Google Cloud

Accelerating generative AI-driven transformation with databases

Your guide to unlocking gen AI's full potential with operational databases.

Table of contents

Chapter 01

Generative AI success starts with your operational data

Your enterprise's success hinges on its data—using it to power applications and extract business value. You're already using operational data—like customer information, financial transactions, and inventory levels—to improve processes and customer experience. And now, gen AI has raised the potential of your data by leaps and bounds.

You already know that gen AI is the key to unlocking further business value. It has the power to transform customer interactions through improved search and personalized assistance. It can supercharge team productivity by assisting developers and administrators in their tasks. It can perform routine tasks to free your staff to innovate and create.

In this paper, we'll show you how to harness the full potential of gen AI with operational databases—and leverage the next generation of AI tools to improve employee productivity.

Gen AI makes modernization more urgent than ever

Leading enterprises are using gen AI in their workflows. And this rapid acceleration of gen AI adoption is creating wider gaps in the market—as businesses that embrace it continue to move ahead, and those that don't are left behind.

As an example, integrating gen AI with your operational data enables relevant and real-time responses that today's customers value. It's the difference between a personalized, accurate response to a customer vs. a generic one.

> 86% of organizations recognize that delivering contextual and relevant user experiences through gen AI integrated databases has a substantial positive impact.[1]

And yet, implementing gen AI isn't as easy as flicking a switch. Many organizations are discovering that their legacy databases are holding them back from the next level of digital transformation.

Only 14% of organizations are satisfied or very satisfied with their legacy databases' support for AI, indicating there is a lot of room for improvement.[1]

Lagging technology and poor user experience are just a couple of the issues caused by legacy databases. Gen AI is bringing new urgency to database modernization because the most popular AI tools for working with vectors, models, and agents run in the cloud. With the right tools, you can harness the power of gen AI within your database to deliver better experiences, drive productivity, and improve data availability.

Let's get started.

Leading enterprises are already using their operational databases with gen AI to improve experiences and drive business value in areas such as:

- Customer support
- Marketing automation
- Product search
- Employee assist

TARGET

In the retail industry, Target utilizes AlloyDB AI to power its next-generation e-commerce search engine. By converting its massive product catalog into vector embeddings, Target can perform similarity searches that better understand a shopper's intent, delivering more relevant results and improving the customer experience at massive scale.

nuro

In the autonomous vehicle industry, Nuro leverages AlloyDB to power the development of its autonomous delivery vehicles. They analyze petabytes of complex simulation and road-test data, running high-performance queries to rapidly identify specific events and accelerate improvements to their driving systems.

NEUROPACE

In the medical device industry, NeuroPace uses AlloyDB Omni to find electrophysiological features that are similar across patients with epilepsy in order to help identify treatment options. They use AlloyDB AI's embeddings function to transform patient iEEG (intracranial electroencephalogram) data into vector representations directly within the database.

[1]Google Cloud Customer Intelligence Data & AI Trends Research, 2024

Chapter 02

Databases power enterprise gen AI apps and agents

Foundation models are large machine learning (ML) models that are trained on generalized data. Off the shelf, they're cost-effective and fast bases for building gen AI apps and agents. However, many enterprises find that these models alone are insufficient for building the contextualized, highly accurate experiences that users demand.

In retail applications, customers increasingly expect agents to provide up-to-date information like stock levels and shipping time estimates. Internally, employees benefit from access to agents that provide self-service, accurate information regarding HR policies. Generically trained models don't cut it for enterprise applications and agents.

That's why many businesses ground foundation models in real-time web information, enterprise data (databases and data warehouses), enterprise applications (ERP, CRM and HR systems), and other sources of relevant information.

Agents are increasingly reaching across data from more parts of your business. And so the more you can ground your models in operational databases, the more powerful your agents and gen AI apps will be.

> "It's a really exciting time for databases, because we're seeing how organizations can bridge the gap between foundation models and enterprise gen AI apps with operational databases to contextualize and personalize the user experience."
>
> **Andi Gutmans**
> VP and GM, Data Cloud, Google Cloud

The most powerful enterprise AI agents are built around three guiding principles:

01

Accuracy

Deliver accurate and up-to-date information.

An operational database stores and processes your data in real time, making it the most reliable source of up-to-date information. And if you aren't integrating this data in your app, it will fall short of its full capability. The retrieval augmented generation (RAG) technique enables you to leverage fresh or domain-specific data into your foundation model—opening up new opportunities to build gen AI apps to deliver answers that are accurate, informative, and relevant to your end-users.

