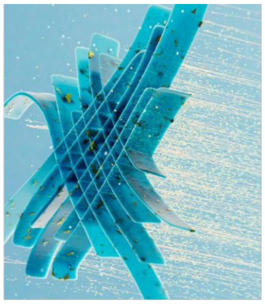
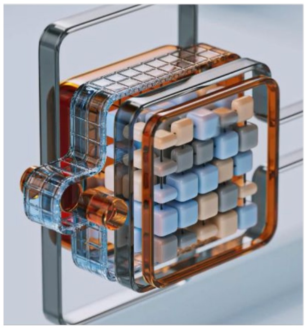
Google Cloud



Accelerating generative AI-driven transformation with databases



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

Table of contents

Generative AI success starts with your operational data

Your enterprise depends on extracting value from your data. You're already using operational data—like customer information, financial transactions, and inventory levels—to improve processes and customer experience. And now, generative 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. And so much more.

In this paper, we'll show you how to harness the full potential 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.
>
> Source: Google Cloud Customer Intelligence Data & AI Trends Research, 2024.

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.
>
> Source: Google Cloud Customer Intelligence Data & AI Trends Research, 2024.

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 data run in the cloud and are based on open source database technologies such as PostgreSQL. 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.

1. **In the financial services industry,** Regnology is using AlloyDB AI to develop a regulatory reporting chatbot. This chatbot is designed to expedite the process of obtaining accurate answers to regulatory inquiries, from both internal and external users.

2. **In the software industry,** Linear uses Cloud SQL for its project management platform for cross-functional teams. They leverage Cloud SQL's pgvector capability for similarity-search features that can easily identify potential duplicates when a user creates a new issue and display related bugs that have already been logged.

3. **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.

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

Databases are at the heart of enterprise gen AI apps

Foundation models are large machine learning (ML) models that are trained on generalized data. They're suited for a multitude of purposes, like content generation, summarization, and simple natural language-based classification. When developing an enterprise gen AI app, using a foundation model as a base is more cost-effective and faster than building the gen AI app from scratch. Yet while they minimize the amount of work required to reach implementation, relying on foundation models also has limitations.

Many businesses find that foundation models alone aren't enough for building the contextualized, highly accurate enterprise gen AI apps required to deliver excellent user experiences. Enterprises are looking beyond foundation models—to grounding them in real-time information and enterprise data. At Google Cloud, we call this "enterprise truth"—the approach to grounding a foundation model in web information; enterprise data like databases and data warehouses; enterprise applications like ERP, CRM, and HR systems; and other sources of relevant information.

The more your business can ground generic foundation models in the enterprise truth that's specific to your products and customers, the more powerful your gen AI app 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**
>
> GM and VP of Engineering, Databases, Google Cloud

Accuracy

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 gen AI model, then your enterprise app 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 that deliver answers that are accurate, informative, and relevant to your end-users.

The most powerful enterprise gen AI apps move beyond generic foundation models and are built around three guiding principles:

Accuracy

Deliver accurate and up-to-date information

Context

Offer relevant user experiences

Simplicity

Easy for developers to build, operate, and modify

Context

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

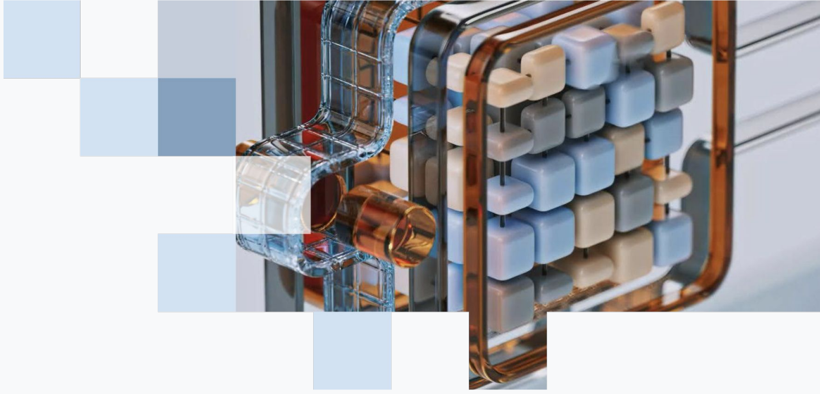
Vector embeddings convert text into numerical representations, allowing a foundation model to understand semantic similarities between words and phrases. Vector search then enables the model to quickly find the most relevant information from vast amounts of data.

