Informatica

Google Cloud

eBook

5 Keys to Deploying Enterprise-Grade GenAI Applications on Google Cloud

Choose the Right Architecture to Overcome Data Challenges

Where data & AI come to LIFE™

Contents

Enterprises Look to GenAI to Support Innovation

Generative AI (GenAI) is a revolutionary technology that leverages advanced algorithms and deep learning models to create new content and provide innovative solutions across various domains. GenAI models learn the patterns and structures of their training data, allowing them to generate text, images, videos and other content in response to specific prompts. By automating complex tasks, optimizing decision-making and fostering creativity, GenAI empowers businesses to achieve unprecedented levels of efficiency, personalization and a competitive edge. In an era where digital transformation is key, understanding and integrating GenAI into business strategies is not just advantageous—it's essential for sustainable growth and success.

Organizations across all industries globally are adopting AI at a rapid pace to transform operations, drive growth and secure positions as industry leaders. Already, many companies are deploying GenAI applications to support a myriad of valuable use cases, helping them:

- Enhance customer self-service and cut down operational expenses by automating replies to customer inquiries through AI-driven chatbots, voice bots and virtual assistants.
- Boost employee productivity by allowing quick access to accurate information, precise answers and summarization via a conversational interface.
- Speed up application development by providing code suggestions based on developers' comments and existing code.
- Streamline logistics and reduce expenses by analyzing and refining various supply chain scenarios.
- Automate the creation of financial reports, summaries and forecasts, thus saving time and minimizing errors.

Recognizing this value, many enterprises are ramping up their use of GenAI-enabled applications. In a recent survey of global data leaders, 45% of respondents say they have already implemented GenAI. An additional 54% of leaders anticipate they will in the future — and 36% of those expect to deploy the technology within the next two years.[1]

89%

of surveyed executives consider AI and GenAI to be a top-three tech priority[2]

[1] Informatica, "CDO Insights 2024: Charting a Course to AI Readiness," 2024.

[2] Boston Consulting Group, "www.bcg.com/publications/2024/from-potential-to-profit-with-genai," 2024.

“By 2026, more than 80% of enterprises will have used generative AI APIs or models and/or deployed GenAI-enabled applications in production environments, up from less than 5% in 2023.”

Gartner, “**What’s Driving the Hype Cycle for Generative AI, 2024**,” November 14, 2024.

Data Quality Correlates With GenAI Success

Deploying GenAI applications is not without its challenges. According to the global survey, 99% of GenAI adopters have encountered roadblocks such as data privacy and protection, AI ethics and AI governance.[3]

However, the leading challenge is data quality, which 42% of data leaders cited as the main obstacle to their success with GenAI.[3] To ensure the technology generates correct content, you need data that is correct, precise and well-governed. All necessary data fields must be present, with no missing values. And your data must be consistent across datasets and time periods.

With these challenges in mind, how can your organization use GenAI technology — which is commonly known for summarizing information from the public domain — effectively and securely? How can you use it behind your firewall and deploy GenAI-enabled applications that rely on your organization's private, trusted data to generate accurate insights?

93%

of CDOs believe a data strategy is critical to realizing value from GenAI[4]

Only 29%

of organizations are completely ready to use data with GenAI, with processes in place to standardize data definitions and maintain integrity[5]

[3] Informatica, "CDO Insights 2024: Charting a Course to AI Readiness," 2024.

[4] Harvard Business Review, "Is Your Company's Data Ready for Generative AI?" https://hbr.org/2024/03/is-your-companys-data-ready-for-generative-ai, 2024.

[5] IDC, "Making the Case: Data Governance for GenAI," August 2024.

Data Quality Correlates With GenAI Success (continued)

Data leaders recognize the value of GenAI

Despite these challenges to GenAI implementation, a majority of data leaders believe the technology is a worthwhile investment.[6]

73%

use or plan to use GenAI to improve time to value with faster data insights

66%

want to drive more productivity through automation and augmentation

[6] Informatica, "CDO Insights 2024: Charting a Course to AI Readiness," 2024

5 Key Requirements for Enterprise GenAI Applications

To improve the success of your organization's GenAI initiatives, you must ensure that enterprise-grade GenAI applications address the following five requirements:

1

Grounded

Grounding prompts and responses with your enterprise's unique data is critical and can be achieved by using a Retrieval-Augmented Generation (RAG) framework. RAG is a hybrid model architecture that combines retrieval mechanisms with generative models to enhance the quality and accuracy of generated content. RAG-based architectures help to ensure precise and relevant responses that are based on your organization's data, offering more accurate and informed answers than a standard generative model might provide on its own. Enterprises that do not use RAG often find that their models hallucinate and create erroneous outcomes.

