BARC

More than Vectors: How Multi-Faceted AI Databases Enable Smart Applications

Author: Kevin Petrie

Publication: July, 2024

Abstract

This research note explores the emergence of versatile AI databases that support multi-model applications. Practitioners, data/AI leaders, and business leaders should read this report to understand this new platform option for supporting modern AI/ML initiatives.

Research sponsored by:

Vespa

Table of Contents

Introduction

Innovative companies across industries are investing in artificial intelligence to increase efficiency and gain competitive advantage. Their initiatives are often multi-faceted, with generative AI complementing outputs from natural language processing or predictive machine learning models. Many early adopters build smart applications and workflows that incorporate two or more AI/ML models. As they do so, they need more than just a data warehouse or vector database.

This report explores the emergence of versatile AI databases that make those smart, multi-model applications successful. It describes AI databases' functionality, must-have characteristics, benefits, and use cases. Practitioners, data/AI leaders, and business leaders should read this report to understand this new platform option for supporting modern AI/ML initiatives. The need for multi-purpose, scalable AI databases will only increase as GenAI-driven experiments and pilots move into production.

To understand the need for these AI databases, let's consider the evolution of data analytics. Three technological waves have shaped this market: business intelligence, artificial intelligence/machine learning, and generative AI (GenAI). (See figure 1).

Figure 1: Evolution of analytics and AI

- **Business intelligence**. In the 1990s, analysts started measuring historical company performance and market trends based on periodic assessments of database records. They used BI tools to build reports and dashboards that inform operational business decisions.
- **Artificial intelligence/machine learning**. In the 2010s, a rising number of data scientists adopted AI to assist decisions and automate actions. They trained ML models to classify, predict, and recommend future outcomes based on historical patterns in diverse datasets.
- **Generative AI**. OpenAI triggered a boom in generative AI with its release of Chat-GPT 3.5 in November 2022. Data science teams and developers are now building applications that create content to enrich functions such as customer service and document processing.

Each phase complements rather than replaces what came before. Building on this trend, most early adopters use GenAI language models alongside other AI/ML models or analytical functions. And many BI tools now include GenAI language models that help analysts query data, visualize outputs, interpret

results, and so on. Multifaceted applications such as these need a multifaceted data platform that manages and delivers all types of data, including structured tables, semi-structured log files, and unstructured text.

Most early adopters use GenAI language models alongside other AI/ML models or analytical functions

Enter the AI database

The AI database can help. This emerging type of platform manages objects such as tables and documents, as well as vectors that assign numerical values to chunks of unstructured data. These chunks might be groups of words, images, audio clips, or video segments. The AI database can apply one or more various AI models to each piece of data, whatever the format, and combine multiple signals to create more accurate AI outputs. For example, a GenAI chatbot might calibrate its text outputs based on the results of a machine learning or natural language processing model. By consolidating models and data types in one multi-purpose platform, the AI database improves computing efficiency and scalability compared with disparate single-purpose platforms.

Three categories of vendors are converging on the AI database market: lake houses, data warehouses, and vector databases. Lake houses and data warehouses are adding vector and text capabilities, while vector databases are adding support for text and tables. For example, Vespa combines text, table and vector database capabilities into one platform.

The AI database market includes lake houses and data warehouses that are adding vector and text capabilities; and vector databases that are adding support for text and tables

The AI database delivers data to AI models, which in turn contribute to smart applications as part of an AI stack. (See figure 2.)

