

Scaling Intelligence: Unleashing the potential of GenAI with vector search

RESEARCHED BY

MDIA

COMMISSIONED BY

ORACLE

Introduction

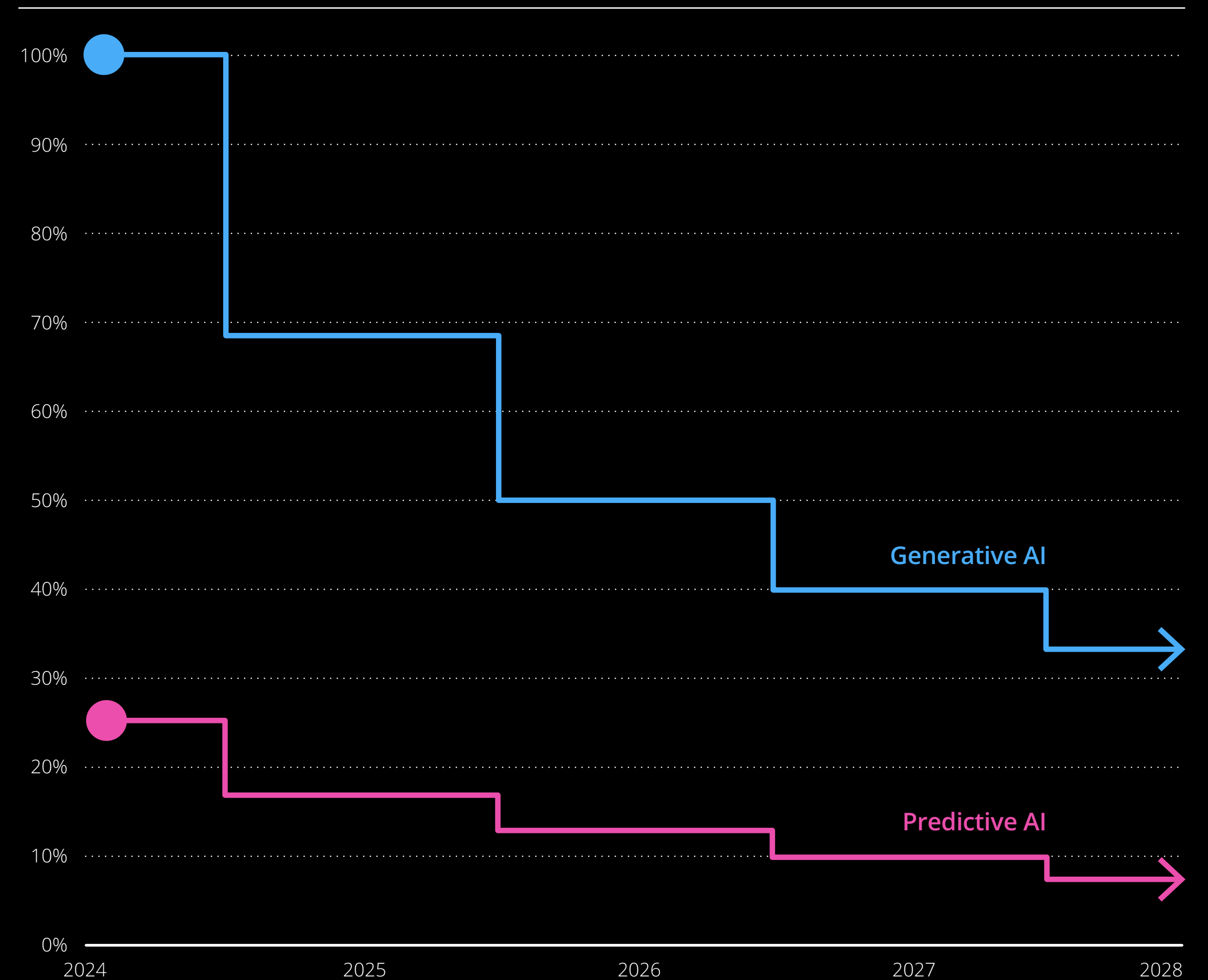
The era of generative AI (GenAI) is upon us, right here, right now. Every day brings new GenAI startups, LLMs, chatbots, and assistants, along with an even more diverse assortment of supportive tools, each vying for attention among enterprise practitioners and promising to elevate the value of GenAI across the business.