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.

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.

Simplicity

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.



Use case example

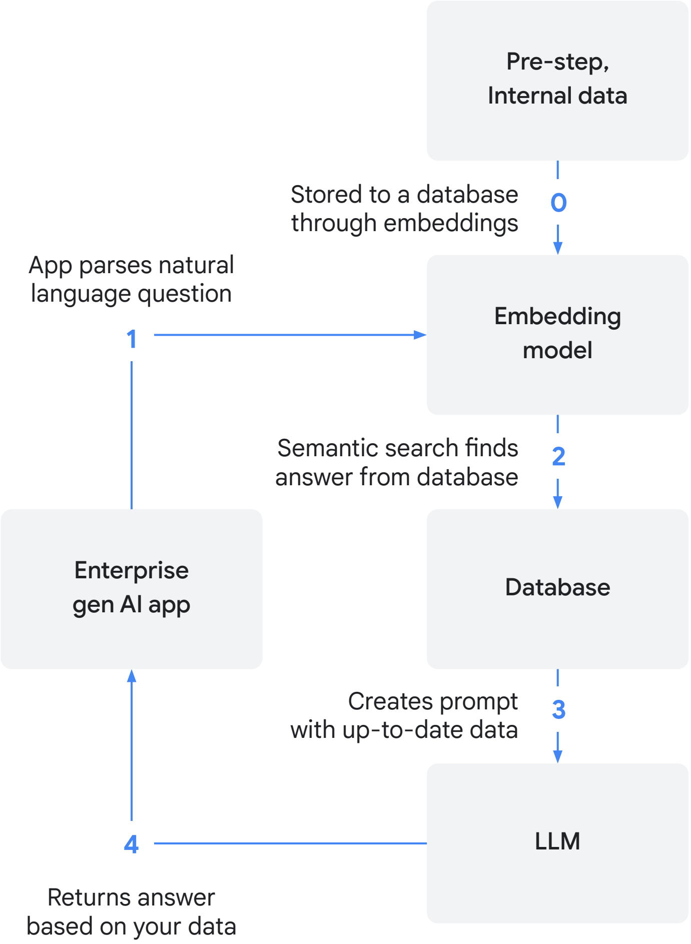
Building a product search app

Customers expect prompt, personalized interactions. And to deliver those unparalleled customer experiences, enterprises are relying on operational databases and gen AI to harness their application data. You can build applications faster with modern databases compared to using legacy systems. And in today's competitive landscape, this speed is a top priority.

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

Now, we have a customer who is looking for some popular toys for kids under five years old. Within the app, they can interact with a chatbot to answer a wide range of questions including availability, pricing, and return policies. When the foundation model is augmented by RAG, the chatbot can answer these questions based on up-to-date information about inventory levels.

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.


Pre-step,
Internal data
Stored to a database through embeddings
0
App parses natural language question
1
Embedding model
Semantic search finds answer from database
2
Enterprise gen AI app
Database
Creates prompt with up-to-date data
3
LLM
4
Returns answer based on your data

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

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

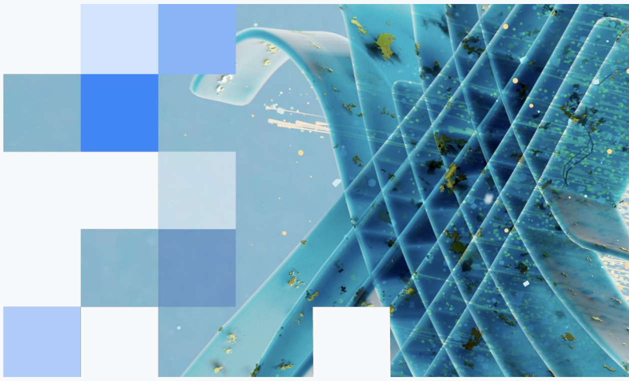
1 Gen AI app uses the embedding model to convert a natural language question ("what are some popular toys for kids under five years old") to a vector.

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

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

4 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."

