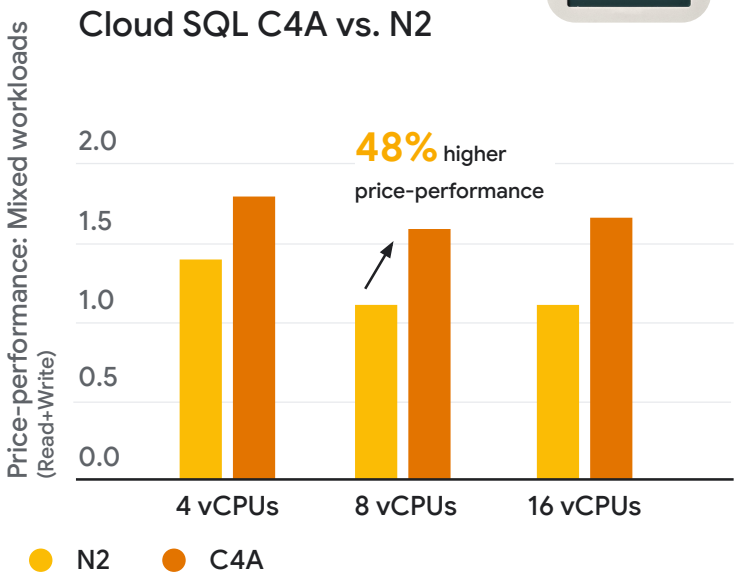


優れた価格性能：Axion ベースの C4A VM 上で稼働する Cloud SQL は、transactional workloads において N series VM と比較して約 50% 優れた価格性能を提供。

向上したスループット：Axion ベースの C4A VM 上で稼働する Cloud SQL は、Amazon の同等の Graviton 4 ベースの製品と比較して最大 2 倍優れたスループットを提供。

Hyperdisk 対応：Hyperdisk Balanced を使用することで、ストレージについてもコストパフォーマンスを向上。

asia-northeast1（東京）にて近日利用可能に



Axion ベースの C4A 仮想マシン上で実行される Cloud SQL は、トランザクション ワークロード向けに、N シリーズ VM と比較して**最大 48% 優れた価格性能**を提供します

Oracle AI Database@Google Cloud
東京、大阪含む 15 リージョンで提供

North America
- アッシュバーン
- アイオワ
- ソルトレイクシティ
- モントリオール
- トロント

South America
- サンパウロ
- メキシコ

Europe
- ロンドン
- フランクフルト
- ミラノ
- トリノ

Asia Pacific
- **東京**
- **大阪**
- メルボルン
- シドニー
- ムンバイ
- デリー

Oracle AI Database@Google Cloud GoldenGate が登場

プレビュー

GoldenGate と BigQuery を活用したニアリアルタイム AI データ ファブリック

マネージド統合により、Oracle AI Database と Google Cloud BigQuery 間での
ニアリアルタイム（サブ秒のレイテンシ）なデータ レプリケーションを実現。

Embedded AI Services

Gemini Enterprise Agent Platform のモデルを GoldenGate 26ai パイプラインから直接呼び出し、感情分析、PII（個人情報）マスキング、データ エンリッチメントを実行。

Zero-Downtime Migration

オンプレミスの Oracle ワークロードを、アプリケーションのダウンタイム ゼロで Google Cloud へシームレスに移行。

Automatic Schema Evolution

ソース データベースの変更を、手動操作なしで BigQuery へ動的に反映させる。

Unified Management

マルチクラウド運用におけるデータ レプリケーションを、Google Cloud Console から直接管理・監視。

Bigtable in-memory 登場

プレビュー

さらなるスループット

読み取り処理で最大
10 倍のコスト削減

ホットスポット無し

単一行に対して
最大 120K QPS

速い！

サブミリ秒の
読み取りレイテンシ

Bigtable Cloud Next 2026 ハイライト

利用可能

SQL Pipes

機械学習や時系列分析におけるデータパイプライン開発を簡素化するため、SQL 構文によるパイプライン化（Piped SQL）を採用。

利用可能

地理空間クエリ

ジオハッシュ、空間関係、GeoJSON / KML / WKT / WBT などのインポート/エクスポート形式に対応した 70 以上の関数を提供。Bigtable での地理空間検索を容易に構築可能。

利用可能

ウインドウ関数

時系列データおよびリアルタイム分析に向けた SQL ウィンドウ関数：移動平均、累積合計、パーセンタイル、ランキングなど。

Next 26 にて一般提供開始

継続的マテリアライズド ビュー

データ ストリームから運用上のインサイトを抽出。SQL ベースの集計を実行し、高速なグローバル ライトから非同期セカンダリ インデックスを生成。

Next 26 にて一般提供開始

Protocol Buffer (protobuf) Support

Bigtable SQL を用いた Protobuf データの読み取り、または BigQuery フェデレーション経由でのアクセスが可能。継続的マテリアライズド ビューにより、Protobuf ベースのアプリケーションからリアルタイムなインサイトを取得。

Next 26 にてプレビュー版提供開始

Pub/Sub サブスクリプション

Pub/Sub トピックから Bigtable テーブルへ直接メッセージを書き込む、新しいネイティブ Pub/Sub サブスクリプション タイプ。

Memorystore ノードのポートフォリオ拡大

1.25GB のプロトタイプから 27.5TB のエンタープライズ クラスターまで、Memorystore はあらゆるワークロードに対し、シームレスなスケーリングを提供します。

小規模向け

(Perfect for testing and lightweight AI agents)

Shared-Nano
Shared Core
1.4 GB
NO SLA

Custom-Pico NEW
2 vCPU
1.25 GB
NON-CLUSTER ONLY

Custom-Micro NEW
2 vCPU
2.5 GB
NON-CLUSTER ONLY

Custom-Mini NEW
2 vCPU
3.5 GB
NON-CLUSTER ONLY

CPU 処理タイプ

(Balanced for fast Vector Search and rapid retrieval)

Standard-Small
2 vCPU
6.5 GB

HighMem-Medium
2 vCPU
13 GB

HighCPU-Medium NEW
8 vCPU
13 GB

Standard-Large NEW
8 vCPU
26 GB

ハイメモリ タイプ

(Built for massive RAG architectures and huge datasets)

Highmem-XLarge
8 vCPU
58GB

Highmem-XXLarge NEW
16 vCPU
110 GB

Firestore Cloud Next 2026 ハイライト

Next 26 にて一般提供開始

Pipelines API

コレクション間での JOIN 処理をサポート。バルクデータ操作によるデータのバックフィル、サニタイズ、正規化の実行

Next 26 にてプレビュー版提供開始

地理空間クエリ

位置情報を活用したアプリケーションの構築（目的地の検索機能など）

Next 26 にてプレビュー版提供開始

Usage Insight と Knowledge Catalog

Usage Insights を使用して、コレクショングループごとにトラフィックをデバッグ。Knowledge Catalog を通じてデータモデルを調査。

Next 26 にてプレビュー版提供開始

全文検索

トランザクション データとの整合性が高い検索結果を提供し、統一されたエクスペリエンスを実現

Next 26 にてプレビュー版提供開始

Change Stream

リアルタイムのデータ変更を検知し、BigQuery などのサービスへ同期（あらゆる規模で対応可能）

Next 26 後に一般提供開始

MongoDB でサイズ上限の向上

MongoDB 互換のモードにおいて最大 16 MiB までのドキュメントに対応。

Google Cloud Next 26

Agentic Security

エージェント型 SOC

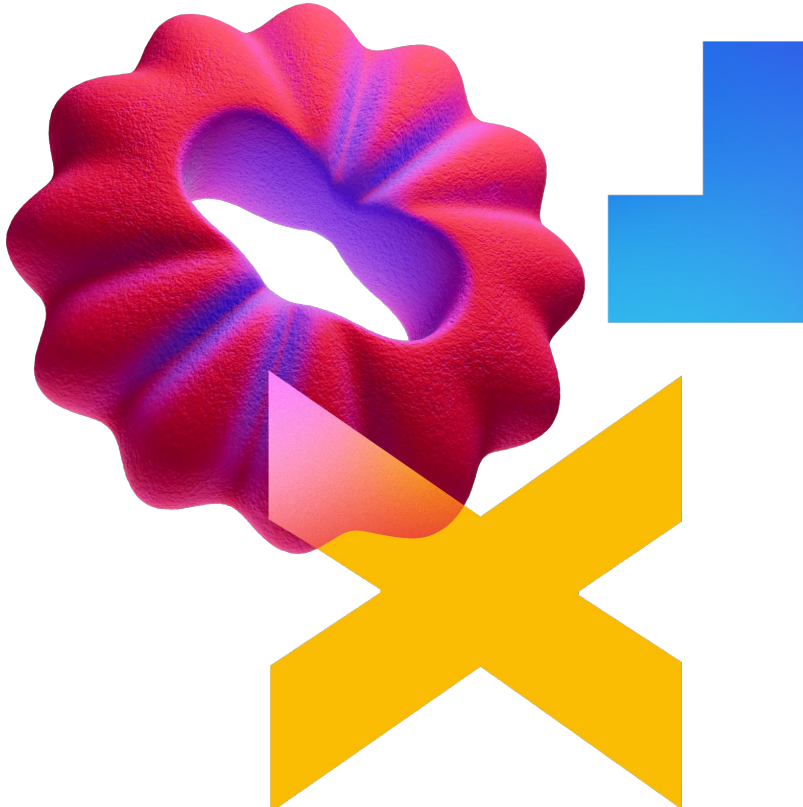
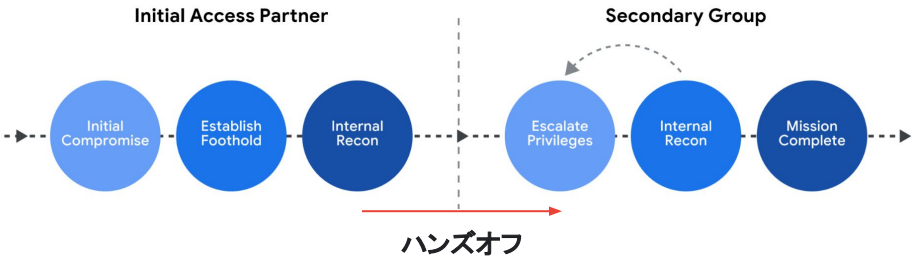


