Informatica

ORACLE | Partner

eBook

Deploying Enterprise-Grade GenAI Applications on Oracle Cloud Infrastructure (OCI)

Best Practices to Overcome the Top 5 Challenges

Where data & AI come to LIFE™

Contents

How GenAI Supports Innovation

Enterprises today are rapidly adopting Generative AI (GenAI), a transformative technology that utilizes deep learning models and advanced algorithms to generate new content and provide innovative solutions across various domains. GenAI models identify and learn the patterns and structures in their training data, allowing them to create text, images, videos, and other content in response to targeted prompts.

GenAI empowers businesses to achieve unprecedented levels of efficiency, personalization and a competitive edge by automating complex tasks, optimizing decision-making and fostering creativity. In an era where digital transformation is a requirement, understanding and integrating GenAI into business strategies is not just advantageous — it's essential for sustainable growth and success.

Organizations across industries worldwide are adopting AI to transform operations, drive growth and secure positions as industry leaders. In a recent Gartner survey, 95% of CIOs expect GenAI to significantly improve their organizations — with 74% citing increased productivity and 49% anticipating enhanced customer experiences.[1]

Already, GenAI applications support a myriad of valuable use cases, helping companies:

- 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

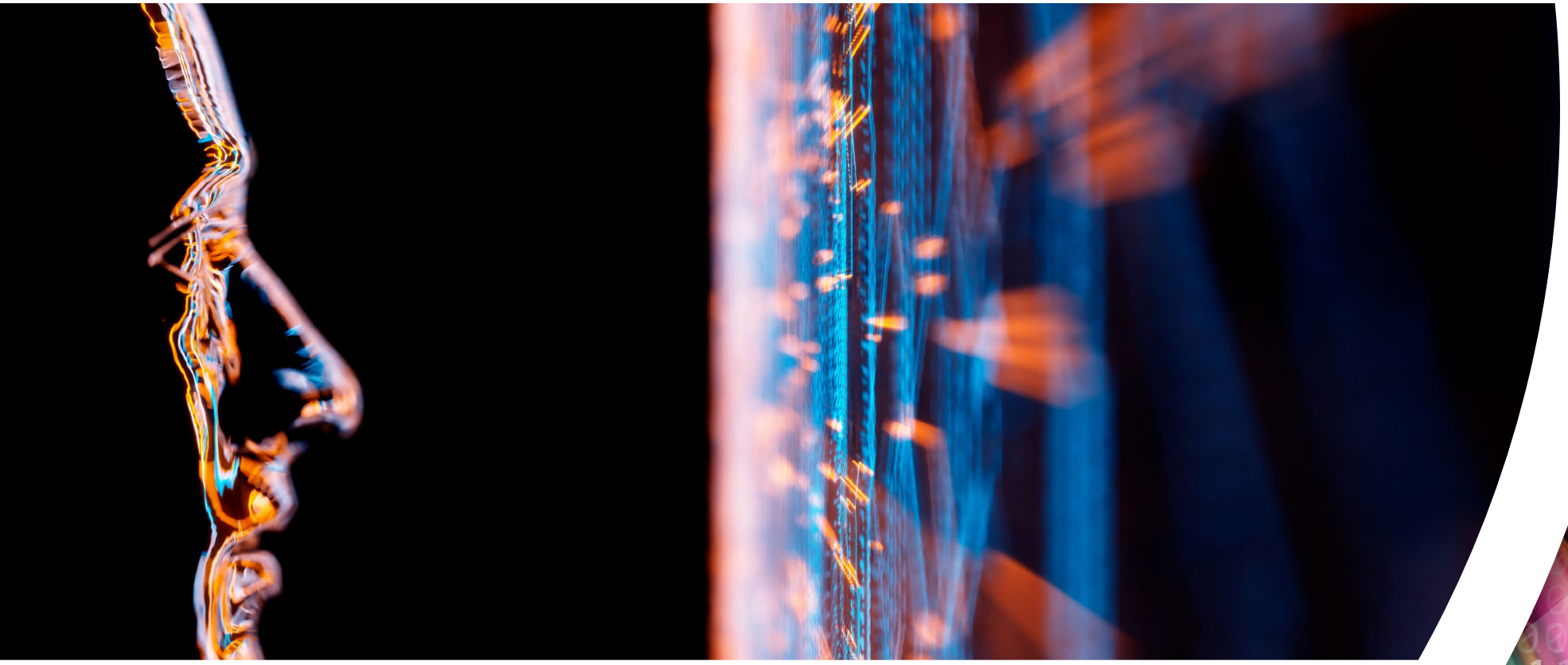
With these early successes, many enterprises are ramping up their use of GenAI-enabled applications. In a recent survey of global data leaders, 45% of respondents say they have already implemented GenAI. An additional 54% of leaders anticipate they will do so in the future — and 36% of those expect to deploy the technology within the next two years.[2]