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 to their data. For instance, built-in vector embedding capabilities in AlloyDB and BigQuery allow users to store and generate embeddings within their data stores to help augment their foundation models and support their gen AI use cases.

Building enterprise gen AI apps faster

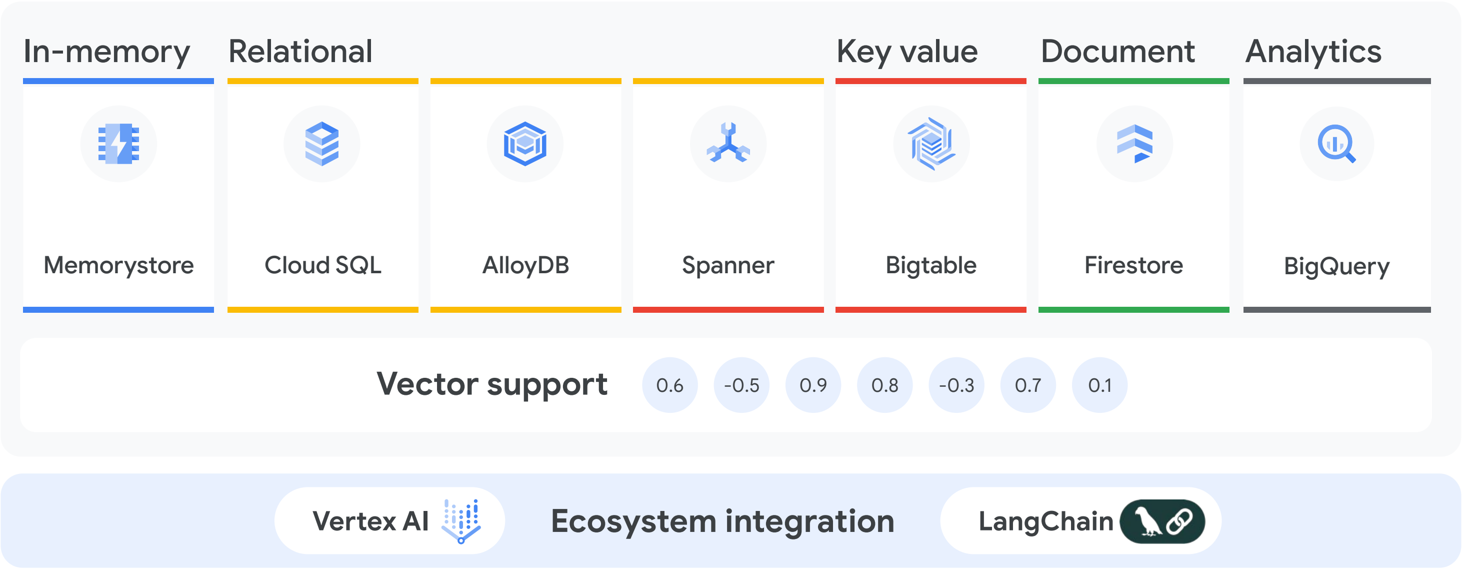
Across our database portfolio, Google Cloud delivers world-class vector embedding and search capabilities. Relational databases and non-relational databases alike offer gen AI features to provide a deeper, more meaningful understanding of your data.

Google Cloud databases are simple to integrate with your developer ecosystem. They support popular open source database standards like PostgreSQL and HBase, making it easy to migrate from legacy databases.

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.

> “All of our databases have vector search capabilities. That means you don't have to deal with complex data pipelines to move your data to specialized vector stores. Furthermore, you can easily perform filter and join operations on your data with your familiar database interface. To top it all, you get the required enterprise-grade data protection, availability SLA, security, and compliance from your database, giving you the peace of mind to future-proof your application.
>
> **Pranav Nambiar**
>
> Director of Product Management, Databases,
> Google Cloud

Infusing Gen AI across Google Cloud Databases



In-memory: Memorystore

Relational: Cloud SQL, AlloyDB, Spanner

Key value: Bigtable

Document: Firestore

Analytics: BigQuery

Vector support

Vertex AI

Ecosystem integration

LangChain

AlloyDB is helping organizations to 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 customers to modernize and innovate wherever they 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 developers 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:

- **Easy embeddings generation.** With a single line of SQL, you can access Google's embeddings models, including both local models and richer remote models in Vertex AI.
- **Enhanced vector support** with up to 10x faster vector queries than standard PostgreSQL. Quantization techniques support four times more vector dimensions and a three-times space reduction.
- **Integrations with the AI ecosystem,** including Vertex AI extensions and LangChain.