Even in the enterprise, where radically new technologies are usually met with some skepticism and restraint, GenAI has already found a welcome home. It's not seen as a new play toy but instead as a fully integrated member of the IT family, one poised to outpace traditional, predictive AI spending among enterprise adoptees in terms of growth over the next five years (see Figure 1). In other words, GenAI has already reached mission-critical status among companies that shape the world's economy.

But GenAI isn't just about models, bots, and tools. Ultimately it's about data. Even with the latest and greatest ChatGPT-scale model at the ready, enterprises must figure out how to use high quality, private data as part of the GenAI workflow. To illustrate, according to an Omdia study of more than 1,600 enterprise AI practitioners, the top area for investment is data quality, security, and privacy (see Figure 2).

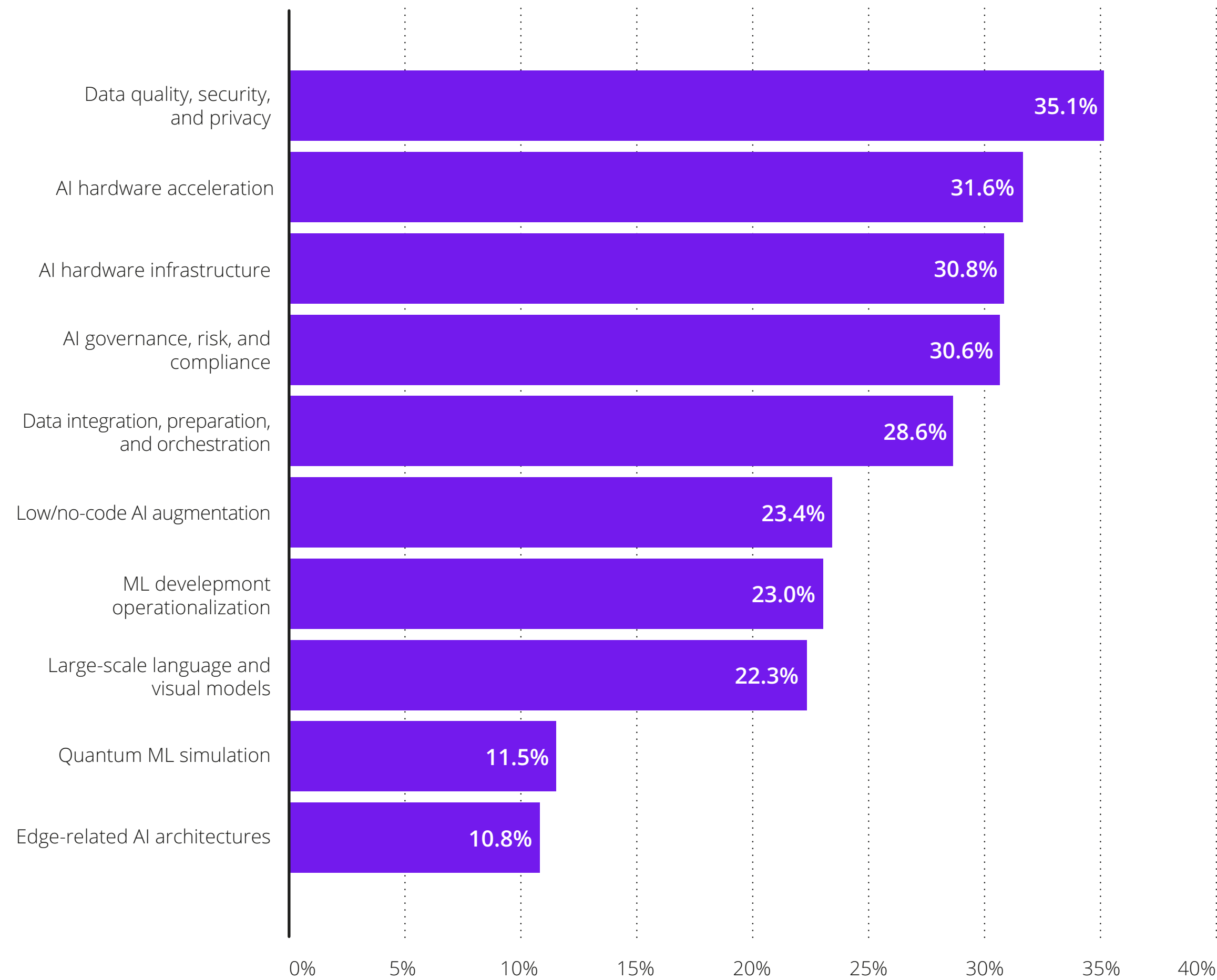
Figure 1: GenAI investments are mission-critical for big business