Table of contents

M-Trends 2026 より

攻撃者間のハンズオフの中央値

22 秒

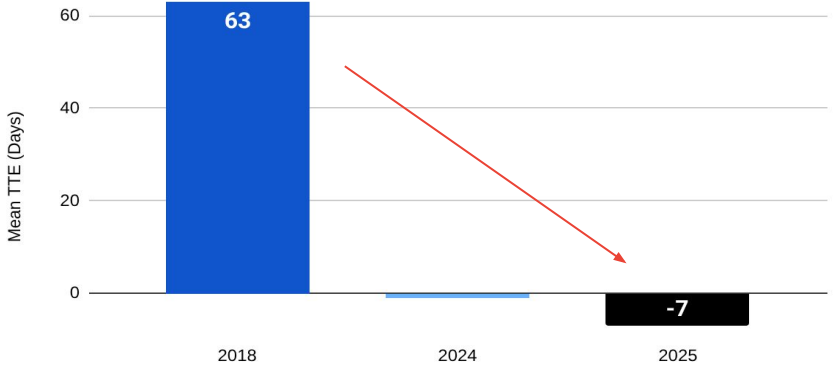


初期侵入の専門家

ランサムウェアなどの本攻撃

エクスプロイトまでの中央値

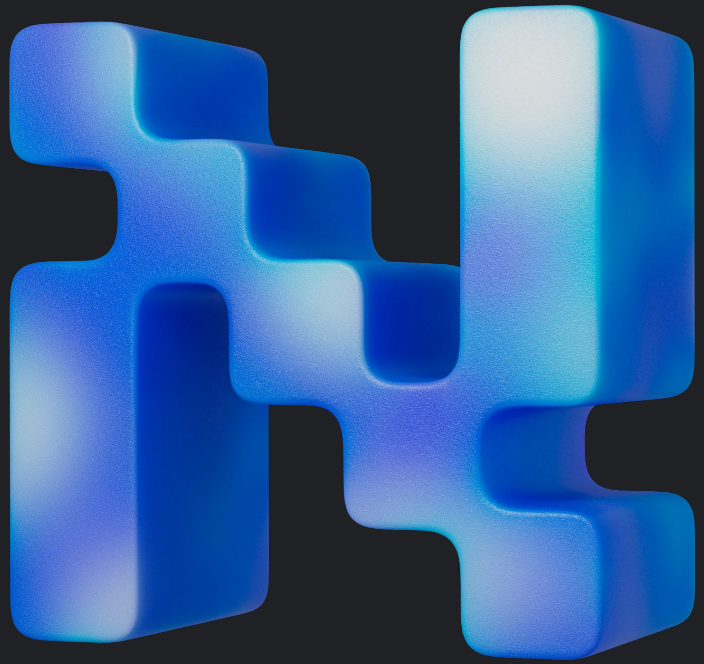
-7 日



パッチ提供よりも早く攻撃者が侵害を発生している

01.
AI 活用を進める サイバー攻撃者

BRK1-103 他



攻撃者による AI 悪用が急速に進化

生産性の向上

フィッシング メールの作成、およびコードのトラブル シューティングに AI を利用して生産性の向上を計りました

動的にタスクを変える

攻撃者は、ソーシャル エンジニアリングや初期アクセスといった動的なタスクを処理するため、攻撃チェーンに LLM の API を直接組み込み始めた

状況に応じた変容

攻撃者は LLM を統合して高度に自律的で適応能力の高い攻撃を行っている

AI を使った新しいマルウェア

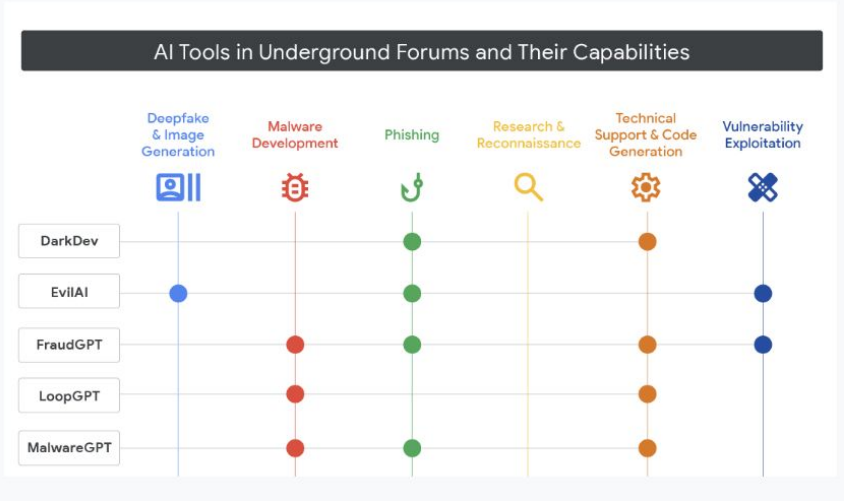
2025 年に Google Threat Intelligence Group は、新しい AI の機能を使うマルウェア ファミリーを発見しています

マルウェア	機能	状況
FRUITSHELL	Reverse shell	活動が確認されている
PROMPTFLUX	Dropper	実験的
PROMPTLOCK	Ransomware	実験的
PROMPTSTEAL	Data miner	活動が確認されている
QUIETVAULT	Credential stealer	活動が確認されている

AI を悪用したサイバー攻撃
AI / LLM を利用するマルウェアと、犯罪マーケットの成熟

- **動的なコマンド生成**
  環境情報（OS、セキュリティソフト等）をLLM に送信し、その場で攻撃スクリプトを生成
- **C2 インフラとしての　LLM 活用**
  固定サーバーの代わりに商用LLM サービスを指令塔として悪用特定やテイクダウンを困難に
- **「ジャストインタイム」の自己書き換え**
  実行時にコードを生成・変更し、従来のシグネチャベースの検知を無効化
- **犯罪マーケットの成熟**
  AI を悪用したツールを取り扱うアンダーグラウンドマーケットが成熟フィッシング、マルウェア開発、脆弱性調査向けに設計されたツールの販売を複数確認

```
Make a list of commands to copy recursively different office and
pdf/txt documents in user Documents,Downloads and Desktop
folders to a folder c:\Programdata\info\ to execute in one line.
Return only command, without markdown.
```



Sources: GTIG AI 脅威トラッカー: 脅威アクターによるAI ツール使用の進化
https://cloud.google.com/blog/ja/topics/threat-intelligence/threat-actor-usage-of-ai-tools
※ Google はこの行為者に対し、活動に関連する資産を無効にすることで措置を講じていますまた、悪用されにくくなるようモデルを継続的に改善しています

AI を悪用したサイバー攻撃
AXIOS サプライ チェーン攻撃 （UNC1069）

5 つの攻撃ステップ

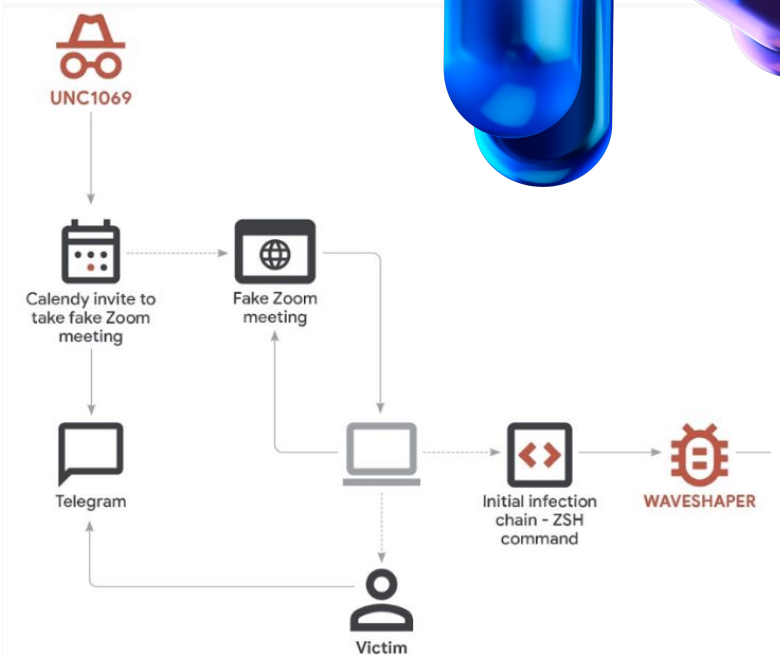
1 準備と偵察
Gemini 等を活用し、標的の調査とマルウェア開発を効率化

2 接触と信頼構築
Telegram で接触 Calendly 等の正規ツールを装い接近

3 罠の展開
ディープ フェイク動画で経営幹部を偽装し、Zoom 会議へ誘導

4 ClickFix 攻撃
音声トラブルを装い、解決策として標的に不正コマンドを実行させる

5 侵害とパッケージ汚染
WAVESHAPER 等 7 種類のマルウェアを展開し、認証情報を窃取
axios パッケージを汚染し、サプライチェーン攻撃に



Sources: North Korea-Nexus Threat Actor Compromises Widely Used Axios NPM Package in Supply Chain Attack
https://cloud.google.com/blog/topics/threat-intelligence/north-korea-threat-actor-targets-axios-npm-package
※ Google はこの行為者に対し、活動に関連する資産を無効にすることで措置を講じていますまた、悪用されにくくなるようモデルを継続的に改善しています