02

Context

Offer relevant user experiences.

Referencing an easily updated knowledge base enables enterprise gen AI apps to provide responses that are more relevant.

Vector embeddings convert text into numerical representations, allowing a foundation model to understand semantic similarities between words and phrases. Vector search then enables the model to quickly find the most relevant information from vast amounts of data.

RAG workflows can use these vector embeddings to retrieve relevant data into foundation model prompts to refine them. This minimizes hallucinations, gives more context to the foundation model's answers, and provides more reliable information.

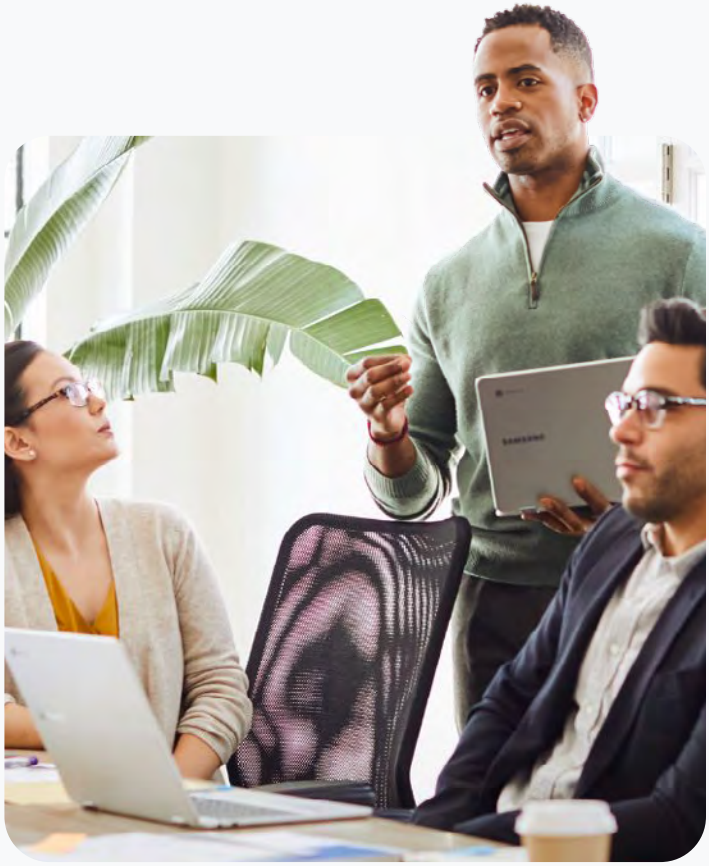
03

Simplicity

Easy for developers to build, operate, and modify.

Any technology relies on the people using it—and so it's imperative that your team work as seamlessly as possible with your databases and gen AI technologies. Application developers know and understand operational databases, and they can interact with the apps under development. The basic framework already exists to make use of your operational data in your enterprise application without having to learn an entirely new system.

Google Cloud databases include support for vectors, meaning you don't need a specialized database. Instead, you can streamline your embedding creation and access processes using your regular database.



Use case example: Building a product search agent

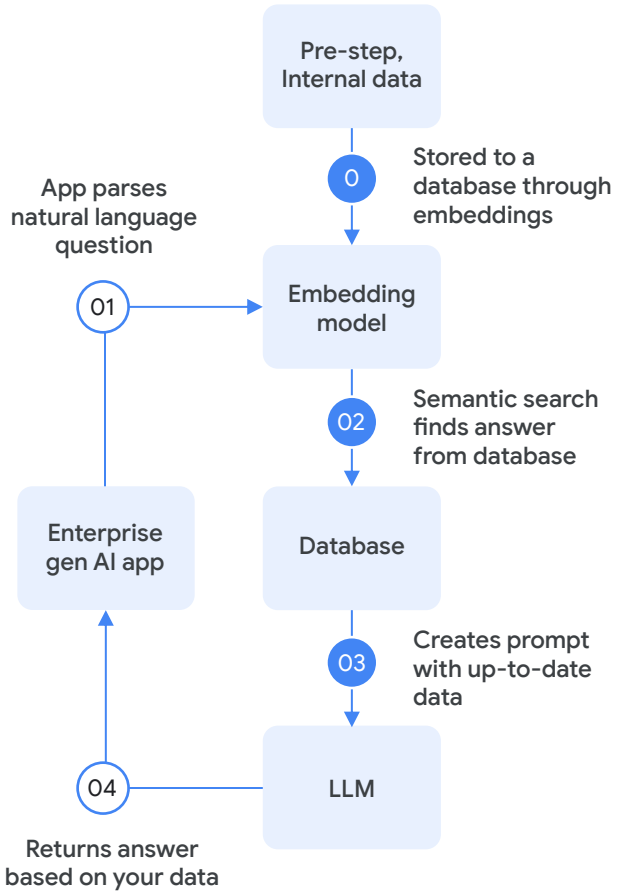
Customers expect prompt, personalized interactions. Using RAG enables gen AI apps to gain access to the information stored in operational databases, in real time.