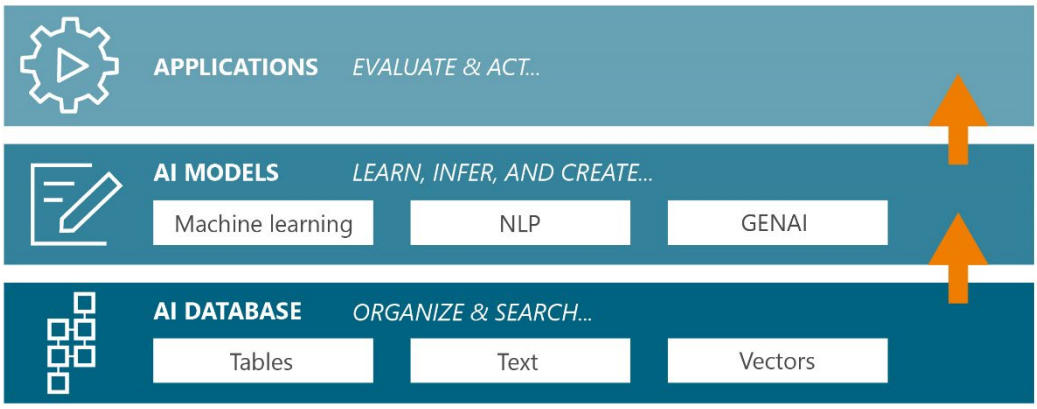

Figure 2: AI Database within the AI Stack

The AI database organizes and searches this multi structured data as described below.

- **Organize**. The AI database organizes its data, for example by placing similar vectors close to another based on assessments of their metadata. It also governs data with role-based access controls, encryption, masking of personally identifiable information (PII), and audit logs that support compliance reporting. This improves accuracy and reduces risk.
- **Search**. The AI database also searches tables, text, and vectors. It queries tables (still the most popular AI input) to find or derive specific values; finds document passages that match keywords; and runs similarity searches on vectors. It then selects and makes inferences in the data using various AI models.

The AI database organizes and searches tables, text, and vectors

AI Models

AI databases support three primary types of AI models: machine learning, natural language processing, and GenAI.

- **Machine learning**. ML uses techniques such as classification or clustering to find patterns in historical data, then predicts events or trends based on those patterns. ML models also identify anomalies and recommend actions. In the context of the AI database, ML models primarily select tables, text, or images for humans, GenAI language models, or other subsystems to consume.
- **Natural language processing**. NLP interprets and creates speech or text to assist tasks such as translation, sentiment analysis, or document summarization. These models primarily consume text files.
- **GenAI**. GenAI language models generate text, imagery, audio, or video based on what they learn from a corpus of existing content. For example, they predict the next word or phrase in a string of text based on the statistical inter-relationships of words in the training corpus.

These various AI/ML models learn patterns, make inferences, and create outputs based on what they receive from the AI database – in fact, the AI database often hosts and runs the models. In some cases, the models are embedded within applications.

Various AI/ML models learn patterns, make inferences, and create outputs based on what they receive from the AI database

Applications

Smart applications evaluate, select, and assign AI models to assist various tasks with inference. They feed the selected model the appropriate data, for example by injecting it into the user prompt of a GenAI language model as part of retrieval-augmented generation (RAG). Then they act on the outputs by sending offers to customers, tailoring website content for visitors, suggesting draft contract language to business partners, and so on. These applications execute chains of interrelated events: one model's output might trigger inference by another model, which in turn triggers a message to the human user.

Smart applications evaluate, select, and assign AI/ML models to assist various tasks

More than a vector database

The AI database is more than a vector database because it organizes and searches more types of data to support various AI/ML models. Table 1 summarizes the key differences between these two platforms.

	Vector Database	AI Database
Data types	Vectors	Vectors, text, tables,...
Ranking / data selection	Vector similarity	Combined signals: keyword matches, SQL queries, ML modeling
Model types	NLP; predictive ML, GenAI	
Text processing	NA	Keyword searches, ranking, linguistic processing

Table 1: Comparison of Vector and AI Databases

And this multi-faceted database approach contributes to larger AI environments that include various complementary technologies. Recent research by BARC and IT Market Strategy found that 20% of survey respondents have vector databases in production, and another 26% are in the testing/POC stage. AI databases also can serve as feature stores, with 24% and 25% adoption rates respectively, and support other technologies such as knowledge graphs shown as shown in Figure 3.