[1] Gartner, **2024 Gartner CIO Survey**, generative AI survey (ID G00820936), September 2024.

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



How GenAI Supports Innovation (continued)

Data Leaders Recognize the Value of GenAI

Despite the common challenges of deploying GenAI, a majority of data leaders believe the technology is a worthwhile investment.[4]

89%

of surveyed executives consider AI and GenAI to be a top-three tech priority[3]

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

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

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

Data Quality Correlates With GenAI Success

Despite this enthusiasm, data leaders need to take steps to effectively and securely use GenAI technology, which is known for summarizing information from the public domain. How can you best deploy GenAI-enabled applications that rely on your organization's private, trusted data to generate accurate insights?

It's essential to prepare for impediments typically encountered while deploying GenAI applications. According to the CDO Insights survey, 99% of GenAI adopters have faced roadblocks such as data privacy and protection, AI ethics and AI governance.[5]

However, the main obstacle cited by 42% of data leaders is data quality.[5] To ensure the technology generates correct content, your data must be 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.

93%

of CDOs believe a data strategy is critical to realizing value from GenAI[6]

Only 29%

of organizations are completely ready to use data with GenAI and have processes in place to standardize data definitions and maintain integrity[7]

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

[6] Harvard Business Review, "Is Your Company's Data Ready for Generative AI?" 2024.

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

5 Top Challenges in Deploying Enterprise GenAI Applications

Let's consider the five most common challenges you may encounter when preparing to deploy GenAI applications and some best practices to overcome them.

Hallucinations

In Generative AI, "hallucinations" occur when the model produces plausible-sounding content that is factually incorrect or nonsensical, as it relies on data patterns rather than true understanding. To ensure that GenAI applications deliver precise and relevant responses and reduce hallucinations, you must ground prompts and responses with your enterprise's unique data. You can achieve grounding 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 offer 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.

5 Top Challenges in Deploying Enterprise GenAI Applications (continued)

Lack of Context

The large language models (LLMs) that power GenAI typically 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. Make sure data is contextualized to avoid having your GenAI applications provide generic responses that fail to meet users' needs.

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

Poor Data Quality

If the data your GenAI applications are referencing is incomplete, low quality, or inaccurate, then the outputs from your applications will be as well. To ensure that your GenAI applications deliver useful insights and accurate responses, it's critical to use high-quality, well-governed data. Clean, complete master data helps models return optimal results. Creating precise, high-quality data may require consolidating multiple records maintained by your company, making sure that your models can use a golden record. You can also select and prioritize the data to be used for prompt enrichment and summarization, which will increase the quality of responses from the LLM.

5 Top Challenges in Deploying Enterprise GenAI Applications (continued)

Complicated to Develop and Deploy

While application development inherently demands specialized skills due to its complexity, if developing and deploying GenAI applications takes too long, it can lead to significant challenges and costs. To keep pace with new and evolving innovations, you need to be able to develop and deploy GenAI applications with limited hand-coding. Look for solutions that democratize GenAI by making tools available to a wide range of users at various skill levels — 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. The right tools can help you rapidly build applications, providing transparency and enabling the reuse and portability of the applications to users across the enterprise.

Security and Compliance

Without proper guardrails, GenAI applications may unintentionally access or disclose confidential or restricted data to unauthorized individuals within an organization. To ensure that GenAI applications share enterprise data appropriately, they must comply with your organization's data-use and access policies. For example, an entry-level employee in the logistics department likely shouldn't be able to access customers' payment and banking information, but a senior accounts receivable analyst might. When data is enriched through the RAG framework, you can specify which data is accessible to whom. Being able to trace the lineage of data outputs from enterprise GenAI applications as they develop and evolve is also critical to protecting data security and ensuring compliance with all data-use policies.