Let's see how RAG works in a common scenario. In this case, let's look at a shopping agent for a toy company that uses a standard foundation model, augmented by real-time inventory and product information from their operational database.

Now, we have a customer who is looking for some popular toys for kids under five years old.

Using a foundation model, a basic chatbot can already answer a wide range of basic questions about availability and return policies. With RAG, the chatbot can answer even more questions based on up-to-date information about inventory levels or current pricing.

So, instead of simply receiving a recommendation for a child under five, the customer would also be given information about the store closest to them that has the toy in stock. That's the type of personalized response that improves the conversion into a sale.



Here's a look at how RAG works with an operational database:

01 Internal data is stored in a database through the embedding model.

02 Gen AI app uses the embedding model to convert a natural language question "what are some popular toys for kids under five years old" to a vector.

03 Embedding model is used to make a semantic search on the database to retrieve relevant products, and order them according to stock levels.

04 Database returns the search results to be used as part of the prompt for the foundation model.

05 Foundation model constructs an accurate answer based on your data, such as "Here is a list of popular toys for kids under five years old in stock."

Chapter 03

Innovate and transform with Google Cloud databases and gen AI

Google Cloud helps organizations build gen AI solutions and simplify the management of the databases they depend upon. With Google's Data Cloud, data teams can use gen AI tools to activate their enterprise data and use built-in features to easily apply AI/ML directly. For instance, built-in vector embedding capabilities in AlloyDB and BigQuery allow users to store and generate embeddings within their data stores to help support their gen AI use cases.

It's also easy to connect your database to external services that provide additional AI inferencing services, such as Vertex AI, and integrate with orchestration frameworks such as LangChain and LlamaIndex.

> A robust suite of industry-leading databases built on planet-scale infrastructure with AI at its core.

