Transformation through Sovereign AI Cloud

Pankaj Shukla
Head of Customer Engineering
Public Sector Google Cloud India

Future of AI Assistance


Golden Stripes

Changing the photography landscape


Prompt: elephant amigurami walking in savanna, professional photograph, blurry background

The science behind the art

2015

Google DeepMind AlphaGo defeats Go champion

2016

Google's DeepMind helps detect eye disease

2017

Google invents Transformer kickstarting LLM revolution

2018

Google's groundbreaking large language model, BERT

T5

2019

Text-to-Text Transfer Transformer LLM 10B P Model Open Sourced

2020

Google LaMDA Model Trained to converse

2022

AlphaFold predicts structures of all known proteins

Bard

2023

A conversational AI Service powered by LaMDA.

3,000 Researchers
7,000 Publications

Responsible AI

Built & Tested for Safety

Privacy in design

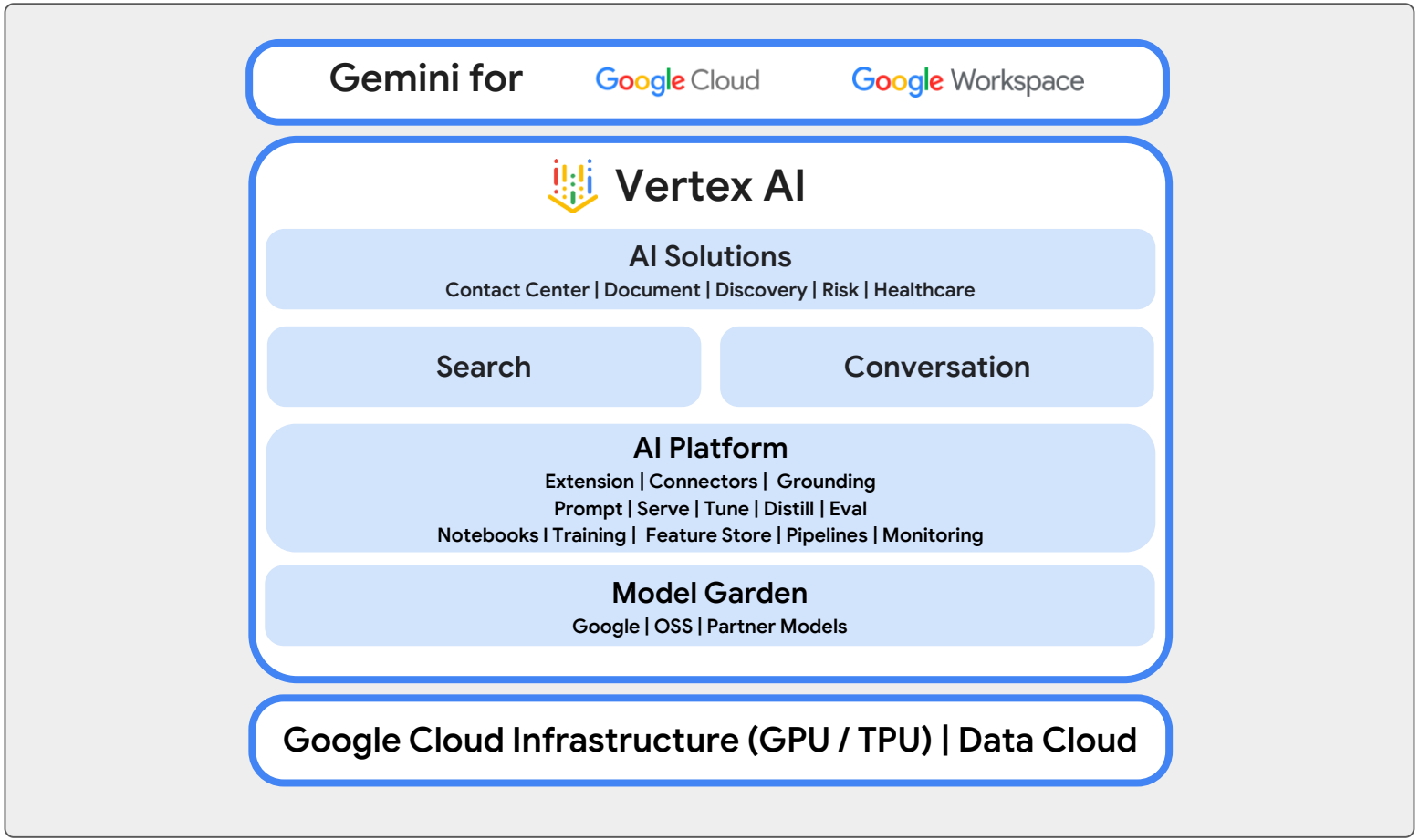
Upholds high scientific standards

Accountable to People

Socially Beneficial

Avoid creating unfair bias

Google Cloud


Gemini for
Google Cloud
Google Workspace
Vertex AI
AI Solutions
Contact Center | Document | Discovery | Risk | Healthcare
Search
Conversation
AI Platform
Extension | Connectors | Grounding
Prompt | Serve | Tune | Distill | Eval