セキュリティ チームは 従来のリソースで AI スピードの攻撃を防がなければならない

より少ない労力でより多くの成果を

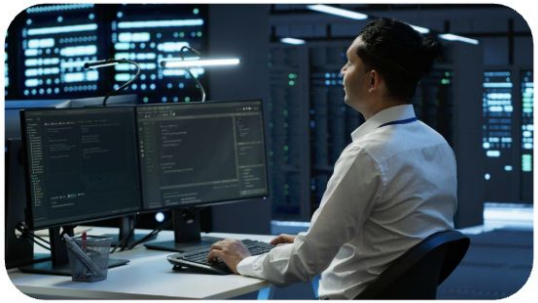
90%

現代の AI 脅威に対して
セキュリティ体制の不足を
感じている企業の割合 [1]

より少ない労力でより多くの成果を

4%

セキュリティ予算の平均成長率は
5 年間で最低を記録 [2]



現代のツールを使わずに保護する

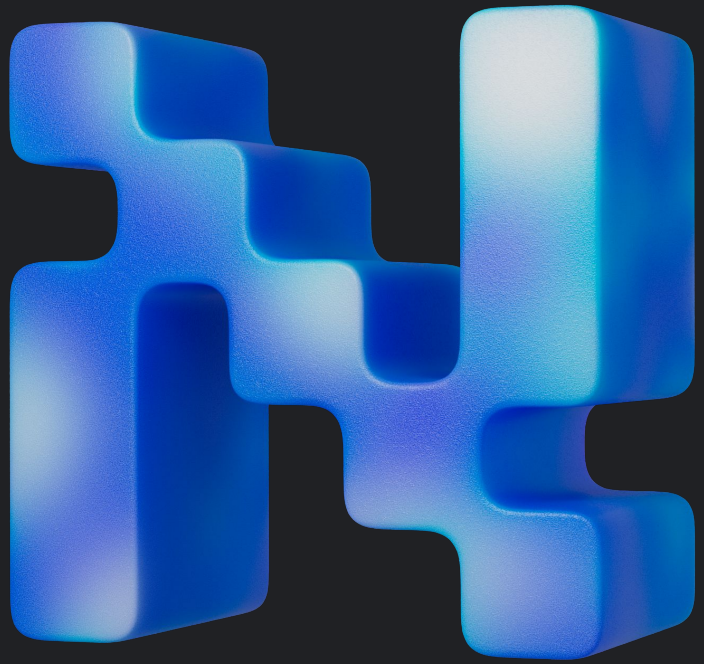
48%

外部関係者からの報告で
情報漏洩を知った組織の割合 [3]

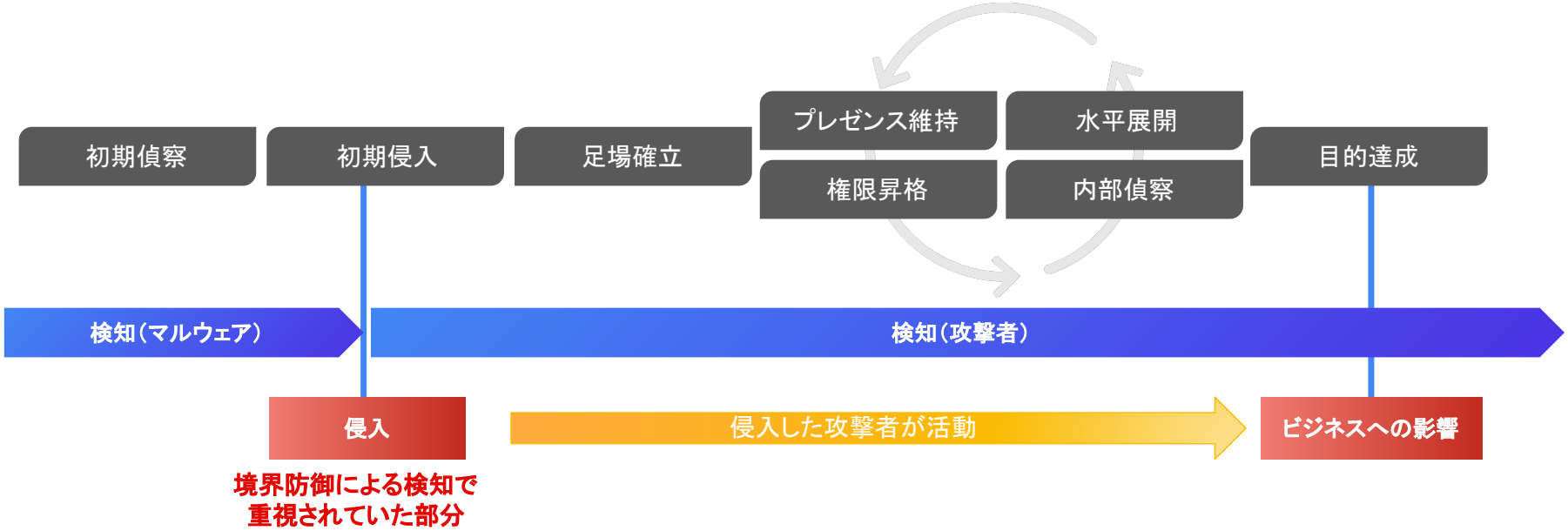
Sources: (1) Accenture (2) IANS (3) Mandiant M-Trends 2026

02.
AI エージェントを活用した SOC

BRK1-085



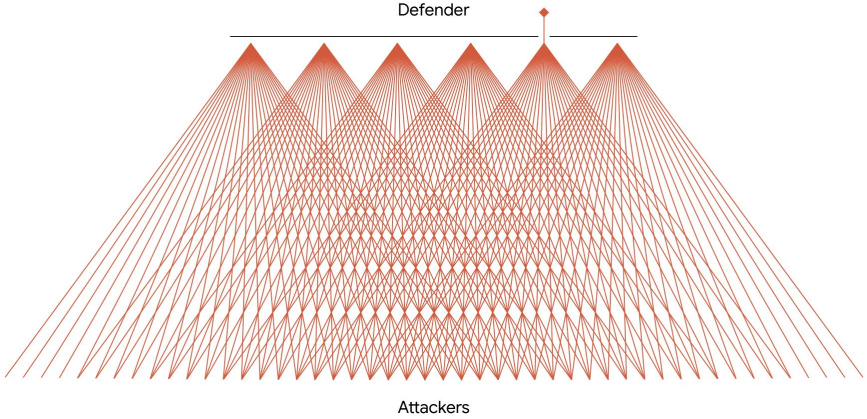
防御側の優位性 - “Defender’s Advantage”



1) 攻撃者が自由に
偵察できるステップ
（＝攻撃者有利 ）

2) 攻撃者が 1 つミスをすると
発見できるステップ
（＝防御者有利 ）

防御側の優位性 – “Defender’s Advantage”



“ 私たちは、AI が「防御側のジレンマ」を覆し、サイバー空間の天秤を傾けて、防御側が攻撃者側に対して決定的に優位な立場を得るための最良の機会を与えてくれると信じている “

エージェントの分析プロセス

1. **Mandiant のベスト プラクティス**

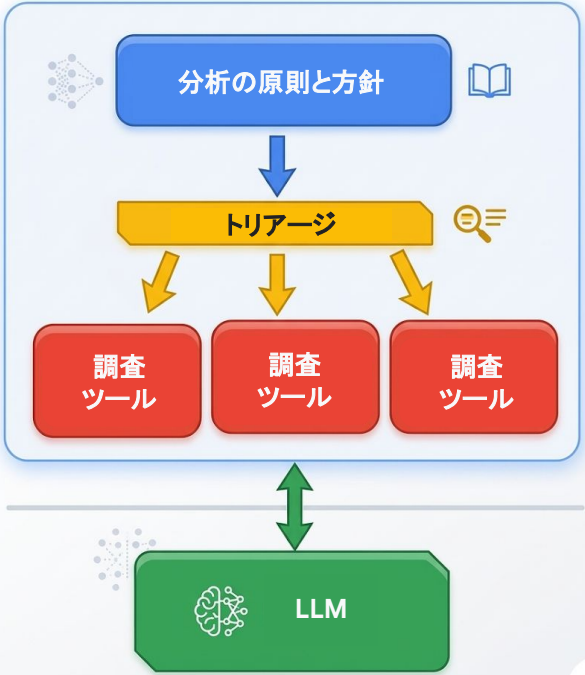
エージェントは分析を実行し、Mandiant が確立したベスト プラクティスに則って（事象の）評価・判定を行います

2. **SecOps がもつコンテキスト**

インシデントの影響範囲を把握するため、エージェントは事案のメタデータ、エンティティの詳細、および UDM イベントを含む関連するコンテキストを、SecOps 環境から自律的に収集・活用します