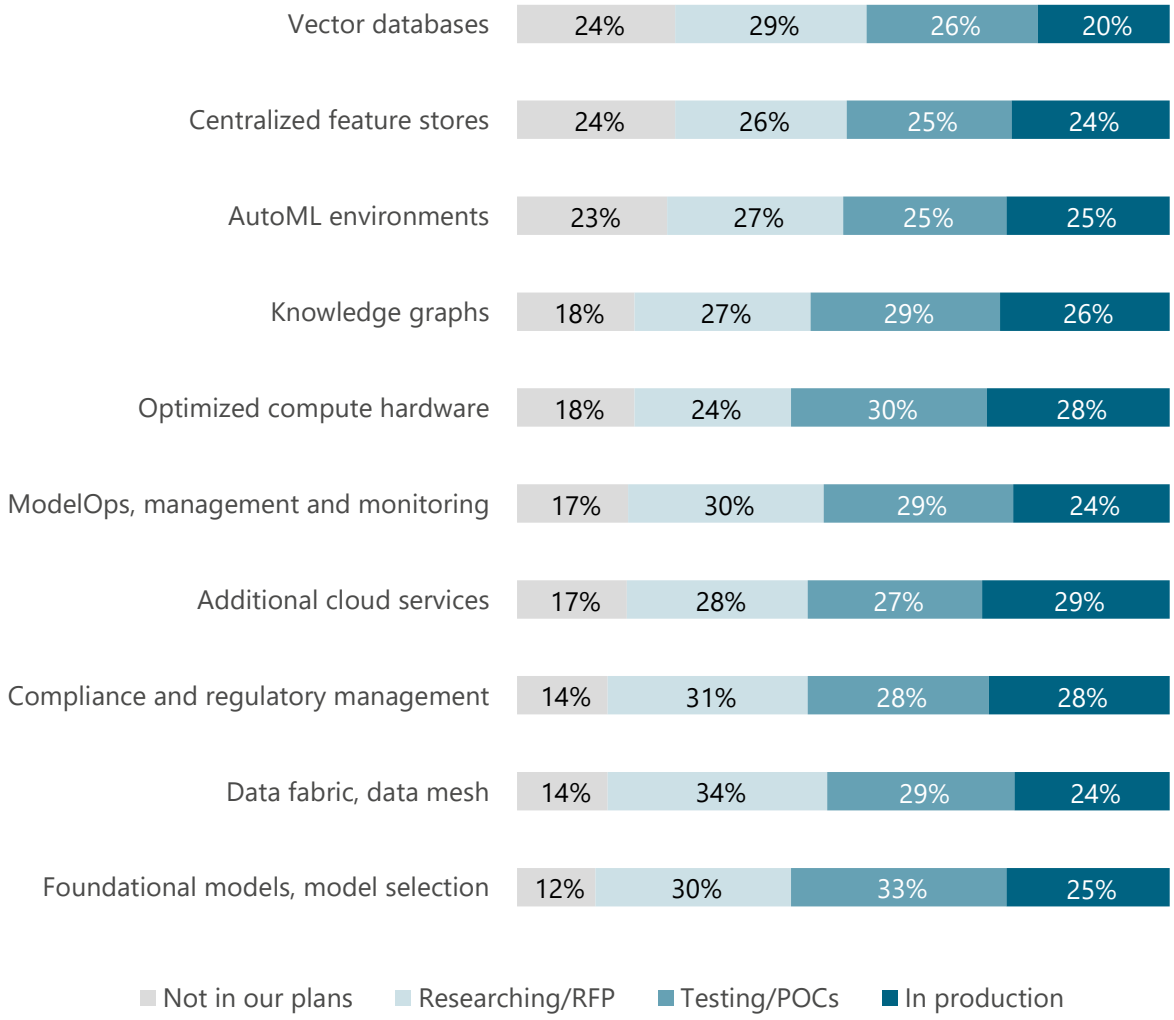

Figure 3: Status of AI Technologies in Existing Environments (n-298)[1]

[1] Optimizing Your Architecture for AI Innovation, Shawn Rogers (BARC) and Merv Adrian (IT Market Strategy), March 2024.

Must-have characteristics

Successful AI/ML initiatives require an AI database that is multipurpose, governed, open, intuitive, fast, and scalable. (See figure 4.)