AlloyDB Omni was built with portability and flexibility in mind. Customers 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 next for AlloyDB

We are continuing to innovate on AlloyDB 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 new pgvector-compatible index based on Google's state-of-the-art approximate nearest neighbor algorithms. In our performance tests, AlloyDB AI offers up to four times faster vector querying than the popular HNSW index in standard PostgreSQL, up to eight times faster index creation, and typically uses three to four times less memory than the HNSW index in standard PostgreSQL.

To facilitate easier management of inferencing endpoints, AlloyDB model endpoint management makes it even easier to call remote Vertex AI, third-party, and custom models. In addition to Vertex AI, model catalog can also be easily configured for third-party services such as Anthropic and Hugging Face.

Finally, we're bringing two new features in AlloyDB AI to support flexible, accurate, and secure natural language experiences. First, we're enabling gen AI developers to build applications that accurately query data with natural language—just like they do with SQL—for maximum flexibility and expressiveness. That means generative AI apps can respond to a much broader and more unpredictable set of questions. Second, we're adding a new type of database view called "parameterized secure view" that makes it easy to secure your data based on the end-users' context enabling you to deliver richer and more flexible natural language experiences. Together, these advances present a new paradigm for integrating operational data into generative AI apps.

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 MySQL supports exact nearest neighbor search and approximate nearest neighbor search. Developers can store millions of vectors in the same MySQL instances they are already using and search against their vector store.
- Cloud SQL for PostgreSQL supports two search approaches for balancing speed and accuracy. Approximate nearest neighbor vector search is ideal for large datasets where close matches suffice, while exact nearest neighbor vector search is used for precision.
- AlloyDB for PostgreSQL offers high-performance, pgvector-compatible search that runs vector queries up to 10 times faster compared to standard PostgreSQL. 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.
- 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.
- Bigtable will soon offer 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 for Redis provides support for vector storage, enabling ultra low-latency queries for your gen AI applications. 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.

Accelerating ecosystem support for LangChain

Integration with specific LangChain components simplifies the process of incorporating Google databases into applications. By leveraging the power of LangChain with our databases, developers can now easily create context-aware gen AI applications, faster.

The LangChain integration provides built-in RAG workflows across developers' preferred data sources, using their choice of enterprise-grade Google Cloud database. Example use cases include personalized product recommendations, question answering, document search and synthesis, and customer service automation.

Supercharge database development and management with AI

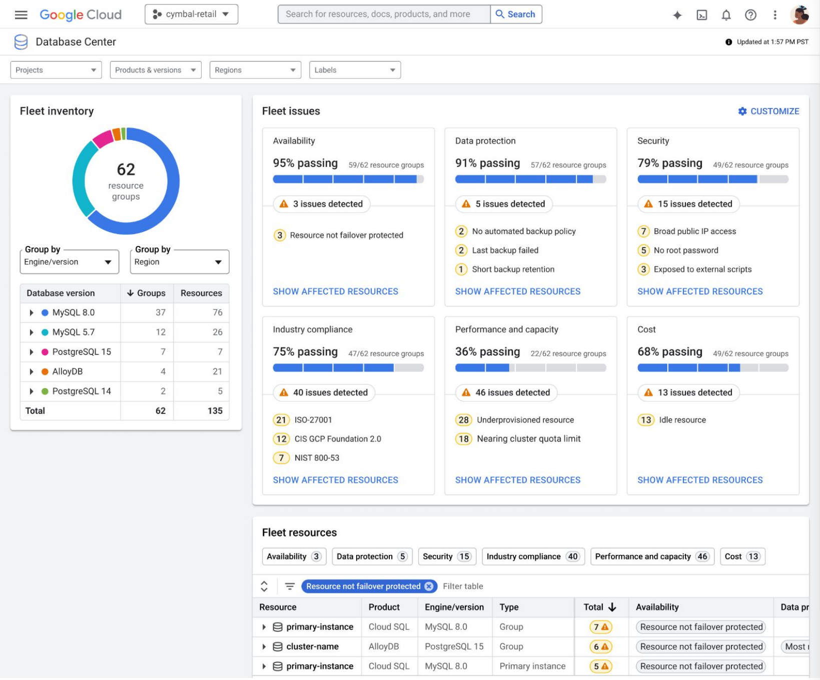
In a business landscape where agility and responsiveness are key differentiating factors for success, you need to be able to move fast when market forces change. Database technology is evolving fast, and database professionals are finding it hard to stay up-to-date—which hampers both programming quality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