Notebooks | Training | Feature Store | Pipelines | Monitoring
Model Garden
Google | OSS | Partner Models
Google Cloud Infrastructure (GPU / TPU) | Data Cloud

Google Cloud brings the best of Google

Google Workspace AppSheet

Productivity & Work Transformation

Insights Platform/ Smart Analytics -as-a-Service	Data Management -as-a-Service	API Management -as-a-Service	Application Platform -as-a-Service

BigQuery · AI Platform · Contact Center AI · Cloud Data Fusion · Cloud SQL · Cloud Spanner · Apigee Sense · Apigee API Platform · Anthos

Infrastructure -as-a-Service

Compute Engine · Google Kubernetes Engine · App Engine · Cloud Storage · Cloud Network · Virtual Private Cloud · Cloud VPN · Google Cloud Armor · Security Command Center · IAM · Application Migration

Public Cloud · Hybrid Cloud · Multi Cloud · Edge Cloud

Why not just use Public Cloud

Regulation

When information is highly sensitive or regulated, and air-gapped sovereignty

Survivability

When the processing needs to continue even if the connection is down

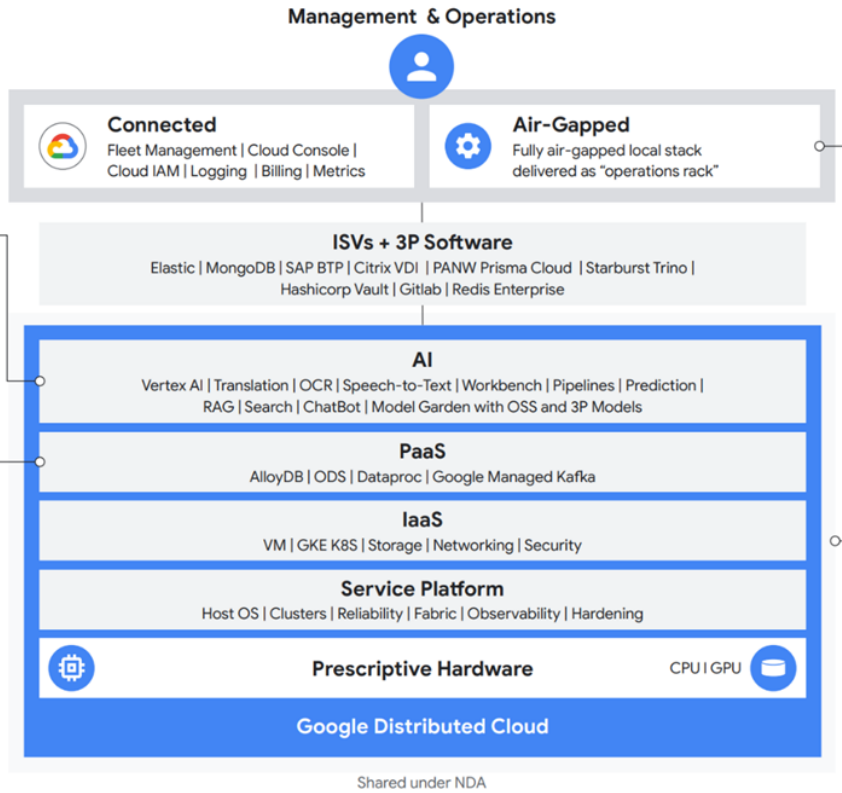
Local data processing

When data is high-volume,noisy, expensive or time consuming to send elsewhere

Latency and speed

When decision time matters, windows of opportunity are small, interactions require locality