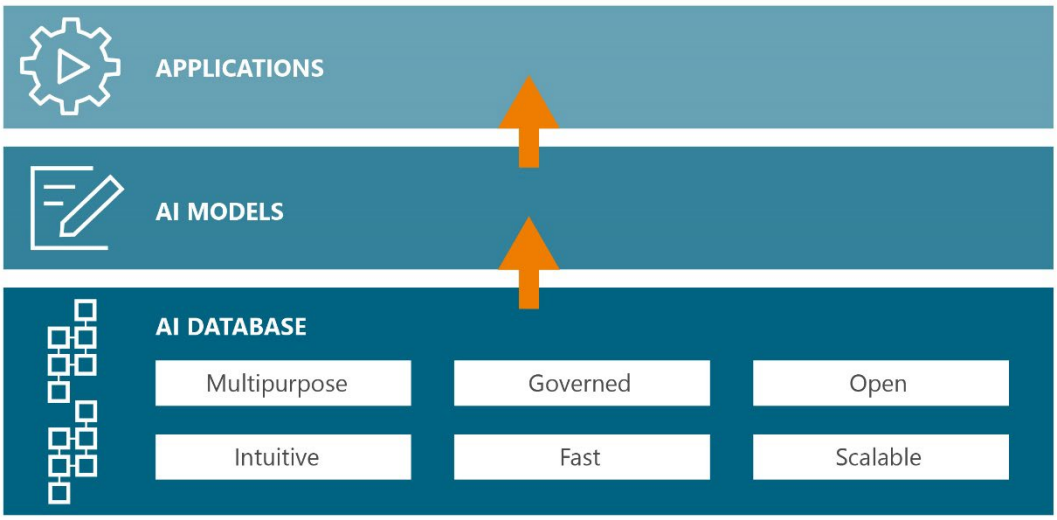

Figure 4: Must-Have Characteristics of the AI Database

- **Multipurpose**. With the rise of GenAI, mainstream companies now want to analyze text in addition to traditional tables and logs. A multipurpose AI database delivers data of any type – structured, semi-structured, or unstructured – to ML and NLP models as well as GenAI.
- **Governed**. AI can threaten business operations, privacy, and compliance. A governed AI database mitigates such risks by enabling data teams to oversee and control user access, data quality, and application tasks that consume its outputs.
- **Open**. Data scientists and developers tap many commercial and open-source elements, including libraries, tools, APIs, languages, and formats, as they build AI applications and workflows. An open AI database provides unfettered access to this ecosystem.
- **Intuitive**. Busy data teams don't have the time to learn complex new procedures. An intuitive AI database simplifies processes to reduce skill and training requirements, which rank as the top obstacle to AI success according to a recent survey report by BARC.
- **Fast**. Use cases such as fraud prevention, security scanning, and customer recommendations depend on real-time processing. A fast AI database meets service level agreements (SLAs) for latency, throughput, and concurrency.
- **Scalable**. As GenAI experiments and pilots move into production, scalability becomes a must-have characteristic. An AI database must automatically scale to support many complex models as they drive up data throughput and user requests per second. And to stay within budget, it must give users insight into and control over the associated compute costs.

Successful AI/ML initiatives require an AI database that is multipurpose, governed, open, intuitive, and fast

Benefits

AI databases that have these characteristics help simplify AI projects by consolidating infrastructure and reducing the need to integrate multiple platforms. This speeds project execution and frees up resources to support additional projects. It also reduces operational risks because each project depends on fewer platforms. Benefits like these improve revenue and costs, boosting company profits and AI's return on investment (ROI). They enable companies to realize the promise of AI: streamlining operations and enriching user interactions.

Use cases

Now we explore how the AI database enables three example use cases: customer service, document processing, and supply chain optimization. Each use case applies a combination of ML, NLP, and GenAI to solve a business problem. (See figure 5).