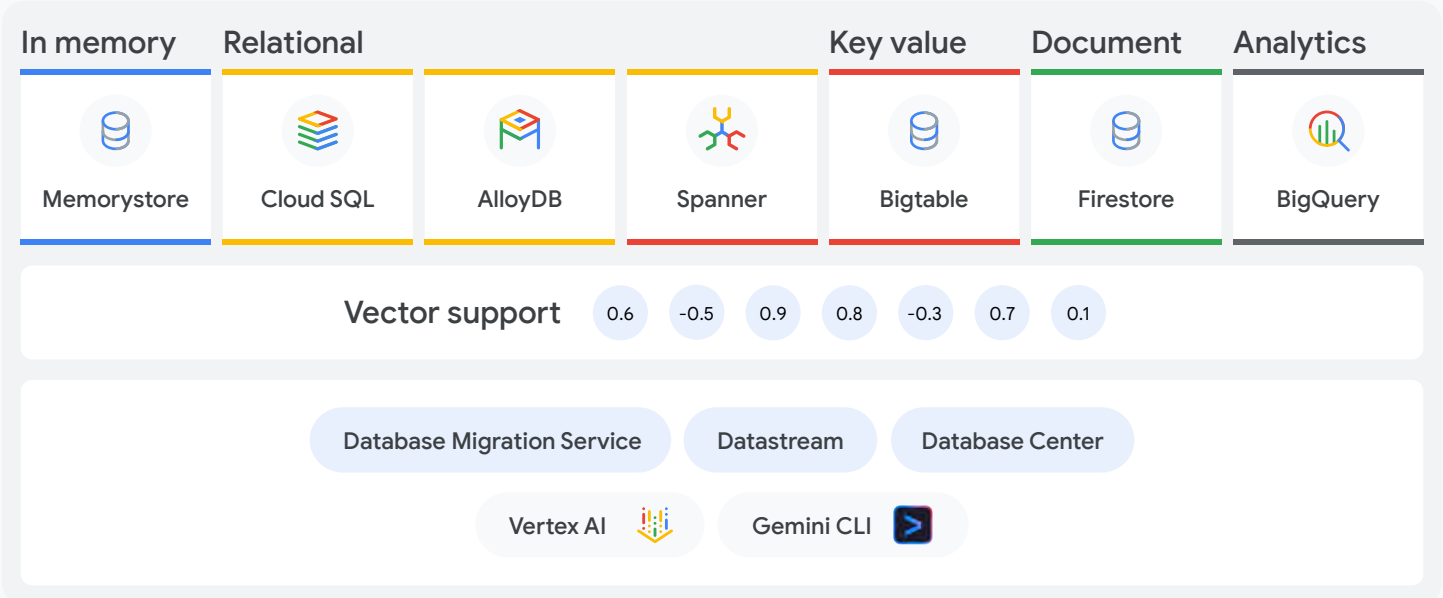
Building enterprise agents faster

Across our database portfolio, Google Cloud delivers world-class vector embedding and search capabilities. Relational databases and non-relational databases alike offer gen AI features to provide a deeper, more meaningful understanding of your data.

Google Cloud databases are simple to integrate with your developer ecosystem. They support the most popular open-source and commercial engines such as MySQL, PostgreSQL, Oracle, SQL Server, Redis, Valkey and Cassandra, and MongoDB—ensuring operational efficiency, accelerated development, and lower total-cost-of-ownership (TCO), allowing you to modernize efficiently at your own pace.

Regnology developed a regulatory reporting chatbot with AlloyDB. This chatbot is designed to expedite the process of obtaining accurate answers to regulatory inquiries, from both internal and external users. AlloyDB acts as a dynamic vector store, indexing repositories of regulatory guidelines, compliance documents, and historical reporting data to ground the chatbot. Compliance analysts and reporting specialists interact with the chatbot in a conversational manner, saving time and addressing diverse regulatory reporting questions.

Infusing gen AI across Google Cloud databases



AlloyDB is helping organizations build gen AI apps

AlloyDB is optimized for enterprise gen AI apps that need real-time and accurate responses. It delivers superior performance for transactional, analytical, and vector workloads. It runs anywhere, including on-premises and on other clouds, enabling you to modernize and innovate wherever you are.

AlloyDB AI is an integrated set of capabilities built into AlloyDB to help developers build performant and scalable gen AI applications using their operational data. It helps them more easily and efficiently combine the power of foundation models with their real-time operational data by providing built-in, end-to-end support for vector embeddings, and offers:

- **Automated embeddings generation.** With a single line of SQL, you can access embedding models, whether they run on Vertex AI or any other platform.
- **Fast, pgvector-compatible vector search** with up to 10x faster index creation, up to 4x faster vector search queries, and up to 10x faster filtered vector search queries than standard PostgreSQL.[2]
- **Integrations with the AI ecosystem,** enabling models to access real-time data and act on external systems through Vertex AI, alongside support for orchestration frameworks like LangChain and LlamaIndex.

AlloyDB Omni was built with portability and flexibility in mind. You can take advantage of the technology in AlloyDB to build enterprise-grade, AI-enabled applications everywhere: on premises, at the edge, across clouds, or even on developer laptops.

What's new in AlloyDB

We are continuing to innovate on AlloyDB and building the next generation of AlloyDB AI, which includes new vector capabilities, easier access to remote models, and secure and flexible natural language support.

At Google, we have more than 12 years of experience innovating on real-world vector algorithms to support some of our most popular services, including Google Search and YouTube. We had to invent new ways of indexing and searching vectors to meet the most demanding use cases. In addition to support for open-source pgvector for our PostgreSQL databases, we are bringing the next generation of tree-based vector capabilities to relational databases.

The ScaNN index is a pgvector-compatible index based on Google's state-of-the-art approximate nearest neighbor algorithms. In our performance tests, AlloyDB can scale to more than a billion vectors with typically less than a 25ms query latency.

AlloyDB now integrates with remote AI models, enabling real-time data transformation and enrichment directly within the database. This is enabled by AlloyDB model endpoint management, a feature that simplifies calling models from Vertex AI, third-party providers like Anthropic and Hugging Face, or other custom services.