Informatica's Framework for Deploying GenAI Applications on Oracle

Informatica can help you address these five key challenges with an architectural framework designed for your Oracle Cloud environment. OCI is optimized for GenAI workloads, offering industry-leading NVIDIA bare metal and virtual machine GPU cluster options powered by an ultrahigh-bandwidth network and high-performance storage to fit your AI needs.

As an Oracle-preferred partner for enterprise cloud data integration and governance, Informatica offers a GenAI blueprint that integrates the Informatica Intelligent Data Management Cloud™ (IDMC) platform with Oracle Autonomous Database, Oracle Database 23ai and any LLMs (see Figure 1). Oracle Database 23ai plays a unique role in optimizing GenAI data pipelines because of its embedded vector processing capabilities, which capture voice, video, image, and language data, and combine it with structured enterprise data (relational, text, JSON, spatial, and graph-data types to enhance your apps)—all in a single database. This allows you to bring AI to your data – not move your data for AI. This innovation allows users to rely on a single database for AI vector generation and high-quality RAG fine-tuning, and eliminates the need to move data to specialized databases.

Oracle Database 23ai integrates AI capabilities directly into the database architecture, specifically with its "AI Vector Search" feature, which allows users to query data using natural language and perform semantic searches on complex data sets. These capabilities make Database 23ai a powerful tool for analyzing and extracting insights from large volumes of information, all while maintaining the traditional strengths of Oracle databases like scalability and performance. By leveraging Oracle Database 23ai as part of Informatica's GenAI blueprint, the resulting architecture is simpler.

You can use this blueprint as a guide to deploy enterprise-grade GenAI applications on Oracle Cloud Infrastructure (OCI).