Google Distributed Cloud Technology Stack



Built for AI Anywhere

Delivering Google's differentiated AI capabilities on-prem

Pre-trained models to build AI-powered applications without ML expertise

Build Gen AI powered agents with our Vertex Platform capabilities (Serving, RAG, Grounding, NLP, Fine-tuning)

Data and Analytics

Open-source compatible database for relational database workloads (built-in LangChain and Vector AI extensions)

Data lake modernization, ETL and secure data science, at scale - on prem

Data streaming service with fully-managed sink / source connectors and SQL-based stream processing

Air-Gapped Operations

Complete local operations suite delivered by Google or Partners

Includes support ticketing, runbooks, deployment and IaC tooling

SecOps suite includes SiEM, WebScan and Vuln Scanning tools

Zero Trust End-to-End

Built-in Zero Trust controls and verification with a zero trust approach

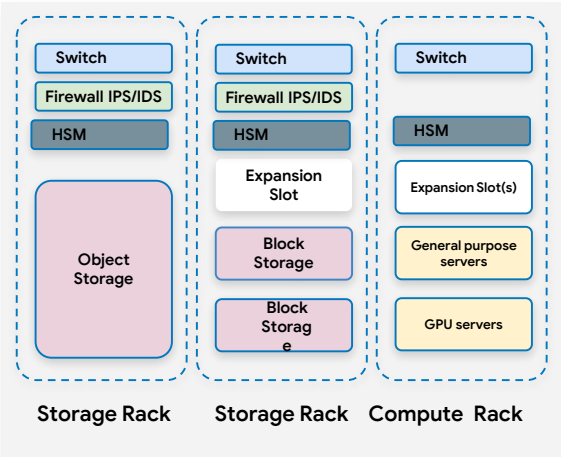
Data Protection: In-transit, At-rest and in-use (CMEK, LUKS)

Comprehensive application protection against threats (DLP, EDR)

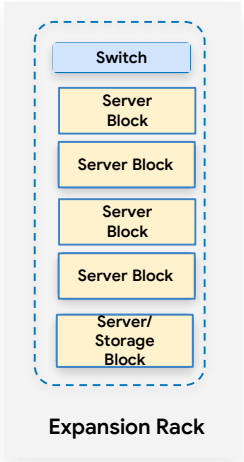
GDCH Configuration Overview - Base and Expansion Racks

Built for Scale and Modularity

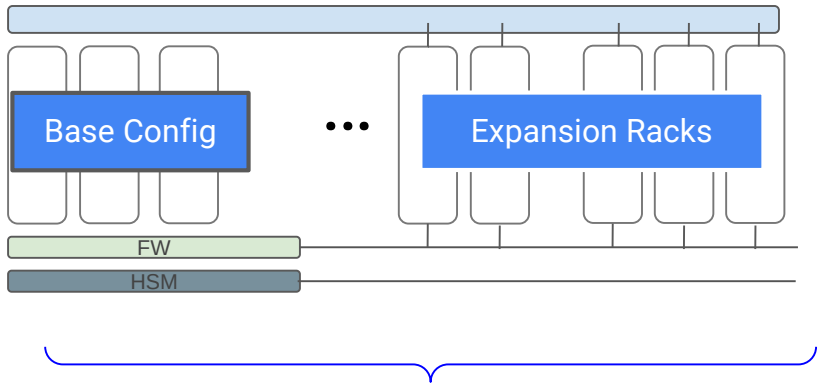
3 Rack Base Config



Expansions



GDCH Deployment Zone



- GDCH starts with a base configuration of 3 racks containing a mix of compute and storage resource

- Customer can select different options for expansion
- Maximize available rack space

Up to 30 racks in a single Zone

GDC air-gapped appliance

16.87” H x 14.69” W x 27.13” L
(428 mm H x 373 mm W x 689 mm L)
~100 lbs (~45.3 kg)

Containerized AI/ML models can be deployed on Kubernetes
Vertex Pre-trained AI models

- Optical character recognition
- Translation
- Speech-to-text

Core GDC air-gapped Services: Kubernetes, VMs, networking, Block & Object Storage.