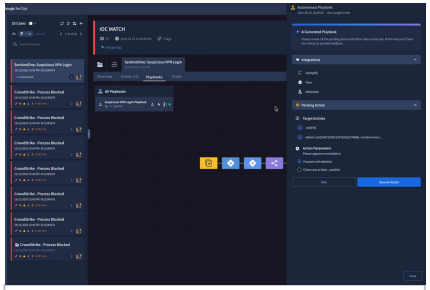
3. **Google Threat Intelligence**

エージェントは、ドメイン、URL、IP アドレス、ファイル ハッシュに関する脅威インテリジェンスを動的に取得し、調査プロセスの一環として分析します



Gemini in セキュリティ エージェント

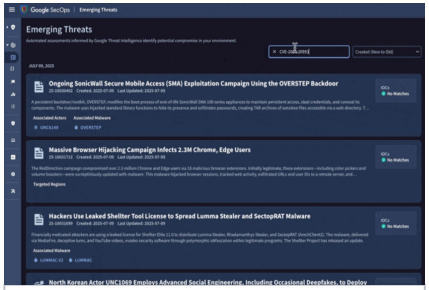
スキルが求められ、かつ人手のかかるサイバーセキュリティの運用を「セキュリティ用の Gemini 」が支援



調査 &トリアージ

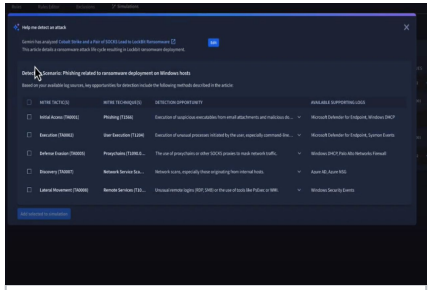
アラートの背後にある一連の攻撃を調査、概要を把握

対応策の提案と、すぐに対策が必要なアラートを選別して優先度付け



脅威ハンティング

新たな脅威を継続的にハンティングし、侵害評価やレポートを提示



検知エンジニアリング

脅威レポートから対策状況を評価

ギャップを特定し、検知ルール作成

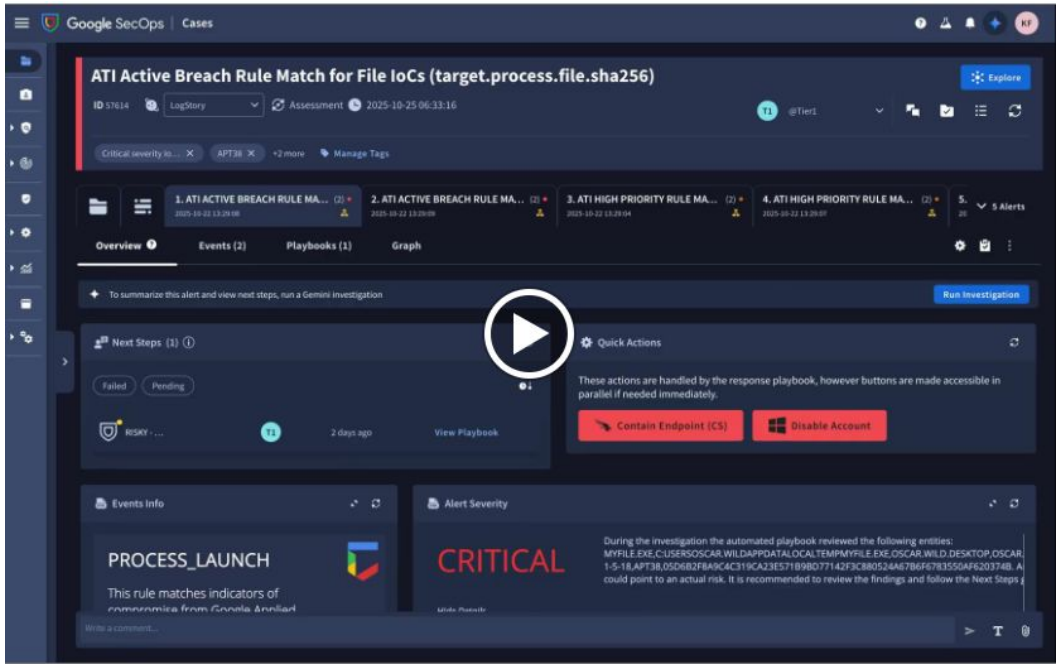
.....

Gemini in セキュリティ エージェント

スキルが求められ、かつ人手のかかるサイバーセキュリティの運用を「セキュリティ用の Gemini 」が支援

調査 & トリアージ エージェント

- アラートの初期調査を行い、真陽性・偽陽性を判定
- アラートの背後にある一連の攻撃を調査、概要を解説
- 次のステップ（追加調査、対応策）の提案
- 数が多いアラートから、すぐに対策が必要なものを選別して優先度付け



Gemini in セキュリティ エージェント

スキルが求められ、かつ人手のかかるサイバーセキュリティの運用を「セキュリティ用の Gemini 」が支援

脅威ハント エージェント

- 特定攻撃者グループによる侵害の兆候がないかを、取り込んだログ情報に対して調査
- 攻撃キャンペーンの対象となった可能性がある場合に有効（例：所有する VPN 装置のゼロデイ脆弱性が公開 APTxx による攻撃キャンペーンが 3ヵ月前から観測されている）
- 侵害の兆候がある場合、追加調査をボタン 1 つで実施可能（攻撃をうけた端末やユーザーの特定、具体的なマルウェアやその通信先の調査など）

Gemini が調査して数秒で応答
・この攻撃者に関連するIOC (IPアドレスやファイルなど)が、過去 3 カ月間のログに12 件 確認
・侵害が疑われる社内PC やユーザ名を表示(jorge, Tim Smith)
・次の詳細調査のためのアドバイス