AlloyDB AI is designed to accelerate intelligent agent and app development. High-performance filtered vector search enables the smart, multimodal data retrieval that modern apps require. And AlloyDB AI query engine unlocks deep semantic insights from enterprise data through AI-powered SQL operators. Finally, AlloyDB AI natural language turns questions from end users or agents into SQL queries that provide answers. This capability helps you build interactive natural language user interfaces that decipher user intent accurately, so you can build highly-accurate mappings of user questions to SQL queries that answer them. These advances are the future of databases—providing proactive insights for agents that anticipate and act decisively, powered by AI-ready data.

[2] Google internal data, March 2025

Vector search across all Google Cloud databases

Vector embeddings and searches are critical for building useful and accurate gen AI-powered applications—making it easier to find similar search results across unstructured data such as text and images using a nearest neighbor algorithm. Because vector searches are so important, we provide built-in vector capabilities across the entire suite of Google Cloud database offerings for greater operational simplicity and efficiency. You can now store and search across vector embeddings using your existing databases without the hassle of copying data to another vector search solution or learning a separate system.

- **Cloud SQL for PostgreSQL** powers vector search at massive scale, supporting both approximate (ANN) and exact nearest neighbor (ENN) search. It leverages industry-standard indexing techniques like HNSW and IVFFlat, enabling efficient search across hundreds of millions of vectors for demanding, high-scale applications.
- **Cloud SQL for MySQL** also provides integrated vector search, supporting both ANN (using IVFFlat indexes) and ENN search. This allows you to easily store and query millions of vectors for similarity search and recommendations directly within existing MySQL instances.
- **AlloyDB for PostgreSQL** offers high-performance, pgvector-compatible search that runs filtered vector queries up to 10 times faster than the HNSW index in standard PostgreSQL.[3] Your apps can perform fast similarity searches on complex data types such as text and images, using approximate nearest neighbor or exact nearest neighbor algorithms.
- **Spanner** supports exact nearest neighbor vector search on datasets containing trillions of vectors for highly partitionable workloads. It can efficiently reduce the search space to provide accurate, real-time results with low latency—leveraging Spanner columnar engine to rapidly process vector data at a massive scale.
- **Bigtable** offers exact nearest neighbor vector search on datasets containing trillions of vectors for highly partitionable workloads. It can efficiently reduce the search space to provide accurate, real-time results with low latency.
- **Memorystore** provides support for vector storage, enabling ultra low-latency queries for your gen AI apps across Redis, Redis Cluster and Valkey. It's an ultra-low-latency data store suitable for use cases such as foundation model semantic caching and recommendation systems.
- **Firestore** supports exact nearest neighbor vector search. Developers can perform vector search on transactional Firestore data without the hassle of copying data to another vector search solution.
- **BigQuery** supports approximate nearest-neighbor search on BigQuery data. This functionality is key to empowering numerous new data and AI use cases such as semantic search, similarity detection, and RAG.

Orchestration frameworks

Gen AI Toolbox for Databases streamlines the creation, deployment, and management of gen AI agents capable of querying databases with secure access, robust observability, scalability, and comprehensive manageability.

Integration with orchestration frameworks, including LangChain and LlamaIndex, simplifies the process of incorporating Google databases into applications. These frameworks streamline the development of gen AI apps by providing structured, reusable components, which significantly cleans up your code and makes it more modular and maintainable. By leveraging the power of these frameworks with our databases, developers can now easily create context-aware AI agents, faster.

This integration provides developers with built-in RAG workflows across their preferred data sources, using their choice of enterprise-grade Google Cloud database. For instance, the LlamaIndex integration is specifically designed to enhance RAG by efficiently handling data indexing and retrieval, making it a strong choice for applications requiring robust search over large datasets.

Example use cases include personalized product recommendations, question answering, document search and synthesis, and customer service automation. These tools empower developers to focus on application logic over boilerplate code, accelerating the development lifecycle.

[3] Google internal data, March 2025

Supercharge database development and management with AI

In a business landscape where agility and responsiveness are key differentiating factors for success, you need to be able to move fast when market forces change. Database technology is evolving fast, and database professionals are finding it hard to stay up-to-date—which hampers both programming quality and productivity.