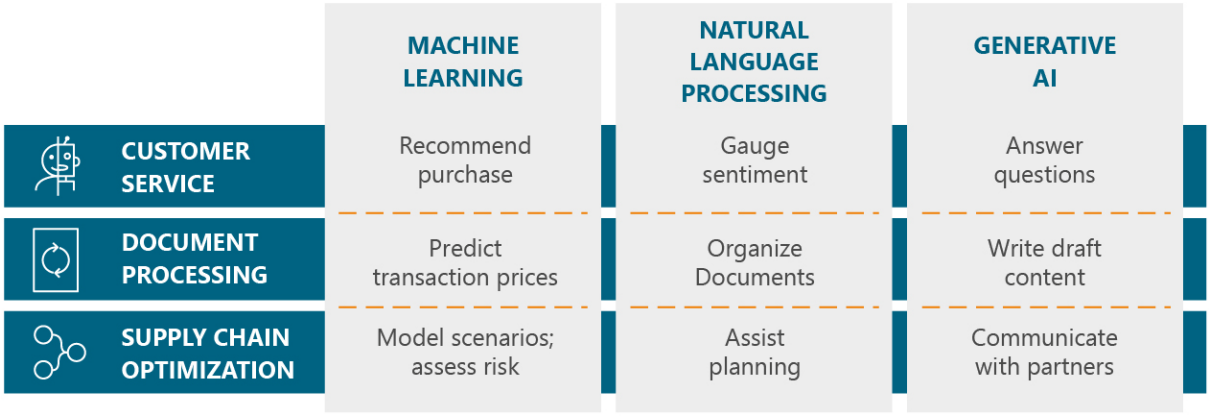

Figure 5: **Use Cases for AI Databases**

Each use case applies a combination of ML, NLP, and GenAI to solve a business problem

Customer Service

Let's first consider a toy retailer that struggles with disappointing web sales and a high percentage of abandoned shopping carts on its website. Working with product managers in the ecommerce division, this retailer's data science team determines that most of the abandoned carts contain toys for toddlers and infants. These toys perform well for in-store sales to first-time parents, many of whom ask sales personnel for advice about items' relative popularity and safety risks. Based on these observations, data scientists engage developers to build a new ecommerce application. This smart application features a chatbot that invites user questions, provides responses, and recommends toys to buy.

The application also contains a GenAI language model that converses with the customer, an NLP model that ranks customer sentiment, and an ML model that recommends a product based on these inputs. The AI database supports each of these activities, starting with retrieval augmented generation (RAG) for the GenAI language model. When the user asks a question, the AI database retrieves relevant sales records from its tables, along with product documentation and service call summaries. It masks PII within this content, then delivers the anonymized content to the application so it can augment the user prompt.

When the user asks a question, the AI database retrieves relevant sales records from its tables, along with product documentation and service call summaries

While the GenAI language model answers user questions through the application chatbot, the AI database handles queries from other models. It finds and retrieves text that helps the NLP model gauge the sentiment of the customer. Is it a concerned parent that needs human advice? If so, the NLP model flags this, and the application routes the conversation to a live customer service representative. The AI database also finds and retrieves relevant purchase histories from its tables, which in turn helps the ML model recommend a specific product for the customer to purchase.

The application, models, and AI database analyze all this data together to help the ecommerce team increase online sales for this product segment.

Document processing

Now we consider an insurance company that struggles with customer churn in its auto policy division. Exit surveys and third-party review sites indicate high levels of dissatisfaction with delays in claims processing. The general manager assigns a data scientist to help claims managers streamline these operations. Working cross-functionally, they identify a major bottleneck: negotiating the price of reimbursements for accident damage. They collaborate with in-house developers to build a new application that organizes relevant documents, predicts repair prices, then drafts letters for claims managers to send to customers and auto service providers.

The application uses NLP to organize and summarize documents, ML to predict prices, and GenAI to draft letters, all based on inputs it receives from the AI database. The AI database uses vector search and keyword matching to pull up documents relating to a given accident, including the car warrantee, service history, and so on. The NLP model organizes these documents, highlighting and summarizing

key passages for the claims manager to review. Meanwhile the ML model, using table columns and records as features, predicts the appropriate market value for repair services. The GenAI language model then drafts letters containing the assessment price. The claims manager reviews this draft, then revises, approves, or rejects it based on their professional judgment.

The AI database uses vector search and keyword matching to pull up relevant documents

Assisted by its smart new application, this insurance company increases the accuracy of its assessments and resolves claims faster, which in turn improves customer satisfaction and reduces churn.

Supply chain optimization

Our final example focuses on supply chain optimization for a container shipping company. Increasingly frequent hurricanes force this company to re-route ships, which leads to cascading delays and issues throughout the supply chain. The chief operating officer forms a tiger team of data scientists, logistics managers, and partner managers to build more flexibility into the system. These team members identify the biggest point of pain: Panama Canal closures during the months of August and September. They determine that if they're able to rapidly re-route container ships from Europe through the Arctic on their way to Asia, they can reduce the hit to operating costs.

