



SEPTEMBER '26 INSIGHTS

Accelerated Computing Enterprise is the Indian arm of Cloud Ace Group.

Cloud Ace was established in Japan in 2014 and became the first Managed Service Provider (MSP) for Google Cloud in the country. Cloud Ace have won the Google partner of the year award in 2018 & 2021 as we continued to grow our business to other part of Asia and now we are expanding into Europe, North and South America as well as Africa.

We can help you in digital transformation by modernizing your infrastructure with best in-class cloud-native and cloud advisory services.

Inside This Issue

- Our Mission, Vision & Commitment
- Specialization
- Our Services
- Our Global Collaboration
- **Article**
- **Production Ready Agentic AI on Google Cloud**

10
Countries

600+
Experts

1000+
Happy Clients

OUR MISSION

Build our work on Honesty.

It is our mission with an unwavering pride that we are always aware of as we work to realize our vision; not just being honest, but “Honesty itself is our work and our service.”

OUR VISION

To contribute to building a world where all people can utilize advanced technology and realize the dreams they have envisioned.

We aspire to build a world where every person in the world can achieve their ambitions through the use of advanced technology. Our vision is to create such a world and help those who want to achieve it.

OUR COMMITMENT

As Google Cloud Partners, we are committed to providing our clients with cutting-edge technologies and exceptional customer service. Our team of experts is trained and certified to help you leverage the full potential of the Google Cloud Platform, from cloud storage and networking to machine learning and artificial intelligence.

SPECIALIZATION

This achievement reflects our continued commitment to delivering exceptional cloud solutions and deep technical expertise on the Google Cloud Platform (GCP). More importantly, it enables us to offer you enhanced support through Google Funding, specifically for any Engineer Support work related to GCP.

What this means for our Clients:

- Access to Google-backed funding to support your cloud engineering initiatives
- Priority support from certified experts across various GCP services
- Improved delivery capabilities and faster time-to-value for your projects
- Recognition that you're working with a trusted, Google-verified partner

We'd love to explore how we can leverage this specialization to add more value to your current or upcoming cloud initiatives. Please let me know a convenient time for a quick call or feel free to reply here — we're excited to help you get started.

Looking forward to continuing our collaboration with even more momentum!



SPECIALIZATION

Infrastructure

Google Cloud

OUR SERVICES



Managed & Consulting Services

Our Managed & Consulting Services Portfolio includes Cloud Infrastructure Services, Cloud Migration Assessment & Planning, Billing & Cost Optimisation, Monitoring, Security, Incident Management and on-going Support Activities.



System Migration, Design, Development & Automation

We provide Application Development, Modernisation Solutions, Cloud Migration Deployment & Optimisation, POCs, Infrastructure Orchestration & Continuous Improvement Activities.



Multi-Cloud, Hybrid Cloud, Backup & Recovery Solutions

If our Customers have Multi-cloud or Hybrid Cloud needs, as Google Cloud Partners, we can help them manage the most Complex Cloud Environments.



AI/ML, Gen AI, Data Engineering, Reporting & IoT Services:

We specialise in providing AI/ML and GenAI Solutions. We can also help our Customers with Data Warehousing, Data Analytics, Dashboarding & Reporting Solutions on top of Google Cloud's Data Services.



Training & Adaptation

Our Certified Professional Google Certified Trainers can help you in Training to ensure that your Team & your Customers are Knowledgeable in GCP & GWS Best Practices, which can inspire Client Confidence.



Google Workspace

We provide Google Workspace Deployment & Migration Business Solutions for setting up Gmail, Google Drive, Google Calendar and other Productivity tools.

OUR GLOBAL COLLABORATION

Expanding Our Global Collaboration to Strengthen Google Cloud Project Delivery in India

At Cloud Ace, we take pride in being one of the leading Google Cloud Partners globally, with a strong footprint across multiple markets. As a Japanese cloud technology company, our commitment to delivering high-quality cloud solutions has always been at the heart of everything we do.

In India, the demand for Google Cloud SI&C (System Integration & Consulting) projects continues to grow rapidly. To ensure we meet and exceed client expectations while maintaining the highest standards of delivery, we are now actively collaborating with our global divisions—including team from Japan.

This strategic collaboration allows us to leverage engineering talent and technical expertise across borders, enhancing our delivery capabilities and ensuring timely execution of projects in India. By aligning with our overseas teams, we aim to bring the best of global knowledge and experience to our local engagements.

This is a significant step forward in our mission to provide reliable, scalable, and future-ready Google Cloud solutions to enterprises in India.

Together, across borders, we continue to build smarter with the cloud.

Production-Ready Agentic AI on Google Cloud

Moving Beyond Prototypes to Secure, Scalable Architectures

Over the past year, enterprise adoption of Generative AI has evolved from static, conversational chatbots to autonomous, goal-driven AI agents. While foundational LLMs provide incredible cognitive reasoning, deploying these autonomous systems into real-world production environments demands strict architectural discipline.