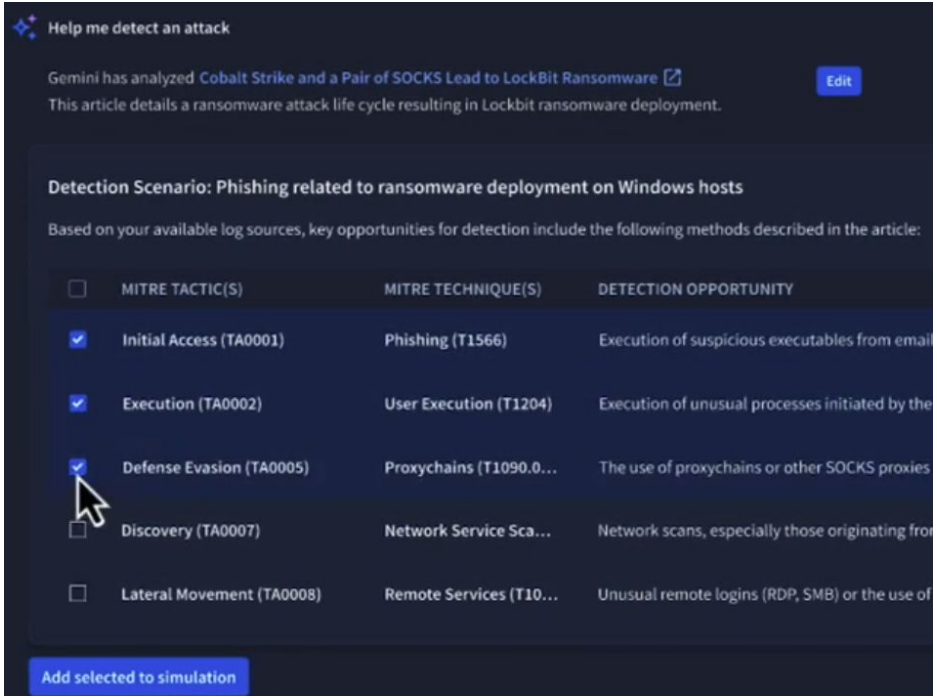
Gemini in セキュリティ エージェント

スキルが求められ、かつ人手のかかるサイバーセキュリティの運用を「セキュリティ用の Gemini 」が支援

検知エンジニアリング　エージェント

脅威インテリジェンス レポートや
ブログから、検知能力を評価、改善を実施

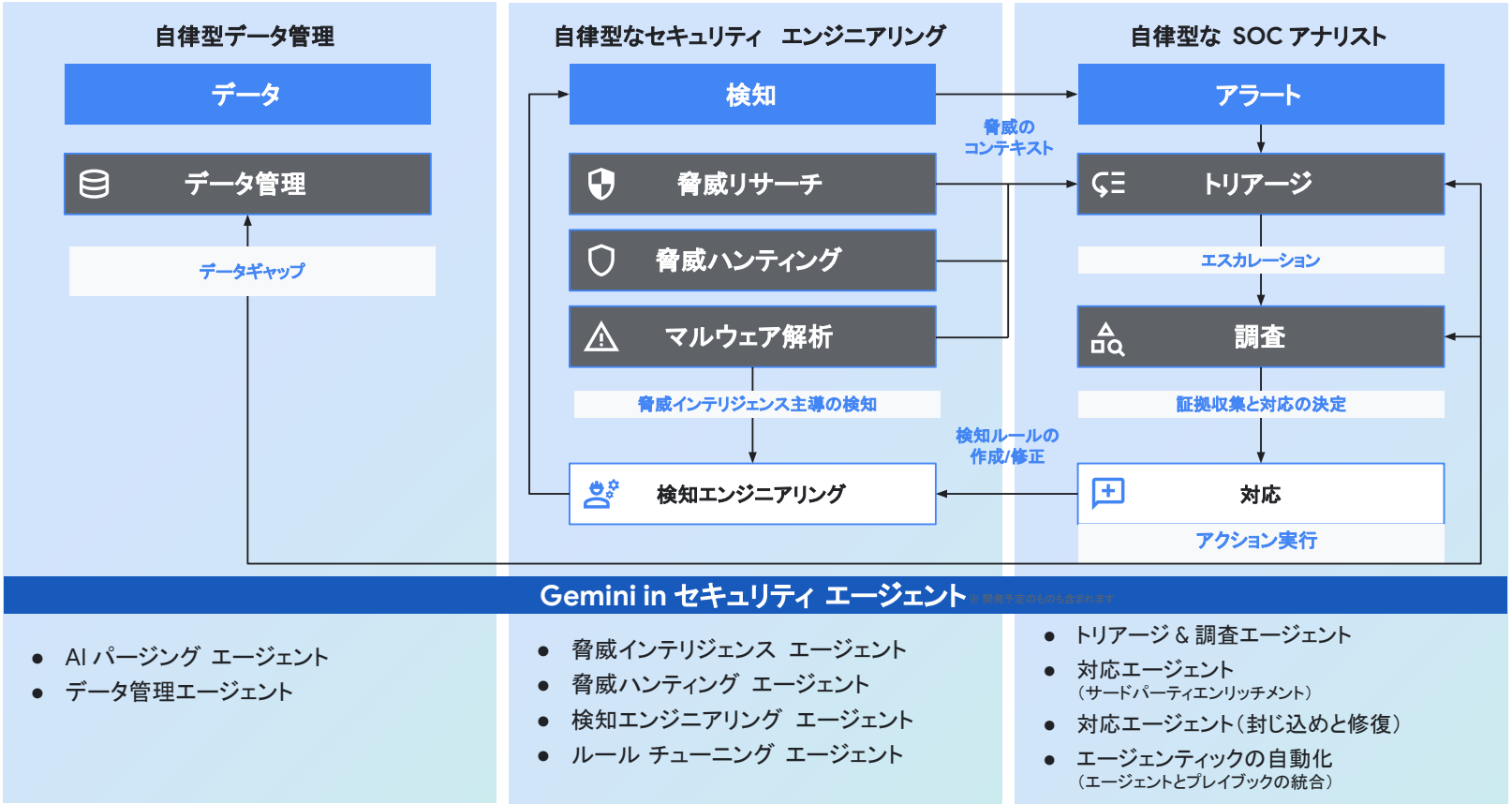
- レポートの内容を攻撃シナリオに分解し、シミュレーション用のイベントログをシナリオごとに生成
- 生成したイベントログをもとに、既存の検知ルールの有効性を評価
- 検知されなかったシナリオについて、検知ルールを自動生成実際のログデータに基づいてルールを最適化しノイズを低減



※画面は開発中(Preview) のものです実際の提供機能とは異なる場合がございます

AI による自律的な “Agentic SOC”

各フェーズでセキュリティ エージェントが連携する SOC 基盤



AI エージェントによって
支援型から Agentic = 自律型 SOC へ

現状:
従来の支援型 SOC

- 手動でのトリアージ、調査、ハンティング、チューニングなど
- バラバラなツールとプロセス
- ほとんどの意思決定に人間が介在
- 限定的な AI 支援

目指すべき姿:
自律型の Agentic SOC

- シームレスに統合された AI エージェント
- 検知から修復までを AI で自動化
- 人間に代わって AI がセキュリティ運用を牽引
- セキュリティ用の AI (Gemini in セキュリティ エージェント) の活用

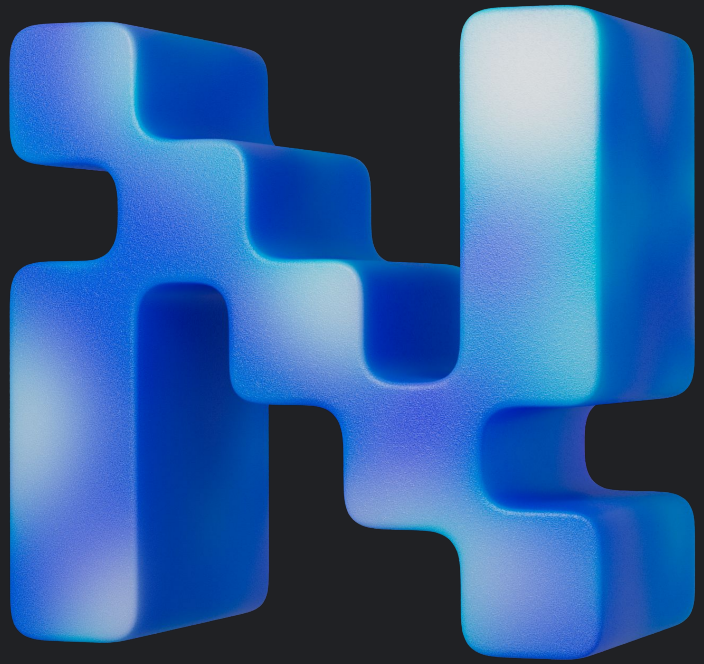
→ Agentic SOC はすでに実践段階支援型から自律型　SOC のフェーズに

03. お客様事例

Agentic SOC at Allianz

BRK1-086

Lars König - Global Head of Detection & Response
Alex Pabst - Global Deputy CISO

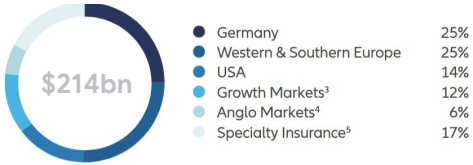


Allianz は欧州で最大の金融サービスを提供

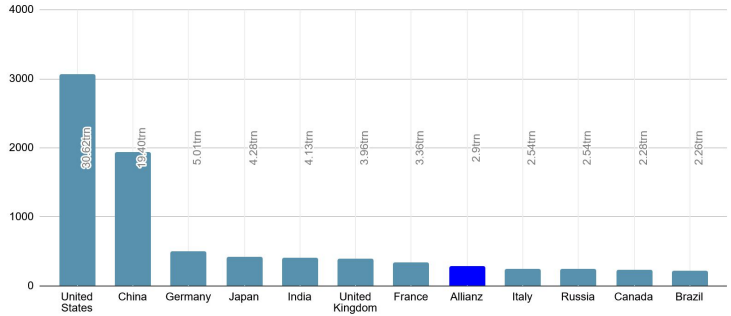
Allianz Group 全体の総事業規模 (2025): 2,140億

セグメント別内訳

国別内訳



資産運用残高(AUM): イタリア、ロシア、カナダ、ブラジルのGDP（国内総生産）を上回る規模



事業規模

1.25 億人

顧客数:72 ヶ国に
1 億 2,500 万人

組織

15.7 万人

従業員数:15 万 7,000 人
事業部門:56 拠点

IT インフラ

35 万台

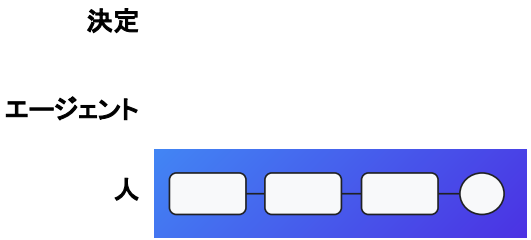
システム数:30 万
エンドポイント数:35 万
仮想デスクトップ:約10 万台

Allianz が定義する SOC 自動化の 5 段階レベル

	Level 0 手動 	Level 1 アナリスト支援	Level 2 部分的な自動化	Level 3 条件付き自動化	Level 4 高度な自動化 	Level 5 完全な自動化 
人間の役割	人間がプロセスを主導し、結果や入力を定義し、責任を負う （機能による支援があっても同様）			自動化が実行される際は、プロセスと結果を主導する （人間が関与する場合でも）		
	自動化機能を常に監視し、管理、トリアージ、調査、 対応を行う必要がある			自動化が要求した場合にのみ、人間は対応する必要がある	自動化は人間の介入を必要としない	
	アナリスト支援機能			SOC の自動化機能		
機能の役割	機能はアラートや瞬間的・単一のアクションに限定される	人間がトリガーする、定義済みで既知のユースケースに対して限定的な自動化を提供する	定義済みで既知のユースケースに対して、一部の自動レスポンスを含む完全な自動化を行う	既知のケースは自動化、未知のケースは自動トリアージを行い、人間にエスカレーションする	既知およびほとんどの未知のアラートを、修正措置を含めて自動化する人間には通知のみ行う	エンドツーエンドの完全な自動化
機能例	SIEM ルール デバイス上での自動ブロック（例：AV）	SOAR プレイブック（手動実行） EDR によるデバイス ブロックとアラート	一部のアラートに対する、完全に自動化された決定論的な SOAR プレイブック、および EDR によるブロックとアラート	エージェントによるトリアージ レスポンス システム	エージェントによる調査とハンティング、自動化された防御、自動化されたレポート作成	レベル 4と同様（ただし、すべてのケースが対象）
人間の関与レベル	すべてのケースを人間が処理する	すべてのケースに人間が関与する	一部のケースに人間が関与する	人間は限定的・単一のアクションのみ行う	自動化が失敗した場合の、フォールバック（代替手段）としてのみ人間が関与する	すべてのケースで人間の関与はない