With this knowledge the team builds a smart application for supply chain optimization. The application uses ML to predict hurricane disruptions and recommend re-routing plans. It uses NLP to assist logistics research processes, and GenAI to generate real-time updates for supply-chain partners. The AI database supports all this by delivering weather table records to the ML model, research documents to the NLP model, and vectors of partner conversations to the GenAI model. Armed with these inputs from the AI database, the smart application enables data scientists and logistics managers to re-route disrupted ships over the North Pole, reducing delays and operating costs during hurricane season.

The AI database delivers tables to the ML model, documents to the NLP model, and vectors to the GenAI model

Conclusion and next steps

Many early adopters these days start with GenAI chatbots. But the most compelling business opportunities for artificial intelligence require diverse families of models that together deliver more value than they would on a standalone basis. And those model families need a solid foundation of accurate, governed, multi-structured datasets. As enterprises realize this, they will turn to AI databases as a viable and scalable option for streamlined data management and delivery. Data, AI, and business leaders should take the following steps to evaluate the role of AI and AI databases in their organization.

- **Prioritize AI use cases.** Create an inventory of your use cases based on conversations with business managers, then stack-rank them. Low-risk, short-term use cases should get highest priority. This might mean you favor a use case that focuses on predictive ML and traditional tabular datasets.
- **Consider your data platform options**. Enlist with your data scientists and data engineers to evaluate how well your existing data platform – perhaps a data warehouse or lake house – will support an AI initiative that focuses on the higher-priority use cases. Can it provide the right data inputs in the right format within reasonable SLAs? If not, consider opting for an AI database.
- **Evaluate AI database offerings**. You need a scalable AI database that is multipurpose, governed, open, intuitive, and fast. Evaluate how well each candidate product meets these requirements during vendor demonstrations and ideally a proof of concept.

About BARC

BARC (Business Application Research Center) is one of Europe's leading analyst firms for business software, focusing on the areas of data, business intelligence (BI) and analytics, enterprise content management (ECM), customer relationship management (CRM) and enterprise resource planning (ERP). Our passion is to help organizations become digital companies of tomorrow. We do this by using technology to rethink the world, trusting databased decisions and optimizing and digitalizing processes. It's about finding the right tools and using them in a way that gives your company the best possible advantage. This unique blend of knowledge, exchange of information and independence distinguishes our services in the areas of research, events and consulting.

Research

BARC studies are based on internal market research, software tests and analyst comments, giving you the security to make the right decisions. Our independent research brings market developments into clear focus, puts software and vendors through their paces and gives users a place to express their opinions.

Events

Decision-makers and IT industry leaders come together at BARC events. BARC seminars in small groups, online webinars and conferences with more than 1,000 participants annually all offer inspiration and interactivity. Through exchange with peers and an overview of current trends and market developments, you will receive new impetus to drive your business forward.

Consulting

In confidential expert workshops, coaching and in-house consultations, we transform the needs of your company into future-proof decisions. We provide you with successful, holistic concepts that enable you to use the right information correctly. Our project support covers all stages of the successful use of software.

BARC

About Vespa.ai

Vespa.ai is the creator of Vespa, a collaborative platform for developing real-time AI-driven applications for search, recommendation, personalization, and retrieval-augmented generation (RAG).

Vespa efficiently manages data, inference, and logic, supporting applications with large data volumes and high concurrent query rates. It is available as a managed service and open source. Designed for low latency and scalability – typically over 100K queries per second – Vespa is the preferred solution for large-scale systems like Spotify, Wix, Capital One, and Yahoo.

Contact info

Vespa.ai

info@vespa.ai

BARC

Data Decisions. Built on BARC.

www.barc.com

Germany
BARC GmbH
Berliner Platz 7
D-97080 Würzburg
+49 931 880651-0

Austria
BARC GmbH
Hirschstettner Straße 19 / I / IS314
A-1220 Wien
+43 660 6366870

Switzerland
BARC Schweiz GmbH
Täfernstr. 22a
CH-5405 Baden-Dättwil
+41 56 470 94 34

United States
BARC US
13463 Falls Drive
Broomfield, CO 80020
USA