3 ruggedized server nodes (temperature, altitude, humidity)

Xeon processors, SSD storage, Nvidia A100 GPU

Offline AI Enabled Signal Analysis

Mission problem:

Too much information to analyze efficiently and thoroughly

- Ubiquitous sensor / signal data is widely used for tasks like traffic monitoring or public safety.
- Privacy concerns prevent public cloud analysis of sensitive government data.
- Manual analysis is time-consuming and doesn't scale effectively.

Solution:

AI Enabled Signal Analysis

- **Rapidly analyze footage for actionable insights**
- **Private AI/ML Ecosystem for fine-tuning**
- **Comprehensive Situational Awareness**

AFA News: Air Force Developing ‘Cloud in a Box’ Device for Aircraft Sustainment

9/16/2024

By Josh Luckenbaugh

Air Force photo

NATIONAL HARBOR, Maryland — The Air Force’s Rapid Sustainment Office has teamed with Google to build a device that will allow forward operators to access important maintenance data even if they are disconnected from the Air Force network.

“Google Distributed Cloud air-gapped appliance will enable the Air Force Rapid Sustainment Office (RSO) to bring the maintenance digital ecosystem to Airmen in austere and forward deployed locations, supporting the Air Force's agile objectives while prioritizing security and reliability”

Col. Nathan Stuckey
Military deputy program executive officer

The Air Force’s Rapid Sustainment Office

U.S. AIR FORCE

Offline AI Enabled Translation

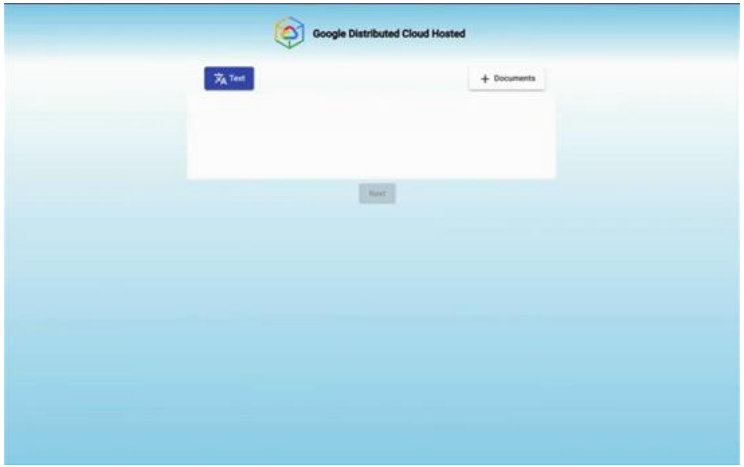
Mission problem:
Overwhelmed by Sensitive Foreign Language Data

- Vast hard copy archives strain timely translation by human linguists.
- Sensitive content prohibits public cloud translation solutions.
- Current offline tools lack the accuracy of those available publicly.

Solution:

High Quality Translation On-prem

- **Translation of text, speech and images while disconnected**
- **Augment existing workflows with intelligent AI**



Secured Offline Gen AI Assisted Search

Problem:

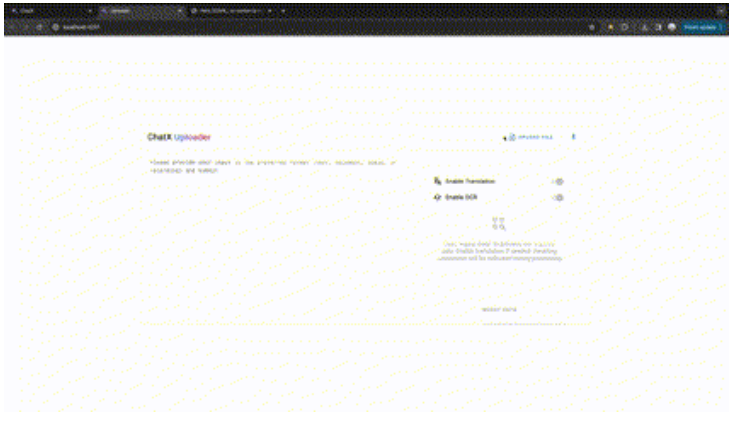
Too much information to analyze efficiently and thoroughly

- Government data silos hinder analysis and retrieval.
- Privacy constraints limit the use of public cloud solutions.
- Multilingual, multimodal data, with continuous updates, poses challenges.

Solution:

AI Enabled Search on GDC

- **Mission-critical Insights**
- **Multi-mode Analysis**
- **Uncompromising Security**



Thank you