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

82%	of developers spend 30 minutes/day searching for solutions
25%	spend more than an hour searching for solutions each day
68%	encounter a knowledge silo at least once a week

Source: Stack Overflow Developer Survey, 2022



Fleet inventory

62 resource groups

Group by: Engine/version
Group by: Region

Database version	Groups	Resources
MySQL 8.0	37	76
MySQL 5.7	12	26
PostgreSQL 15	7	7
AlloyDB	4	21
PostgreSQL 14	2	5
Total	62	135

Fleet issues

Availability
95% passing 59/62 resource groups
3 issues detected
3 Resource not failover protected
SHOW AFFECTED RESOURCES

Data protection
91% passing 57/62 resource groups
5 issues detected
2 No automated backup policy
2 Last backup failed
1 Short backup retention
SHOW AFFECTED RESOURCES

Security
79% passing 49/62 resource groups
15 issues detected
7 Broad public IP access
5 No root password
3 Exposed to external scripts
SHOW AFFECTED RESOURCES

Industry compliance
75% passing 47/62 resource groups
40 issues detected
21 ISO-27001
12 CIS GCP Foundation 2.0
7 NIST 800-53
SHOW AFFECTED RESOURCES

Performance and capacity
36% passing 22/62 resource groups
46 issues detected
28 Underprovisioned resource
18 Nearing cluster quota limit
SHOW AFFECTED RESOURCES

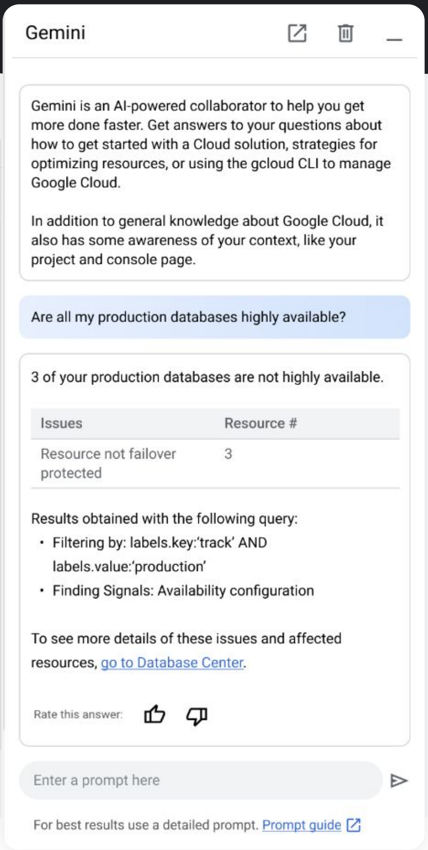
Cost
68% passing 49/62 resource groups
13 issues detected
13 Idle resource
SHOW AFFECTED RESOURCES

Fleet resources

Availability 3 | Data protection 5 | Security 15 | Industry compliance 40 | Performance and capacity 46 | Cost 13

Resource not failover protected ⊗ Filter table

Resource	Product	Engine/version	Type	Total ↓	Availability	Data pr
primary-instance	Cloud SQL	MySQL 8.0	Group	7 ⚠	Resource not failover protected	
cluster-name	AlloyDB	PostgreSQL 15	Group	6 ⚠	Resource not failover protected	Most
primary-instance	Cloud SQL	MySQL 8.0	Primary instance	5 ⚠	Resource not failover protected	


Gemini

Gemini is an AI-powered collaborator to help you get more done faster. Get answers to your questions about how to get started with a Cloud solution, strategies for optimizing resources, or using the gcloud CLI to manage Google Cloud.

In addition to general knowledge about Google Cloud, it also has some awareness of your context, like your project and console page.

Are all my production databases highly available?