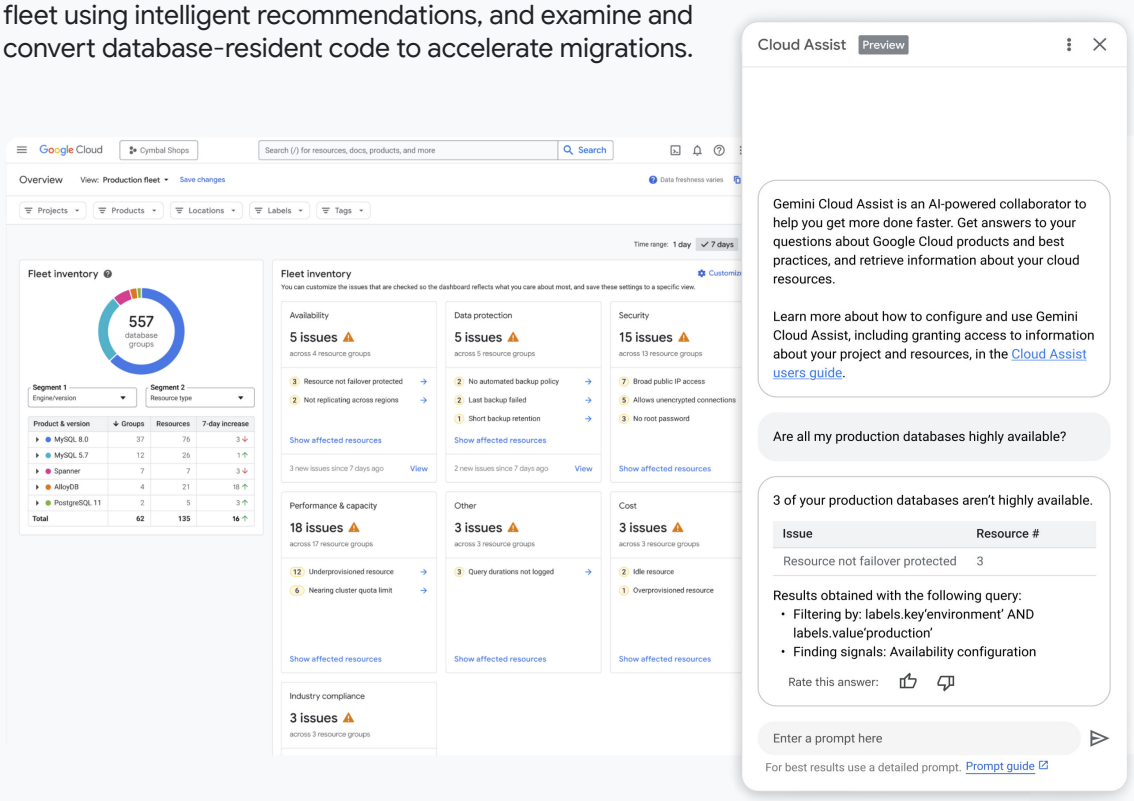
Your operational database is key to managing your organization's data and applications. You want to ensure that data can flow in and out smoothly, and keep your application performing well. Managing databases is a job that comes with a lot of challenges. Many platform engineers, database administrators, and developers juggle ill-fitting tools, complex scripts, and error-prone workflows to complete their tasks.

Database Center changes that. It helps you be more productive and creative—and supercharges your database development and management. AI-powered assistance simplifies all aspects of your database journey, helping you focus on what matters most. With Gemini Cloud Assist enabled, developers, operators, and database administrators can build applications faster using natural language, manage and optimize their entire database fleet using intelligent recommendations, and examine and convert database-resident code to accelerate migrations.

AI can transform how developers operate, by solving critical productivity blockers

> 76% of developers are using or are planning to use AI tools in their development process this year
>
> 81% agree that increasing productivity is the biggest benefit that developers identify for AI tools
>
> 68% spend 30 minutes/day searching for solutions

Source: Stack Overflow Developer Survey, 2024



Database Center allows users to ask ad-hoc questions on their database health and get tailored responses, ultimately enhancing productivity.

Database Studio: Developers can build and deploy applications faster while meeting security and high availability needs with Gemini's ability to generate, fine-tune, and summarize SQL code with simple natural language instructions.

AI-assisted performance troubleshooting: Operators and developers can address database performance issues through an easy-to-use interface, providing visibility into all database metrics in a single view, saving time and enhancing productivity. Database Insights automatically analyzes your workloads, highlights problems, and provides recommendations to resolve them.