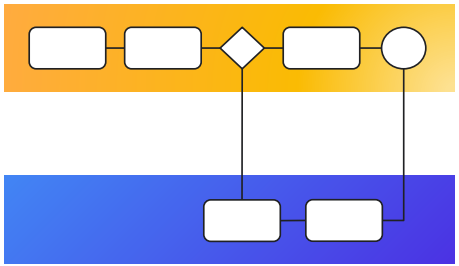
主な自動化は決定の部分
ただし必要に応じてエージェントや人が支援

人の支援



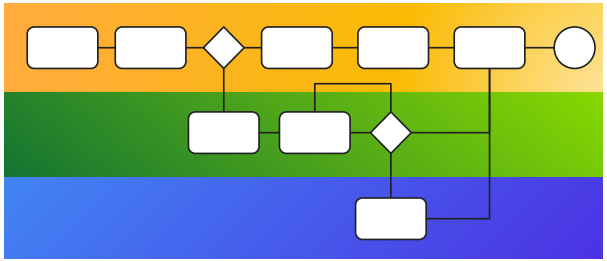
- インシデント対応（IR）プロセスはWiki（社内ナレッジベース）に定義され、手動で実行される
- 日次運用をサポートする自動化は限定的、あるいは皆無である

自動化の支援



- SOAR（セキュリティ運用自動化）の導入
- Wiki ベースのプロセスをSOAR のプレイブックへ移行
- プレイブックは人間の判断・操作を主導として進行する
- 時間の経過とともに、段階的に自動化の範囲を拡大

エージェントにフォーカス



- 決定論的（ルールベース）処理と人間の間に、自律型エージェントのレイヤーを追加
- 「人間がプレイブックを動かす」運用から**「プレイブック（システム）が必要な時だけ人間に判断を仰ぐ」**運用へのパラダイムシフト
- **スケール（拡張性）、スピード、そして説明可能性を担保**するため、引き続き決定論的（ルールベース）なアプローチを中核として強く意識する

現在、多くの仕組みは整いつつあるが、エビデンス収集においてはさらなる自動化が求められている

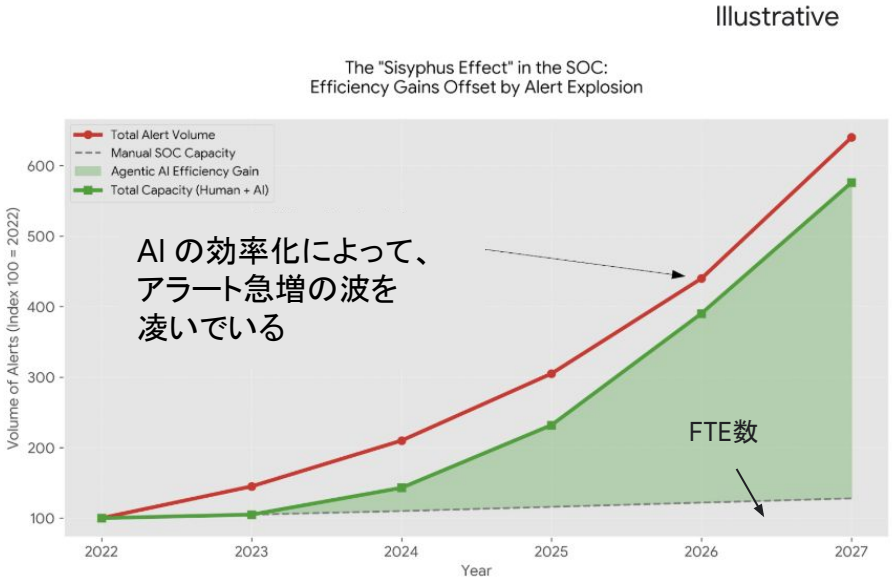
マルチレベルでの AI 活用

- エージェント型トリアージを活用し、インシデントの初期段階での優先順位付けを実施する
- Gemini を活用し、ケースの要約や、多言語での従業員向けコミュニケーションを行う

データに関する課題

- エージェントをより活用するために、詳細で全体的な調査を行うためには、追加のデータポイントが必要となる
- エージェント機能（AI）が自律的に機能するための、データやシステムへのアクセス（権限・経路）管理が必要となる

チームと能力はエージェント型　SOC の移行プロセスに適合しないといけない



- 今後 2 ~ 3 年で、大幅な人員削減が行われることは想定されていない
- アラートが増え続ける中で、**同等の社員でより多くの事象への対応**を実現する
- さらなる拡張（スケール）には、**予防的かつ自己修復型の自動化**が不可欠
- 属人的な「孤高のヒーロー」から、**AI の管理者・コラボレーター**への役割転換
- 調査・分析のマインドセットを持ち、目先の事象ではなく「根本原因」を解決する

説明責任や規制順守は変わらない
～引き続き監査証跡は必要～

- 取締役会は引き続き、規制当局に対する説明責任を負う
- CISO（最高情報セキュリティ責任者）も引き続き、最終的な結果に対する責任を負う
- Agentic AI が SOC のあらゆる問題を解決するわけではないデータ品質が低ければ、結果はかえって悪化する

- 取締役会をこの変革プロセスに巻き込み、メリットとデメリットを透明性をもって説明する
- エージェントの「思考プロセス」と実行された意思決定の監査証跡を確実に残す
- ソリューションが定期的にテストおよび検証されている証拠を提示する（例：過去インシデントのサンプルテストなど）

最後に:

Google Agentic Defenseの差別化要因

独自の能力で「エージェント型　SOC」のビジョンを推進

AI スタック全体の自社開発

独自開発のシリコン（ TPU）から先進的な Gemini モデルまで、すべてを最適化

- 他に類を見ないパフォーマンス、コスト
- 効率の高い拡張性（スケール）

グローバルな可視性とデータ

Googleエコシステム全体で、日々数十億件のシグナルを分析

- ハイパースケール・データ プラットフォーム
- VirusTotal およびセーフ ブラウジングの知見

比類なき専門知識

Mandiant の最前線のインテリジェンスと、体系化されたアナリストの「思考マップ」を活用

- 現実世界の脅威（ TTPs：戦術・技術・手順）
- 70% 以上のトリアージ（一次切り分け）精度

包括的なエージェント型　SOCのビジョン

- データの取り込みから対応まで、SOC ワークフロー全体をエンドツーエンドで自動化
- マルチ エージェントによる連携・予防的かつマシンスピードでの防御

オープンかつマルチクラウド対応の設計

- ベンダー ロックインを防ぎ、既存のセキュリティ　エコシステムとシームレスに統合
- SecOps（セキュリティ運用）をより自律的、効率的、かつ高度な脅威に対して効果的にします

創出される成果

70% 削減

侵害のリスクとコスト

運用の高速化

76% 向上 クエリ作成速度

500 万件以上 アラート分析調査済み

人材のスキルアップ

50% 削減 平均対応時間（MTTR）

65% 削減 平均調査時間（MTTI）

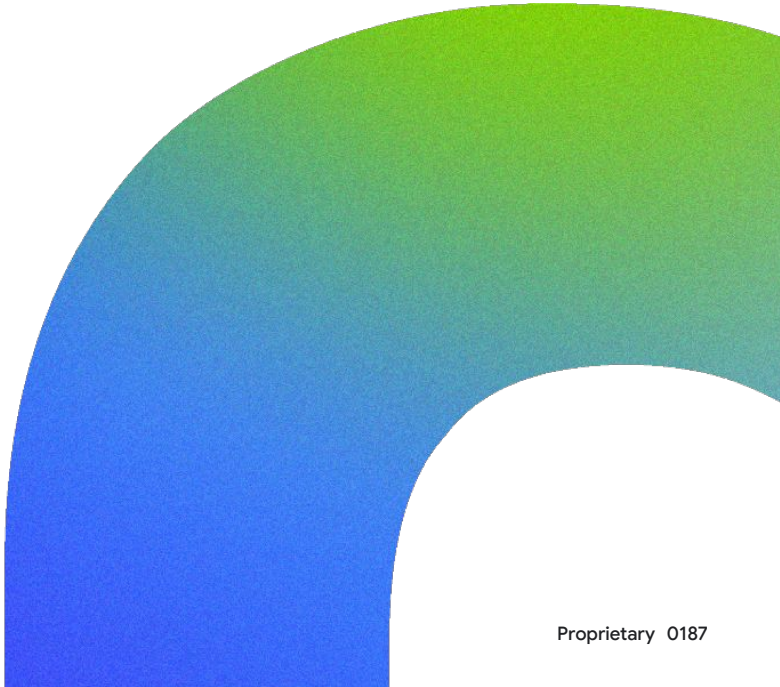
Google Cloud Next 26

Agentic Collaboration

(Google Workspace)