2

Contextualized

The large language models (LLMs) that power GenAI rely on extensive datasets, which are often sourced from publicly accessible knowledge bases, such as the internet. Many of these datasets lack the specialized knowledge needed for industry- or company-specific tasks. Because of this, the LLMs often fail to understand enterprise terminology and semantics provided by users in their prompts. When data is not contextualized, your GenAI applications can create generic responses that fail to meet the user's needs.

Contextualization helps ensure that prompts are enriched with your business context and can produce rich summarizations applicable to your business. By bringing the semantic meaning of your specialized terminology and the language of your business into your models, you can create far more effective conversations between users and your data.

5 Key Requirements for Enterprise GenAI Applications (continued)

3

High data quality

Not all data is created equal. To ensure that your GenAI applications deliver useful insights and accurate responses, you need to ensure that they use high-quality and well-governed data. Models are best able to return results based on accurate, high-quality data when you use clean, complete master data. This may require that you take steps to consolidate multiple records maintained by your company, making sure that your models can use a golden record. You also can select and prioritize the data to be used for prompt enrichment and summarization, which will increase the quality of responses from the LLM.

4

Simple to develop and deploy

To realize maximum value, you need to be able to develop and deploy GenAI applications with limited hand-coding. You also want to be able to rapidly adopt new and evolving GenAI capabilities. Achieving these goals requires GenAI capabilities that are accessible to people at various skill levels. Look for solutions that democratize GenAI by making tools available to a wide range of users — from pro-code roles such as data scientists and data engineers to low-code or no-code workers such as business analysts, citizen integrators and beyond. With these tools, you can rapidly build applications, providing transparency and enabling reuse and portability of the applications to users across the enterprise.

5

Governed and secure

When you bring enterprise data to your GenAI applications, it's crucial to ensure that you comply with your organization's data use policies so that these applications only share appropriate data with users based on their job title, function and data access level within the organization. For example, an entry-level employee in the logistics department likely shouldn't be able to access to customers' payment and banking information, but a senior accounts receivable analyst might. In addition, when data is enriched through the RAG framework, you need to be able to specify which data is accessible. Being able to trace the lineage and governance of data outputs from enterprise GenAI applications as they develop and evolve is critical from a business perspective.

Informatica's Framework for Deploying GenAI Applications on Google Cloud

Addressing these five key requirements can be simple when you use an architectural framework designed for your cloud environment. Informatica's GenAI blueprint for Google Cloud integrates the Informatica Intelligent Data Management Cloud™ (IDMC) platform with Google Cloud Vertex AI and Gemini LLMs (see figure). You can use this blueprint as a guide to deploy enterprise-grade GenAI applications on Google Cloud.