Database Center: Database administrators and platform engineers can manage an entire fleet of diverse databases using the intelligent dashboards built with AI, proactively assessing availability, data protection, security, and compliance issues without any custom tools or processes. With the integrated AI assistant in Database Center, database teams can interface with the system using natural language, making it easier to find the information they need and troubleshoot problems.

Database Center assists organizations in achieving compliance with prevalent industry standards, including NIST 800-53, ISO-27001, and PCI-DSS, for their database resources.

Get started with gen AI in your databases

Gen AI is driving a new wave of database modernization—and it's accelerating rapidly. And organizations who fail to modernize will find themselves on the sidelines.

Modernizing is a crucial strategy, but moving from legacy systems can seem like a daunting prospect—as for many, the road to modern systems can seem unclear and intimidating.

Here are the initial steps we recommend when you're starting out on your transformation journey:

01 **Look at the possibilities**. Gen AI is rapidly reshaping the landscape—and implementation can be the difference between falling behind or leaping ahead. Research what your competitors are doing, and get inspiration from how other organizations are using gen AI.

02 **Put together a development team**. Then, align your key decision makers on your goals to move forward. Consider augmenting your existing team with assistive technologies to lighten the load. For example, with Database Center and Gemini Cloud Assist, you don't need to hire specialist database administrators and platform engineers.

03 **Start small**. Identify simple scenarios and create use cases, such as cleaning up your support ticket queue. Gen AI can identify duplicate support tickets for your team, or pull out previously solved tickets that are similar to the one a support person is looking at, to provide precedent and guidance on what to suggest.

04 **Search for opportunities to improve**. This might include automating tasks like maintenance and background business processes, or personalizing the customer experience so your users get a better, more fulfilling interaction with your organization. Think about your operational data, and how you can use it to add context and relevance to your application.

Google Cloud

Start your transformation with Google Cloud

We're here to guide you. With our years of experience in implementing these systems for ourselves and others, we have a solid understanding of the challenges and opportunities you're facing.

To help simplify your modernization journey, Google Cloud offers a database modernization program that combines the best databases, migration tools, expert guidance, best practices, and financial incentives. It's designed to help you move from Oracle and SQL Server to Google Cloud databases, helping your organization meet its gen AI goals.

Contact us to talk about migrating your database from a legacy system, developing a new application, or simply finding the best way forward for your organization.

Talk with a database specialist

The Google Cloud database portfolio

Google Cloud provides an intelligent, open, and unified data and AI cloud to support your gen AI future. Revolutionize customer experiences with operational databases you know and love, in virtually any environment—whether in the cloud or on-premises.

- Cloud SQL is a fully managed relational database service for MySQL, PostgreSQL, and SQL Server.
- AlloyDB for PostgreSQL is a PostgreSQL-compatible database service for your most demanding enterprise workloads. We also offer a downloadable edition—AlloyDB Omni—designed to run anywhere: in your datacenter, your laptop, and on any cloud. Use AlloyDB AI to easily build enterprise agents and gen AI apps.
- Database Migration Service simplifies migrations of MySQL, PostgreSQL, SQL Server, and other databases to the cloud.
- Spanner is a cloud-native database with virtually unlimited scale, global consistency, and up to 99.999% availability. It processes a sustained load of over four billion queries per second.[3] Its PostgreSQL interface simplifies migration from databases like DynamoDB.
- Bare Metal Solution allows you to lift and shift Oracle workloads to Google Cloud.
- Bigtable is a highly performant, fully managed NoSQL database service for large analytical, operational, and time series workloads. It offers up to 99.999% availability, and processes more than five billion requests per second at peak, with more than 10 Exabytes of data under management.[4] Migrate from databases like HBase and Cassandra.
- Firestore is a highly scalable, massively popular enterprise-ready document database service for mobile, web, and server development now featuring MongoDB compatibility to make migration even easier. It offers rich, fast queries and high availability up to 99.999%. It has a thriving developer community of more than 600,000 monthly active developers.[4]
- Memorystore offers fully managed in-memory Valkey, Redis, and Memcached service that offers sub millisecond data access, scalability, and high availability. Memorystore for Valkey and Memorystore for Redis Cluster is a fully managed service which can easily scale to terabytes of keyspace and tens of millions of operations per second.
- BigQuery is a fully managed, AI-ready data analytics platform that helps you maximize value from your data and is designed to be multi-engine, multi-format, and cross-cloud.

[4] Google internal data, March 2025