AI SOFTWARE REVENUE ANNUAL GROWTH FORECASTS



SOURCE: OMDIA, ARTIFICIAL INTELLIGENCE SOFTWARE MARKET FORECASTS – 2H23 DATA

Figure 2: The growing concern over data quality as a top risk to AI
AI INVESTMENT PLANS



SOURCE: OMDIA, IT ENTERPRISE INSIGHTS: IOT, CLOUD, AI, 5G, AND SUSTAINABILITY – 2024. N = 1,625

As borne out over the past two years of innovation, enterprises have adopted a two-pronged approach to leverage their data with GenAI technologies:

- Using semantic search to rapidly locate and deliver highly relevant enterprise information based on the meaning of stored data and requests. Examples include recommender systems that suggest similar products to online shoppers, quality assurance systems that find flaws in objects by analyzing their images, and fraud detection systems that identify anomalous financial transaction patterns.
- Greatly increasing the relevance and accuracy of LLM responses by supplementing them with semantic search of private or domain-specific information. This approach typically involves the use of in-prompt techniques such as retrieval augmented generation (RAG) that provide additional content and context to LLM queries, so they generate better results.

The recent rise of vector databases has made it easier for enterprises to implement both approaches, enabling them to both search huge swaths of unstructured corporate information and infuse LLMs with corporate data. Searching data based on its semantic value can accelerate the generation of insights, improve the quality of GenAI results, and categorize tremendous amounts of information at speed.

But getting the most value out of a vector database can be a tricky business, especially at scale. As outlined in the remainder of this report, while there are many single-purpose vector databases that enable search and deliver value, enterprise buyers seeking an edge in performance, security, and simplicity, should consider a converged, multi-model database, such as Oracle Database, that can natively store and search vector data alongside corporate data models such as graph, time series, and geospatial data.