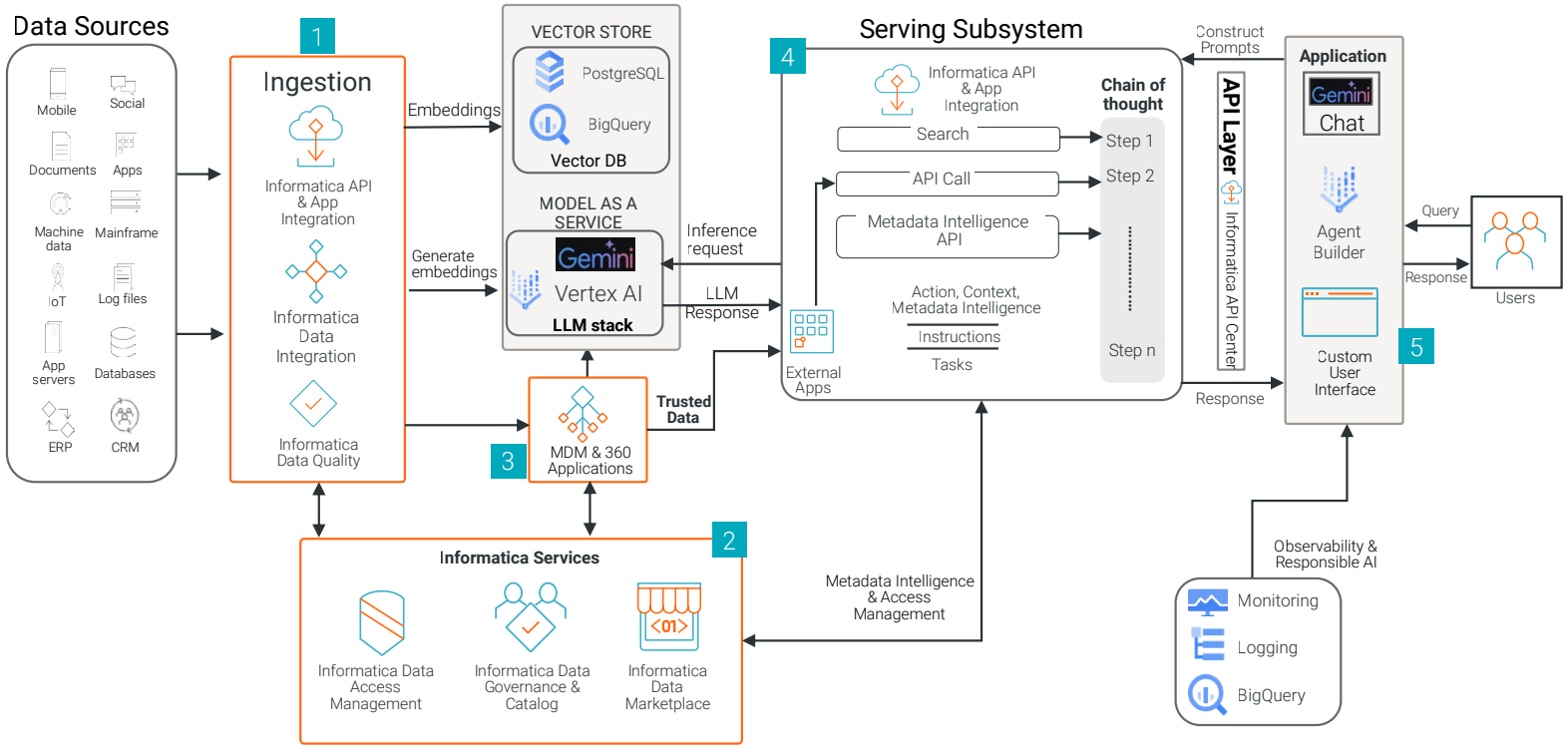

Figure 1: Informatica's Framework for Deploying GeneAI Applications on Google Cloud

Informatica's Framework for Deploying GenAI Applications on Google Cloud (continued)

This framework provides a strong foundation to help you address the five key requirements for enterprise-grade GenAI initiatives. It increases your chances of success by enabling the following processes:

1. **Data ingestion:** In this framework, IDMC ingests data from various sources, including applications such as Salesforce, SAP and Workday; on-premises databases and data warehouses such as Teradata or SQL Server Database; any cloud data warehouse; and streaming sources such as Kafka. Once the data is ingested, Vertex AI generates the embeddings, which are stored in the Google BigQuery vector database. You can schedule auto-ingestion at a desired frequency to ensure your latest enterprise data is available to be used by the GenAI application.
2. **Metadata intelligence and access management:** IDMC includes robust data governance, catalog, policy-based data access management and master data management (MDM) capabilities. These features enhance GenAI applications by ensuring the use of high-quality, well-managed data with secure, role-specific access and rich metadata for context. Together, these capabilities boost the precision and innovation of RAG models.

Informatica's Framework for Deploying GenAI Applications on Google Cloud (continued)

3. **Trusted data:** IDMC provides trusted data using the Informatica MDM solution. MDM consolidates and maintains a single, high-quality record for data entities such as customers, products and suppliers even when integrating data from various sources. This unified master data provides a reliable, consistent foundation that enables accurate reporting, reduced errors, elimination of redundancy and informed decision-making across your organization.

   Informatica also offers the MDM Extension for Google Cloud BigQuery, which provides reliable and trusted master data from Informatica MDM directly in Google BigQuery. The MDM Extension can reduce the time to onboard high-quality customer master data from weeks to minutes. You can rapidly develop and deploy a customer data platform and GenAI applications on Google Cloud, driving improved marketing strategies, accurate forecasting and deeper customer insights. By consolidating key master and transaction data from multiple sources, you can develop enterprise-grade GenAI applications grounded in trusted, high-quality master data across key domains.

4. **Serving subsystem:** The serving subsystem is built using Informatica Cloud Application Integration (CAI), which offers a low-code/no-code development experience that accelerates and democratizes the development of GenAI applications. CAI builds the chain of thought using API calls to external systems for metadata or master data.