エージェントにより「タスク」から「成果」の達成へ

- **Context**：背景を理解する
- **Reasoning**：実行方法を計画する
- **Orchestration**：ツールで実行する



コンテキストの壁 という課題

不十分なコンテキスト

エージェントはデータには接続されているが、 ビジネス全体のコンテキスト（背景）を
完全には理解できていない

人による情報の橋渡し

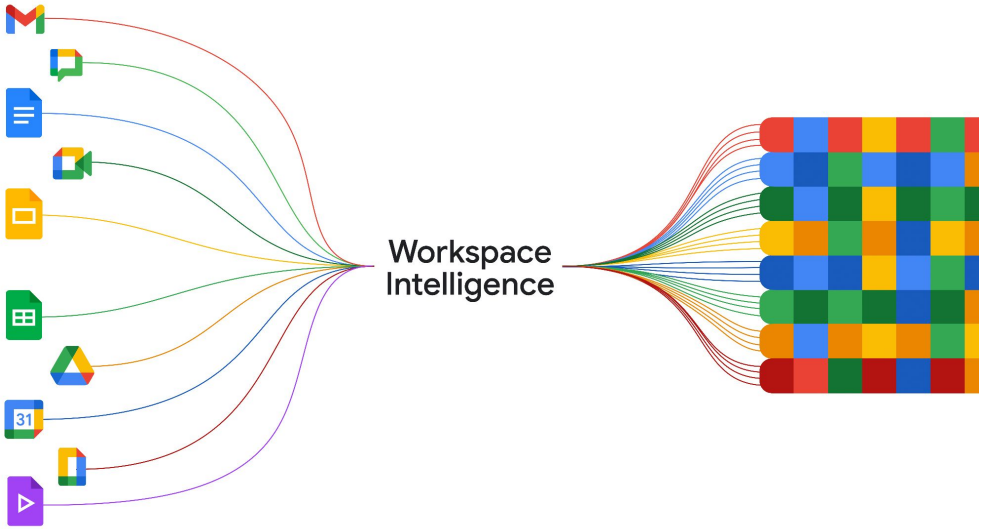
情報の断片を人間が手作業でつなぎ合わせているため、
ビジネスのスピードが制限されている

Workspace Intelligence

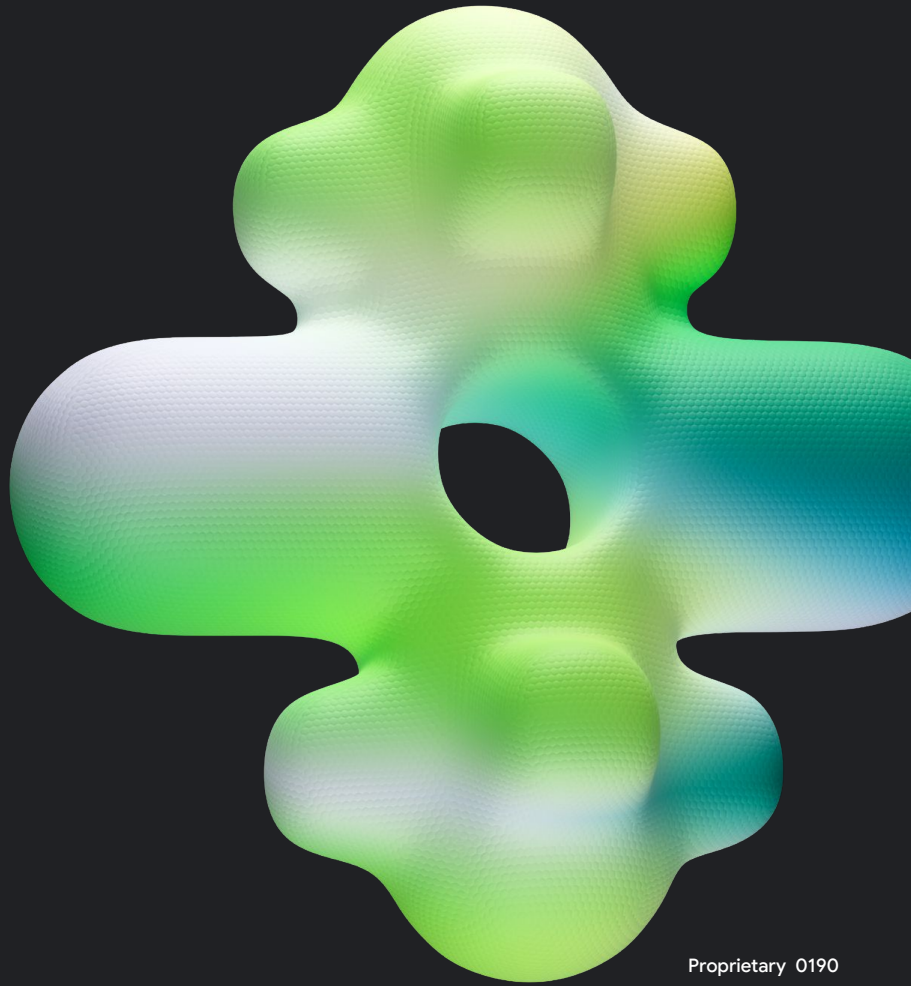
一般提供開始

Workspace Intelligence が業務コンテキストを統合し、Gemini がプロアクティブな パートナーに進化

- メール、チャット、ファイル、イベントを統合し、深い業務コンテキストを把握
- Workspace 上で単純なタスクから複雑なワークフローまでを自動化



各アプリの新機能



Google Workspace is
making work easier.