Unlike basic retrieval-augmented generation (RAG) which simply queries a model, autonomous agents actively use enterprise tools — reading databases, calling external APIs, and executing background code — to solve complex tasks. Our engineering engagements show that deploying these systems safely requires building robust, simple infrastructure boundaries around the models. This article outlines the three core pillars of a production-ready agentic architecture on Google Cloud.

Pillar 1: Zero-Trust Tool Execution & Security

When AI agents are given the ability to execute tools, such as querying databases or updating IT service tickets, securing their operational boundaries is critical. Broad permissions can lead to privilege escalation or accidental data exposure.

Ephemeral Access & Least Privilege

Agents should never run with static, highly privileged administrative keys. Instead, utilize Google Cloud Workload Identity Federation to generate short-lived, single-purpose tokens. If an agent is tasked only with looking up customer status, its temporary token should limit its access to a single BigQuery view.

Human-in-the-Loop (HITL) Checkpoints

Operations should be split into read-only actions and mutative actions. While searching information can happen autonomously, any action that updates a database, initiates a transaction, or provisions infrastructure must halt for human approval before execution.

Sensitive Data Protection

Incorporate Google Cloud's Sensitive Data Protection (Cloud DLP) API to automatically mask personally identifiable information (PII) and compliance-regulated data in both agent inputs and model responses.

Pillar 2: Quota Governance & Runaway Loop Prevention

A unique challenge of multi-agent systems is the recursive loop. If Agent A encounters an error from Agent B, it may continuously retry the call with more context, creating an infinite processing loop that exhausts system quotas and generates sudden, unexpected cloud expenses.

Hard Session Depth Ceilings

Configure the orchestrator (such as LangGraph or Vertex AI Agent Builder) with a strict maximum turn limit (e.g., maximum of 5 tool turns per session). If a task isn't resolved within this threshold, the agent must gracefully terminate and flag the session for review.

Circuit Breakers

Implement rate limiters on your tool execution APIs. If an agent triggers more than 10 tool invocations in a single minute, the circuit trips immediately to halt execution.

Environment Segregation

Keep development, staging, and production environments strictly separated in Google Cloud. Apply strict, lower Vertex AI API quota limits in sandbox projects so that unexpected testing loops can never disrupt production workloads.

Pillar 3: Smart FinOps & Context Caching

Agentic workflows consume significantly more tokens than simple conversational queries due to repeated prompt injections, multi-turn tool outputs, and historical scratchpad retention. Left unoptimized, cost scaling can become unpredictable.

Vertex AI Context Caching

For agents that rely on extensive system instructions, API documentation, or large enterprise knowledge bases, enable Google's context caching. This stores the frequent prompt prefix in memory on Vertex AI, reducing input token costs by up to 75% and significantly lowering latency.

Strategic Model Routing (Flash vs. Pro)

Not every sub-task requires the most expensive reasoning engine. Use Gemini

1.5/2.0 Flash for routine classification, metadata parsing, and simple parameter extraction, reserving Gemini 1.5 Pro strictly for complex reasoning and final synthesis.

Conversational Pruning

Periodically compress long-running execution history. Instead of sending raw, multi-page transcripts back to the model with every turn, use a lightweight background task to summarize conversational state.

Simple Checklist for Production Readiness

Before releasing your enterprise AI agents to users, verify you have implemented these essential guardrails:

Authentication: Ephemeral IAM credentials generated dynamically via Workload Identity Federation.

Operational Control: Clear human authorization gateways for database writes or critical modifications.

Cost Efficiency: Vertex AI Context Caching configured for system prompts larger than 32k tokens.

Data Security: Automated Cloud DLP checks to inspect inputs and outputs for sensitive data.

Runtime Safety: Active recursion caps and circuit breakers enabled on the orchestration layer.

Full Observability: Custom Cloud Monitoring dashboards tracking token usage with metric-threshold alerts.

Cloud Ace Advisory & Engineering Support

Designing, securing, and optimizing autonomous AI workflows requires a balanced approach to cloud infrastructure, database design, and machine learning. As a premier Google Cloud Partner, Cloud Ace collaborates with organizations to transition AI prototypes into secure, enterprise-grade cloud systems.

Contact our advisory team today at support@in.cloud-ace.com or visit <https://c-ace.co/> to schedule your technical architecture review.



Accelerated Computing Enterprise Pvt. Ltd.

Delhi NCR

www.c-ace.co | in@cloud-ace.com

Cloud Ace Global | Japan, China, Hong Kong,
Taiwan, Thailand, Vietnam, Singapore, Indonesia,
India, and Brazil.

View Terms of Services: <https://c-ace.co/cain-terms-of-service/>