

Transform your generative AI applications with seamless data processing

Enhance your AI capabilities with Weaviate's vector database on AWS

Efficient data processing is essential to develop generative AI applications, manage multi-tenancy environments, and enable semantic search capabilities. However, the complexity of data types, scalability demands, and resource constraints can make data processing challenging.

Weaviate is an open-source vector database on Amazon Web Services (AWS) that enables you to store and search both vector data and source objects. This simplifies your generative AI application development by removing the need to host and integrate separate databases. With Weaviate, developers of all skill levels can accelerate data processing and build, iterate, and scale AI capabilities faster.

Weaviate overview

Weaviate offers a streamlined approach to building and scaling AI applications. By integrating vector databases with large language models (LLMs) and foundation models (FMs), Weaviate enables the creation of secure and explainable AI for your applications. This helps reduce hallucinations and make conversational apps more reliable, without compromising data privacy.

Weaviate's modules also allow for seamless integration with embedding and LLMs and can handle both indexing and querying in parallel. Weaviate is equipped with built-in hybrid and semantic search capabilities and can manage vector searches at scale.

In collaboration with:



Transforming customer service tasks with generative AI

Neople is a startup that uses generative AI to create digital co-workers, called Neople Assistants, that help businesses achieve their customer service goals. Neople needed more control over the company's data and quick, reliable tools to process customer support cases, without the burden of building and maintaining its own database. This would allow the team to focus on developing features for Neople Assistants.

Neople decided to leverage Weaviate to replace its original Postgres database. With Weaviate's AI-native vector database, Neople gained full control over how it builds AI applications and manages data storage. The company now benefits from significantly faster query results. Weaviate's plug-and-play modules enhance search accuracy and reduce developer workload by removing the need for custom-built components.

Key features

Retrieval-augmented generation (RAG)

Build trustworthy generative AI applications using your own data while maintaining privacy and security. Surface relevant and accurate answers using your LLMs with Weaviate's vector database and search capabilities.

Built-in hybrid search

Improve search experiences by combining vector and keyword techniques to leverage exact term matching and semantic context. Seamlessly deliver contextual, precise results across all of your data in any modality, including documents and images.