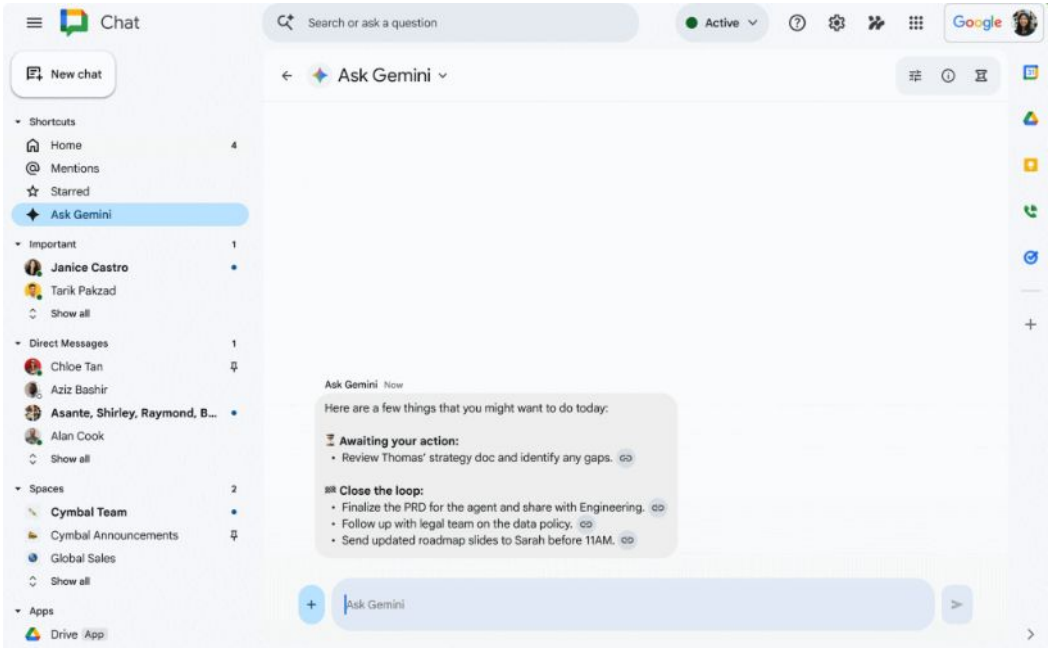
Ask Gemini in Chat

Workspace Intelligence

プレビュー

あらゆる業務を統合する、一元的なコマンドライン

- プロアクティブな「今日やることリスト」で1日をスタート
- メール、メッセージ、ファイルを迅速に検索
- 洗練されたドキュメントやスライドを一つの場所で作成



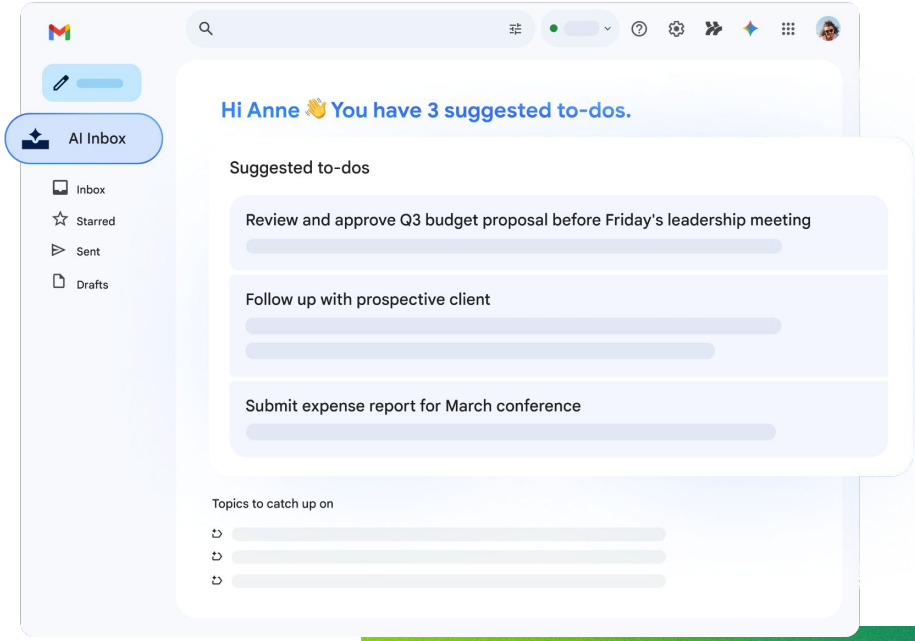
Gmail の AI Inbox と AI Overview

Workspace Intelligence

プレビュー

受信トレイを整理し、
必要な情報を迅速に見つけ出す

- 優先順位付けされた To-Do リストで業務を効率化
- キーワードではなく自然言語で検索
- Gmail スレッドや検索結果の要約により、状況を素早く把握



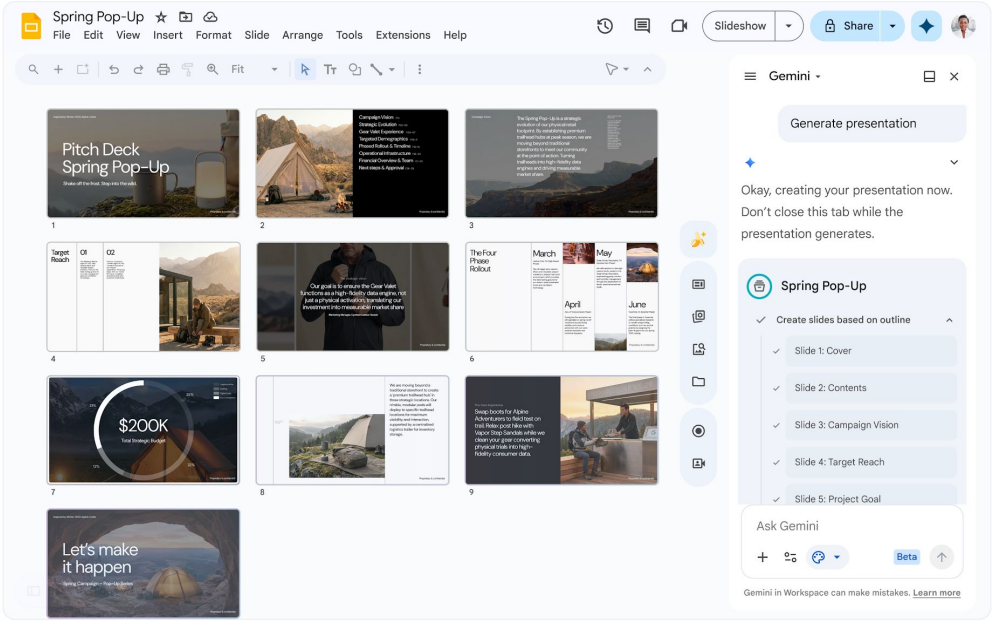
Workspace Intelligence

プレビュー

複数枚のスライド作成

手作業なしで、アイデアを魅力的なプレゼンテーションに変換

- アウトラインに沿ってストーリーを構築・編集
- レイアウト、テキスト、画像を含むプロ品質のスライドを、一発で作成
- 編集可能なスライドを人の手で改善
- 既存のブランドテンプレートからも作成可能

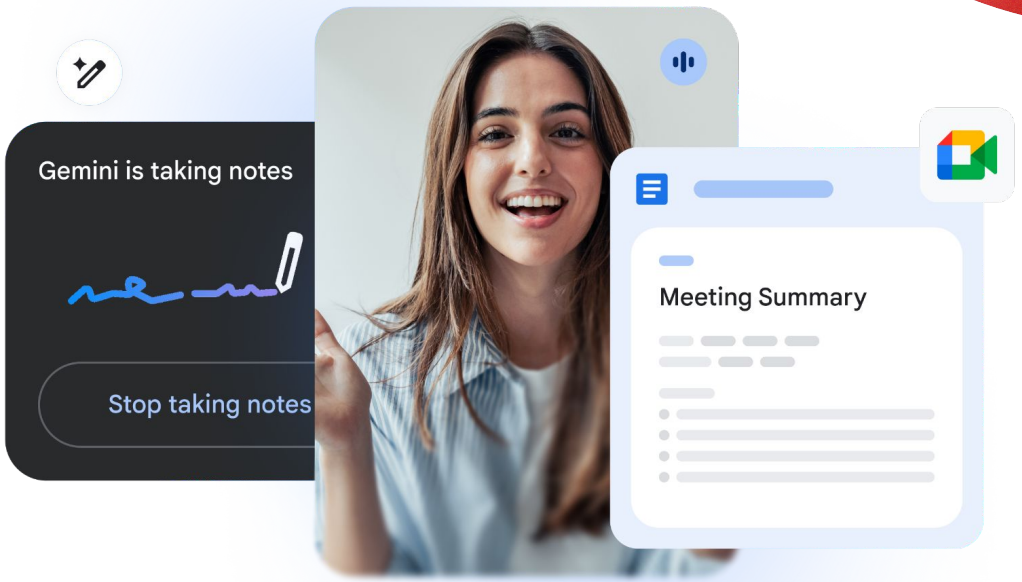


Gemini 議事録をあらゆる会議で

プレビュー

対面または他のプラットフォームでの会議でも、Gemini が議事録を作成

- 議事録作成ではなく、会話に集中
- あらゆる会話に AI のサポートを
- アクション アイテムを記録し、タスクを確実に実行

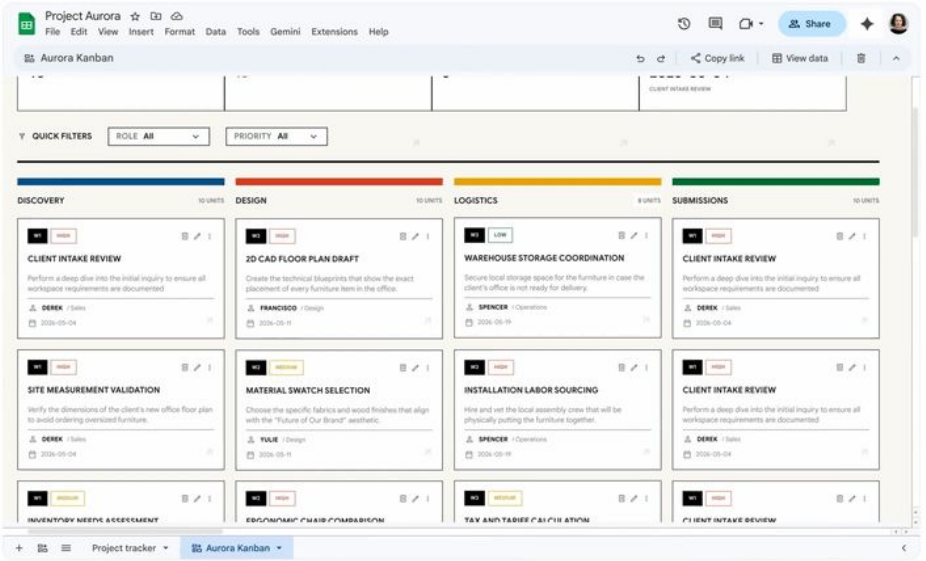


Sheets Canvas

プレビュー

コーディングなしで、シート上にカスタムのインタラクティブなアプリを構築

- ダッシュボード、ヒートマップ、かんばんボードなどのミニアプリを作成
- 生成したアプリは、他のシートと同様に共有可能
- データは常に同期されており、リアルタイムに更新



Sequences shortened. Results may vary. Check responses for accuracy. Availability in select countries and languages.

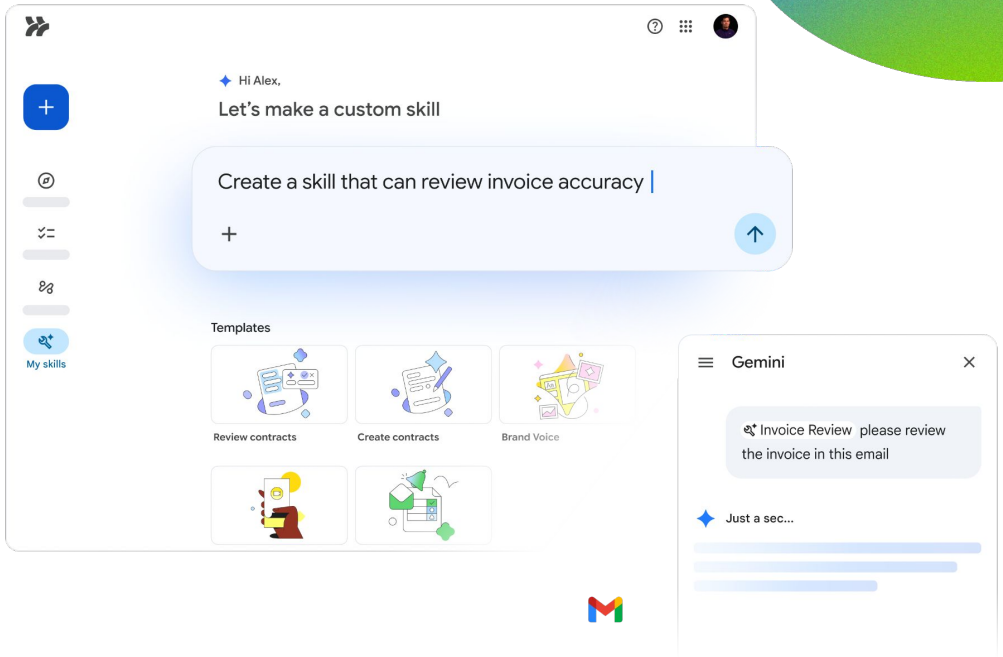
Workspace の「スキル」機能

Workspace Intelligence

プレビュー

定型業務を「組織の頭脳」として自動化・効率化

- よくあるタスクを自動化し、共有可能な「スキル」として登録
- Google ドキュメント上でチームや Gemini と共同でスキルを作成
- Workspace 内のあらゆる Gemini インターフェースからスキルを利用可能



データ インポートによる GWS への移行

追加コストなし

移行の計画

データ量や移行にかかる時間を確認

包括的なソリューション

MS Teams、暗号化コンテンツ、
Outlook ルール、カレンダーと
メタデータ、MIP ラベルなど、
より多くのデータ ソースをサポート

スピード

並列処理とアルゴリズムの改善により
移行を高速化し、データのインポートを
より短時間で完了

使いやすさ

管理コンソールに組み込まれた
スケーラブルなソリューション

Thank you

Google Cloud Next 26