3 of your production databases are not highly available.

Issues	Resource #
Resource not failover protected	3

Results obtained with the following query:

- Filtering by: labels.key:'track' AND labels.value:'production'
- Finding Signals: Availability configuration

To see more details of these issues and affected resources, go to Database Center.

Rate this answer:

Enter a prompt here

For best results use a detailed prompt. Prompt guide

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

Gemini for Google Cloud changes that. It provides an AI-powered assistant that helps you be more productive and creative. It can be your writing and coding assistant, creative designer, migration expert, or even your database administrator. For database users, it can help you in multiple aspects of the database journey across development, performance optimization, fleet management, governance, and migrations. Gemini for Google Cloud can help you move away from legacy databases and migrate your data to Google Cloud databases. It can help your developers, database administrators, and platform engineers do their jobs more efficiently, and become better at what they do, with a suite of AI-assisted features for managing, and tuning your database.

Migration: Leverages foundation models to assess and convert the schema or database resident code before migrating data. Easily learn new PostgreSQL dialects, optimize SQL code, and enhance readability for better productivity, easier migrations, and higher efficiency.

Development: 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.

Performance optimization: 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.

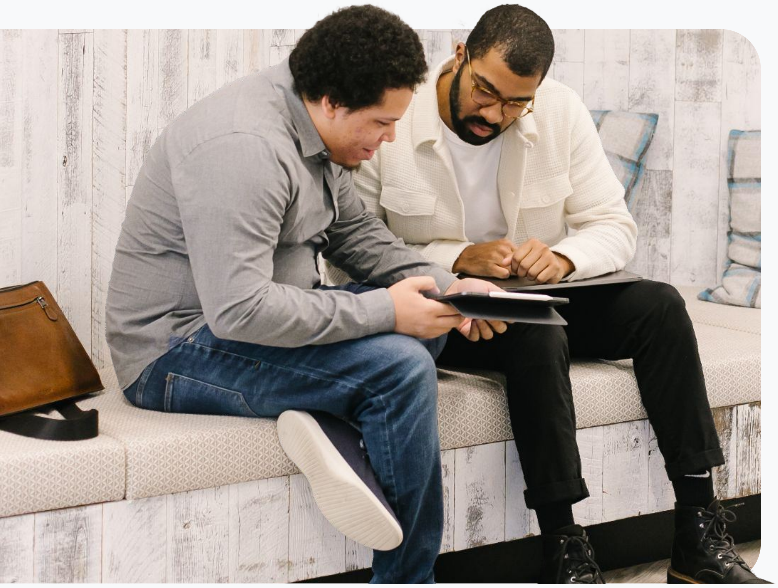
Fleet management: Database administrators and platform engineers can manage an entire fleet of diverse databases using the intelligent dashboards built with AI, proactively assessing availability, data protection, security, and compliance issues without any custom tools or processes. With the integrated 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.

Data governance: Set data policies to improve security, regulatory compliance, and control. Manage all your data, across data silos, in one centralized location. Use built-in data intelligence tools to check data validity and compliance.

Get started with gen AI in your databases

Gen AI is driving a new wave of database modernization—and it is 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:

1. **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.
2. **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, Gemini for Google Cloud can assist with database management, so you don't need to hire specialist database administrators and platform engineers.
3. **Start small.** Make 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.
4. **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.

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.
- Database Migration Service helps simplify migrations from legacy MySQL, PostgreSQL, SQL Server, and Oracle databases.
- 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 generative AI applications, and simplify migrations to AlloyDB with Database Migration Service.
- Spanner is a cloud-native database with virtually unlimited scale, global consistency, and up to 99.999% availability. At peak performance, it processes over four billion queries per second. Migrate from databases like Oracle or 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 and operational workloads. It offers up to 99.999% availability, and processes more than seven billion requests per second at peak, with more than 10 Exabytes of data under management. Migrate from databases like HBase and Cassandra.
- 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 multi-cloud.
- Firestore is a highly scalable, massively popular document database service for mobile, web, and server development. It offers rich, fast queries and high availability up to 99.999%. It has a thriving developer community of more than 500,000 monthly active developers.
- Memorystore offers fully managed in-memory Redis and Memcached service that offers sub millisecond data access, scalability, and high availability. 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.