Solution Brief

Customers Power Modern Generative AI Applications with SingleStore

Introduction

Generative AI is a form of intelligence that discovers complex relationships in large sets of training data, collectively called a corpus, and generalizes from what it learns to create new data. From customer support chatbots and conversational AI assistants to facial recognition and video surveillance, gen AI addresses a myriad of use cases.

Generative AI Applications Powered by SingleStore Today

While many solutions providers are just getting up to speed with vector data, and the most prominent single-function vector database startups have existed for only a few short years, SingleStore was founded in 2011 and originally built vector capabilities into its core data platform in 2017 to meet the needs of one of the world's largest financial institutions. SingleStore is a real-time, distributed SQL database with a unique rowstore/columnstore architecture that natively manages transactions (OLTP), analytics (OLAP), vector, and all relevant data models such as geospatial, JSON, time series and more in a single platform. Using SingleStore instead of a purpose-built vector database delivers:

- **Fast vector capabilities** with exact next neighbor match
- The ability to **join vector data with all relevant data types** with a single SQL query
- The ability to **rapidly ingest and process** data with no ETL, driving the millisecond response times
- The ability to effectively manage **multiple data types** in a single data platform
- **Enterprise-grade capabilities** including HA, DR, PITR, multi-AZ failover and more
- Flexibility to deploy in the cloud, on-premises or in a hybrid environment
- Substantial **data compression** (often 70-90%) to drive significant TCO savings

SingleStore supports vector similarity search using **dot_product** (for cosine similarity) and **euclidean_distance** functions, and these can be leveraged in SQL queries to perform similarity calculations efficiently on large volumes of vector data. Today, these functions are used extensively by SingleStore customers for applications including semantic search, chatbots, facial recognition, image matching, and more. Following are examples of a few of the many customers using SingleStore to power generative AI use cases today:

Adobe

Martech/Adtech

Real-time Resource, Content, and Campaign Management in its Workfront Product

Adobe is changing the world through personalized digital experiences. Adobe solutions orchestrate entire marketing campaigns, accelerate end-to-end workflow, and measure results and ROI.

Adobe is using SingleStore's vector database capabilities to drive generative AI functionality in its Workfront product, including:

- An AI Assistant that generates campaign content
- An Assignee Recommendation Engine to allocate optimal resources to tasks
- A Smart Form Generator that recommends the ideal forms to enhance workflows

Adobe's original data infrastructure, built on PostgreSQL, routinely delivered 5-minute query results and timeouts, offering limited analytical capabilities and poor customer experience. The global martech/adtech leader next tried Elastic for analytics, which yielded fast but not exceptional performance. Any benefits were outweighed by complex data flattening and unsatisfactory price/performance; onboarding new customers required too much memory.

Adobe considered Snowflake and Databricks, but their slow ingestion and query performance were not built for Adobe's demanding, high-concurrency environment, which supports more than 3,600 enterprise customers and 500,000 monthly active users with 60-70 terabytes of data, 250+ tables and 10 billion records that must be queried and joined. Adobe's business is growing 20% per year, and its data infrastructure could not even effectively support existing workloads.

Adobe had targeted 5-second ingest and query latency to deliver good user experiences. With SingleStore, times are a fraction of that: 1-2-second ingest, most queries complete in <1 second and even the most complex queries yield results in 1-2 seconds, which is 300x faster than before. Adobe has truly democratized data: any user can build and refine reports in SingleStore, and users can see results immediately as they build. Along with screaming-fast system performance, it has also attained massive price performance: 50% lower TCO.

Learn more:

Video: How Adobe built its user-facing app, Workfront, with SingleStore

SIEMENS

Industrial Manufacturing & Automation

Analytics as a Service

Semantic Search on Survey Responses for Siemens Pulse Analytics

Siemens is driving sentiment analysis and semantic search using vector capabilities within SingleStore to analyze and gain deeper insights into the responses from company-wide HR surveys across its 200,000 employees worldwide. Siemens Pulse Analytics gathers data across many sources including employee and customer surveys, and learning data (supplemental courses attended by employees).

Its previous data infrastructure based on Microsoft SQL Server limited Siemens to searching for exact "text-based" search terms to extract insights from its vast collection of feedback data sources. By creating vector embeddings of the text feedback using OpenAI's embeddings API and storing the resulting embeddings in SingleStore, Siemens and its customers can now perform semantic search, gaining deeper insights into employee satisfaction, company culture, and retention.

Learn more:

Blog post: AI-Powered Semantic Search in SingleStore

Impact story: SingleStore Delivers 10-100x Performance Gains to Help Siemens and its Customers Make Smart Decisions in Real Time

RightSense

Role- and function-based KPIs

Vector Similarity Search via dot_product Rings Up Actionable Insights for Retailers

RightSense.ai was founded on a mission to democratize data. The goal: to help teams, technical and business users alike, take advantage of the data their organizations already have, and combine it with third-party data, to drive real-time insights. RightSense offers Automated Data Stories (e.g., analysis of customer segmentation, sales trends or inventory shortfalls) integrated with real-time chat powered by large language models such as GPT-4 to offer dynamic, actionable insights for specific organizational KPIs.