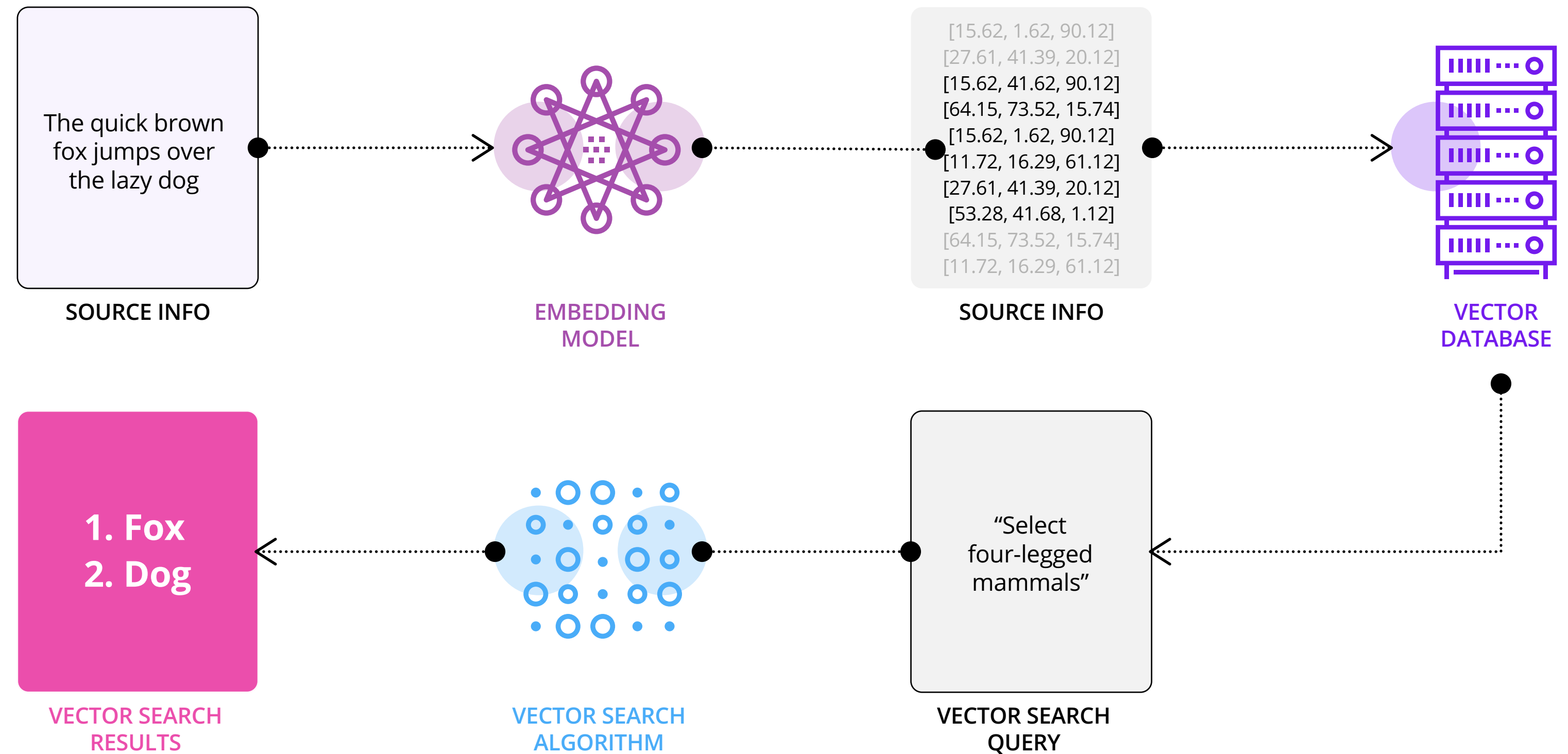
Enter the vector database

From a mathematical perspective, vectors are just numbers, often a very long list of numbers that represent a given piece of information. Often found in scientific and technical applications, vectors can be efficiently compared to identify if a given vector is similar to or dissimilar from other vectors.

In semantic search, vectors are used to represent the “meaning” of a given piece of information which could be text, images, audio, video, or other types of data. Vectors are stored, processed, and searched in what’s known as a vector database. In querying a vector database, users search that database to find the stored vectors (often referred to as embeddings), and hence the input data, that are most similar to an input value. Semantic search helps users gain an understanding of the context and relationships among similarly clustered vectors and by extension the information those vectors represent (see Figure 3).

These vector-based semantic search engines now power many of the recommender systems that

Figure 3: Vectors as numerical representations of various objects



SOURCE: OMDIA

consumers encounter online while viewing media or shopping. Likewise, within the enterprise, they now power information retrieval and document discovery capabilities within apps ranging from customer support to experience forums and human resource management systems.

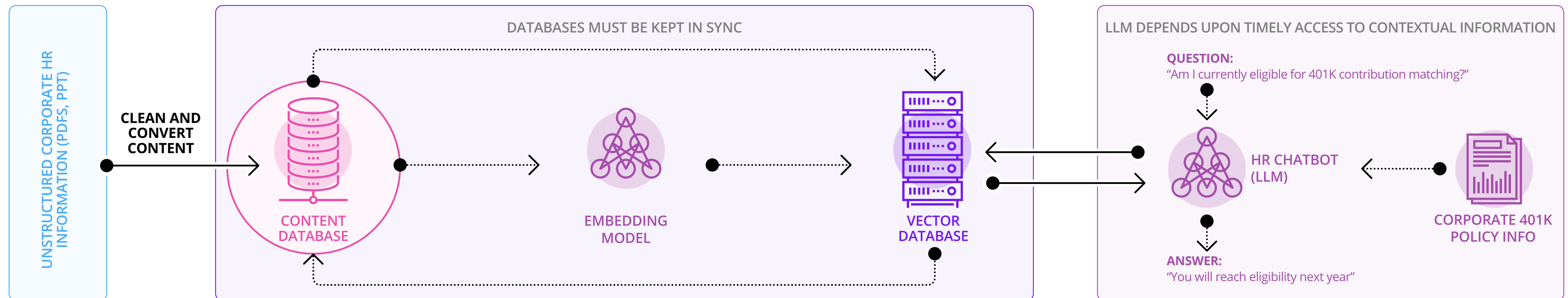
More recently, enterprises have started using additional GenAI technologies such as RAG to feed the required context and reliable facts into an LLM so it can more accurately respond to a given user's prompt. Enterprises are finding that RAG-enabled applications

are becoming crucial because most LLMs, even those that have been fine-tuned to work within a given context such as customer support for a given company, can generate erroneous answers (referred to as hallucinations) when pressed to return knowledge not found in the LLM's training and fine-tuning data.