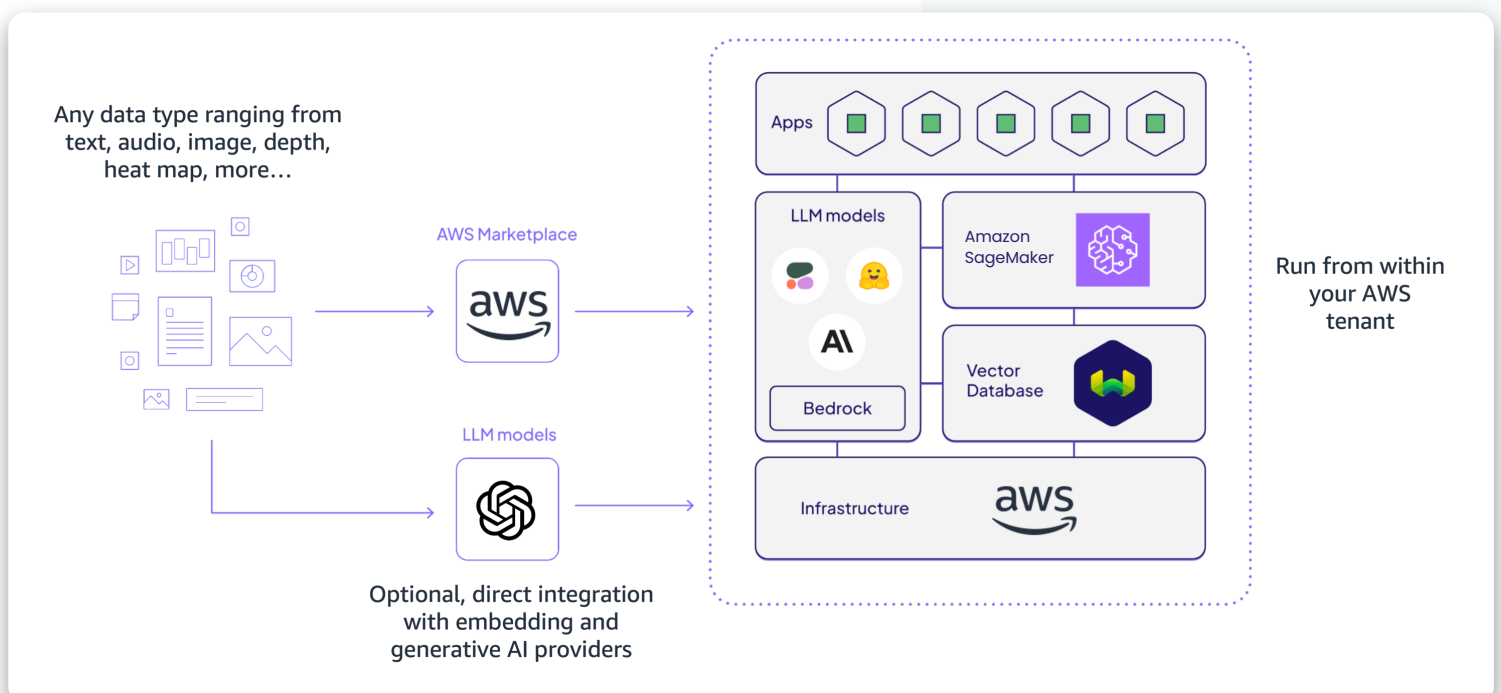
Generative feedback loops

Make your dataset smarter by enriching it with AI-generated answers. Use data produced by generative models to enhance user experiences and automate manual tasks. Improve personalization and spend less time on manual data cleaning.

Developing secure, reliable AI applications fast

FactSet, a financial data company, sought to empower its engineering teams to rapidly build and scale secure, reliable AI applications. To achieve this, FactSet decided to leverage Weaviate's AI-native, open-source vector database for its robust out-of-the-box features, flexible deployment options, strong performance, and reliable support.

Weaviate's modular architecture enabled seamless integration with various data types, sources, and popular LLMs. FactSet also deployed Weaviate on Amazon Elastic Kubernetes Service (Amazon EKS) through a pre-vetted Kubernetes blueprint for internal usage, fulfilling security requirements and allowing teams to move quickly. In less than a year, FactSet successfully commercialized AI, allowing employees to use the company's proprietary chat tool to ask questions about their files.



Outcomes

Increase operational efficiency

Reduce your response time from 10 seconds to under 1 second, achieving 90 percent faster search.¹ Chatbot assistants can find answers quicker than humans, saving time for support teams managing high volumes of customer cases.

Seamlessly scale data storage

Expand the number of data objects in your Weaviate database by a factor of 1000 while improving query response times.¹ This ensures seamless scalability, allowing you to grow without worrying about vector database performance.

Reduce developer strain

Eliminate the need for a custom-built re-ranking mechanism by using Weaviate's out-of-the-box modules. This frees up engineers to focus on improving your offerings and driving innovation.

Featured AWS products



Amazon Bedrock

The easiest way to build and scale generative AI applications with foundation models.



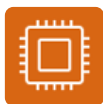
Amazon EKS

Run Kubernetes on AWS without the need to install, operate, and maintain your own Kubernetes control plane.



Amazon SageMaker

Build, train, and deploy ML models for any use case.



AWS Graviton processors

Achieve optimal price performance for cloud workloads.



Amazon Elastic Compute Cloud (Amazon EC2)

Get secure and resizable compute capacity for virtually any workload.



Amazon Elastic Block Store (Amazon EBS)

Get scalable, high-performance block storage resources that can be used with Amazon EC2 instances.

[Learn more ›](#)

Develop and scale AI-powered applications with Weaviate on AWS.
Accelerate your generative AI development with a vector database.

About Weaviate

Weaviate, an AI-native vector database, seamlessly harnesses the capabilities of Amazon Bedrock and Amazon SageMaker, simplifying the development of generative AI and search applications. Weaviate enables users to make sense of vast datasets of unstructured data effortlessly. Weaviate brings powerful vector and keyword search functionalities that scale seamlessly across all data types.

1: [Transforming Customer Service with Generative AI, Weaviate](#)