RightSense's current customer base is concentrated mainly in retail, which means it needs an agile data architecture to quickly add large networks of stores. RightSense originally was using AWS Redshift but found its performance unacceptable. The team then tried MariaDB, but found that as it started adding more customers, stores, and users, it had to rework everything. RightSense could not effectively support its current data workloads, much less its 25% YoY forecast growth.

RightSense’s FUSION platform runs in SingleStore on AWS. Customers have their own KPI data lake in the RightSense Cloud and all data is isolated from other customers. If they have JSON, Excel, CSV or other source files, they can upload those directly, and SingleStore Pipelines perform millisecond ingest of Parquet files into SingleStore. SingleStore uses real-time Retrieval-Augmented Generation (RAG) on data from all sources to enrich the data with context. FUSION generates vector embeddings on uploaded data and stores them in SingleStore. FUSION uses SingleStore’s vector similarity search using dot_product to perform semantic search. Automated Data Stories are automatically generated by the LLM based on KPIs and tailored by role, e.g., Store Manager, Regional Manager or In-Store Sales.

Before RightSense, retailers depending on other systems were forced to wait 1-2 days to get a reading on their business. With RightSense, powered by SingleStore on AWS, retailers can now get a finger on the pulse of their business in real time, or more than 86,000x faster. RightSense customers can measure the effectiveness of promotional campaigns faster, resulting in a 15-20% increase in their average return on campaign spend.

SingleStore-powered TCO savings accrue to RightSense, which can now operate its fast-growing business on a single database, and to its customers, who save on average an equivalent of 1-2 FTE with RightSense. With SingleStore, RightSense can now confidently serve large customers knowing it does not face data size limitations, seamlessly scaling to support current and future customer data volumes.

Learn more:

[Coming soon] Impact story: RightSense Uses SingleStore Vector Similarity Search via dot_product to Ring Up Actionable Insights for Retailers

THORN

Non-Profit
#dataforgood

Real-Time Facial Recognition to Fight Child Sexual Abuse

Thorn is a nonprofit that builds technology to defend children from sexual abuse. Thorn creates products that identify child victims faster, provides services for the tech industry to play a proactive role in removing child sexual abuse material (CSAM) from their platforms, and works directly with youth and their communities to empower them to prevent abuse.

Thorn uses SingleStore’s vector search capabilities based on artificial intelligence and machine learning in real time to help identify and find missing and exploited children so law enforcement authorities can find and help them faster.

Learn more:

Impact story: Ashton Kutcher Races to to Help Thorn Defend Children

Forbes Podcast: Ashton Kutcher on Thorn, SingleStore and the Ethical Responsibility of Tech Leaders | hosted by SingleStore CEO Raj Verma directly apply

Careers

Catalog, Customer 360 and Personalization Matching Job Seekers with Roles

DirectlyApply is a job discovery platform and vertical search engine that connects job seekers with employers. The platform fetches thousands of vacancies from job sites, portals, and career sites, intelligently filtering the results to provide a curated list of up-to-date opportunities at real employers that are actively hiring. It currently has more than 30 million job listings and six million distinct titles.

DirectlyApply is using SingleStore's vector capabilities to perform semantic search for matching job titles with openings. It stores vector embeddings generated from job titles, and performs vector similarity search (using dot_product), to match job openings with the more than 3,000 standard International Standard Classification of Occupations (ISCO) job titles. The company uses multiple vector models, comparing OpenAI vector models to its own models trained with Tensorflow, and also uses SingleStore's spatial data models.

Learn more:

Impact story: DirectlyApply Saves $100K/Year while Helping Job Seekers Find Opportunities 400X Faster with SingleStore on AWS

Lumana

Security

AI-Driven Video Monitoring + Surveillance for Safety and Security

Lumana is a SaaS platform provider that offers a visual intelligence platform for real-time video monitoring and surveillance driven by vector similarity search.

Lumana's previous data infrastructure built on PostgreSQL was too slow for real-time analytics use cases. The company considered a purpose-built vector database, Pinecone, but Lumana has to do a lot of other filtering, tagging, and joins in addition to vector search, so it chose SingleStore to power its video monitoring platform.

The Lumana solution is a multi-tenant SaaS app with more than 10 customers in production; it trains its own models for image recognition and derives vectors from that. Lumana uses SingleStore's vector functionality to perform image matching and video monitoring for a variety of use cases, including:

- **Occupational safety** to alert if, for example, someone in a manufacturing facility or distribution center is at risk because they are standing too close to a forklift
- **Security and surveillance** to identify, for example, if a blue Tesla with a license plate that contains the numeral 9 drove past a given location

Other customers using SingleStore vector capabilities today to support generative AI apps include:

- **LiveRamp**, which is exploring upgrades to entity resolution with its own vector embedding model
- **Outreach**, which discussed its current implementation and its plans to use SingleStore to drive its generative AI functionality at Gen AI Dev Day & Happy Hour @ IBM Toronto
- **Cognism, Datakubes, EduMe, Instantly, MonitorBase, Numerade, OpenDialog, Orca Security, POSCO** and others

To learn more about SingleStore for generative AI:

Blog post: Why Your Vector Database Should Not be a Vector Database

Blog post: Getting Started with OpenAI Embeddings Search + SingleStore

Webinar: Build a ChatGPT App on Your Own Data

Contact us today to discuss generative AI or any of your data needs +1 855.463.6775 or team@singlestore.com