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

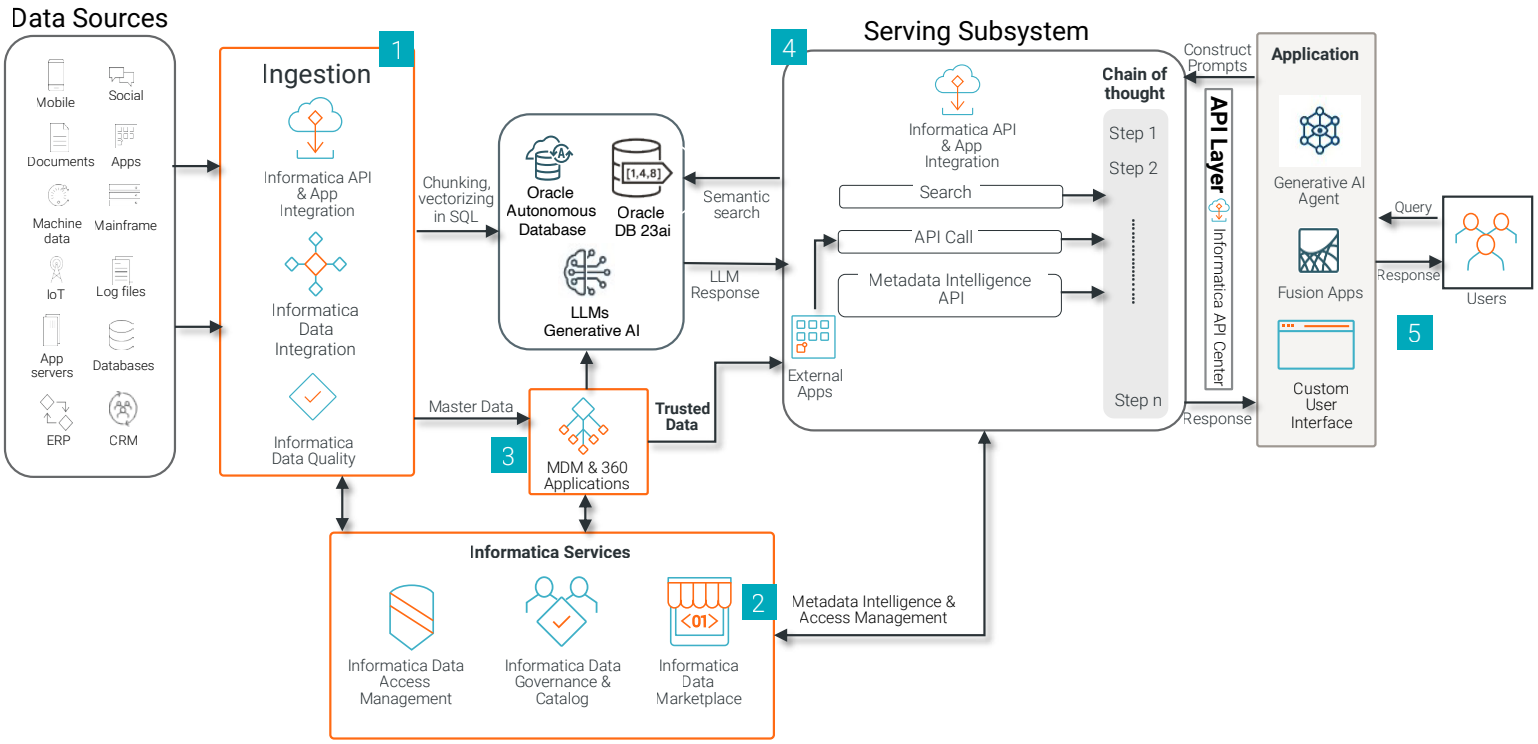


Figure 1: Informatica GenAI Architecture for Oracle

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

This framework provides a strong foundation to help you address the challenges of 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 E-Business Suite, Siebel and Workday; on-premises databases and data warehouses such as Oracle Database, Exadata or SQL Server Database; any cloud data warehouse; and streaming sources such as Kafka. Once the data is ingested, Oracle Database 23ai orchestrates chunking and creates vector embeddings. Database 23ai is a RAG vector database that provides AI Vector Search, broad LLM support, analytics, native support for all modern data types, and the latest development paradigms—all built into one product. The next step is to create a GenAI profile, which specifies an LLM for the ingested data.

2. **Metadata Intelligence and Access Management:** IDMC includes robust data governance, catalog, policy-based data access management and marketplace capabilities, and offers native scanners for Oracle GoldenGate. 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.

3. **Trusted Data:** IDMC provides trusted data using the Informatica Master Data Management (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, reduces errors, eliminates redundancy and supports informed decision-making across your organization.

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

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. For prompt engineering CAI enhances the user's natural language query by supplementing additional context from IDMC data governance and catalog. Translating natural language from user query to SQL statement is one of example of chain of thought. Oracle Autonomous Database Select AI easily translates the natural language prompt to SQL statements, then executes the generated SQL statements and returns the result as a response to the user's query. This enables natural language search across your private business data using RAG to guide the LLM of your choice and steer it away from hallucinations.

IDMC can provide contextual information for RAG pipelines using mastered golden records from MDM and metadata from our data catalog and data governance capabilities. With this process, CAI is able to provide additional context to the query and send it to the LLMs, ensuring more accurate responses 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 OCI Generative AI Agent or Fusion Apps. 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 Accelerate GenAI Application Development

Informatica has developed prebuilt integration recipes that support common use cases, helping you get started quickly developing enterprise-grade GenAI applications. Recipes include preconfigured assets such as process objects, app connections and processes for each use case. By eliminating the need for hand-coding and quickly creating a process for your specific use case, recipes can help democratize GenAI app development.

Our Joint Commitment Drives Your Success

Informatica Intelligent Data Management Cloud (IDMC) empowers businesses to unlock the full potential of their data on OCI accelerate the adoption of game-changing analytics and AI use cases.

As Oracle's preferred partner for Data Integration and Data Governance on Oracle Cloud Infrastructure and the recipient of Oracle's 2024 Partner Award in Business Impact, Informatica enables clients to achieve substantial business results as they navigate and govern their data within OCI.

Together, Informatica and Oracle are committed to empowering customers to achieve meaningful business outcomes with GenAI. For more information, visit the partner pages on the **Informatica** and **Oracle** websites.

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

IN19-5109-0225