   Using mastered golden records from MDM and metadata from our data catalog and data governance capabilities, IDMC can provide contextual information for RAG pipelines. With this process, CAI can add more context to the query and send it to the LLMs, ensuring more accurate responses that are rooted in your organization's unique data and use cases. Additionally, the Informatica API manager can manage API traffic for optimal performance.

5. **Front end:** The user interface can be a custom application built by your organization or any existing front-end GenAI application, such as Gemini Chat or Vertex AI Agent Builder. When an end user inputs a query, the front-end application calls the CAI API through the API manager and initiates the RAG chain.

Prebuilt Integration Recipes Democratize GenAI Application Development

You can get started quickly with developing enterprise-grade GenAI applications by using prebuilt integration recipes that support common use cases. Each recipe is a set of preconfigured assets such as process objects, app connections and processes for each use case.

Recipes help democratize GenAI app development, since they eliminate the need for hand-coding while quickly creating a process for your specific use case. Informatica's iPaaS recipes for GenAI fall within the following three general use case categories.

AI agents

Agent systems are AI programs capable of autonomous decision-making on behalf of users, systems or other programs. Informatica offers pre-built recipes to support popular AI Agent frameworks, thus accelerating project initiation.

Recipes:

- **AI agent for connecting with Salesforce:** Uses the Gemini AI Agent framework to interact with Salesforce and address user queries autonomously.
- **Simple react agent/function calling:** Illustrates building of a simple react agent using LLM that can autonomously generate and execute tasks for user queries.

Prebuilt Integration Recipes Democratize GenAI Application Development (continued)

Prompt engineering

The task of crafting clear prompts for AI language models to enable the generation of accurate and helpful responses is called prompt engineering. Informatica iPaaS offers a simplified way to orchestrate and govern LLM calls/prompts within a low-code/no-code environment through out-of-the-box LLM connectors.

Recipes:

- **Prompt chaining:** Designs prompt chains and resolves them in sequence so that the LLM provides the desired responses.
- **Chat with history:** Provides a user's prior chat history in a file and uses it as context for the next query asked and the LLM's response.
- **Chat with file:** Allows users to ask questions to the Gemini LLM based on an uploaded file's contents.
- **Chat with file using guide:** Uses Gemini LLM to upload a file with context, read the text and answer the user's questions based on the file's contents.

RAG consumption

Retrieval Augmented Generation (RAG) is a technique for integrating authoritative or proprietary data sources into GenAI models, thereby improving the accuracy, trustworthiness and context of LLM responses. As part of Informatica's GenAI blueprint for Google Cloud, Informatica's iPaaS recipes for Google Gemini orchestrate automated, low-code LLM calls/RAG pipelines seamlessly via out-of-the-box LLM connectors to data sources for RAG, such as vector databases.

Recipes:

- **Loan processing with GenAI:** Uses RAG to validate and evaluate loan requests and approve or reject them based on the applicant's credit score.
- **Simple RAG consumption with Pinecone database:** Converts user queries into vectors, and uses them to search for similar vectors in Pinecone Vector DB to form a context. This context and the original query are passed to Gemini LLM to generate and return a comprehensive response.

Our Joint Commitment Drives Your Success

Informatica Intelligent Data Management Cloud (IDMC) empowers businesses to unlock the full potential of their data on Google Cloud and build a trusted data foundation for organizations to accelerate the adoption of game-changing analytics and AI use cases. Together, Informatica and Google are ready to help you successfully deploy GenAI applications within your enterprise. To learn more, visit our **partner page**.

About Us

Informatica (NYSE: INFA), a leader in enterprise AI-powered cloud data management, brings data and AI to life by empowering businesses to realize the transformative power of their most critical assets. We have created a new category of software, the Informatica Intelligent Data Management Cloud™ (IDMC), powered by AI and an end-to-end data management platform that connects, manages and unifies data across virtually any multi-cloud, hybrid system, democratizing data and enabling enterprises to modernize their business strategies. Customers in approximately 100 countries and more than 80 of the Fortune 100 rely on Informatica to drive data-led digital transformation. **Informatica. Where data and AI come to life.™**

Where data & AI come to

Worldwide Headquarters
2100 Seaport Blvd.
Redwood City, CA 94063, USA
Phone: 650.385.5000
Fax: 650.385.5500
Toll-free in the US: 1.800.653.3871

informatica.com
linkedin.com/company/informatica
x.com/Informatica

CONTACT US

IN19-5102-0125