As a simple HR-related example, RAG might take a user prompt such as "Am I eligible for 401k contribution matching?" and perform a semantic search on the user's company retirement contribution policies in their geography (see Figure 4).

Corporate HR policies are represented in a vector database created from company-specific information and the contextual information identified by the vector search is passed to the LLM along with the original query. Without this supporting information, the LLM may create a response based solely on the typical industry policies where the company is based. The resulting hallucination may seem logical, but it could be incorrect, and the user wouldn't know it.

Figure 4: An advanced RAG-enabled application in action



SOURCE: OMDIA

The crucial role played by vector databases in supporting GenAI in the enterprise

In order to realize a simple RAG pipeline, enterprise practitioners have at their fingertips a raft of rapid prototyping solutions with open source community projects like LlamaIndex, which incorporates several semantic search and RAG-specific integration and orchestration tools. Projects such as LlamaIndex, LangChain, and Haystack can make it easy for practitioners to stand up and use different vector search techniques and vector databases with very little code and practically no expense.

When it comes to putting this kind of solution into production, enterprises ultimately need more than a purpose-built vector database, they need one that's capable of meeting mission-critical demands such as real-time data inserts, the ability to process vector and non-vector data, supporting multiple programming and query languages, scalar filtering, and many other capabilities only found in enterprise-grade databases (see Figure 5).

Figure 5: What goes into an enterprise-grade database with vector search?

Performance	<ul style="list-style-type: none"> ● Highly performant and adaptive similarity search algorithms ● Vertical scalability (e.g., scalable database servers and smart storage) ● Index caching and smart updating without downtime ● Support for real-time (or near-real time) queries ● Horizontal scalability for performant queries regardless of database size (e.g., clustering and sharding)
Security and governance	<ul style="list-style-type: none"> ● Data encryption at rest and in motion (e.g., TLS/SSL) ● Access control and authentication on a row-by-row basis ● Auditing, logging, and data lineage management ● Built-in compliance and reporting for regulatory requirements ● Native, low-level data privacy safeguards
Management	<ul style="list-style-type: none"> ● Full ACID database functionality for all data types and operations including vectors ● Automated resource provisioning and deprovisioning ● Resource monitoring with built-in cost estimation and tracking ● Comprehensive load balancing, backup, and failover capabilities ● Automated patching and upgrade facilities
Development	<ul style="list-style-type: none"> ● Unified query support using SQL as well as the option to drop down to language-specific hooks ● Low/no-code tooling that simplifies application development and supports vectors and other data types ● Simple approaches to application development and vector search queries that minimize code changes as embedding models, indexes, and searches evolve ● A complete and consistent API supported by detailed documentation and cookbook assets ● Hybrid search capabilities combining vector search and traditional keyword/hash lookups and enabling joins between data coming from different sources.

On the proper care and feeding of vector databases

It can be tempting to think of semantic search and its use in RAG workflows as just a different way of analyzing traditional corporate data or a string of numbers—but nothing could be further from the truth. The vector databases behind semantic search can reach incredible levels of complexity with hundreds or thousands of values per vector. As a result, their storage and processing can be incredibly resource intensive.

At a high level, semantic search using vector data types looks a lot like traditional database queries, with both using unique indexing methods to find and quickly return the required information. And while that's true once the data is in vector formats, the original data the vectors are based on must pass through multiple steps before indexes can be generated and searches can commence. First, users must convert data into a form that can be processed by embedding models—for instance, an object store filled with PDF product descriptions must be

converted into text and chunked up into a series of reasonably sized objects – each one of which will end up having its own meaning. Depending on the use case, these chunks may have some degree of overlap, so the semantic context isn't lost.

Practitioners will then send these chunks to an embedding model (e.g. the popular open source text-embedding-3-large model created by OpenAI) which uses a tuned neural network to convert the chunks into vectors with the desired number of dimensions. In general, vector embeddings with higher dimensionality can lead to more nuanced vector searches.

Embeddings together with the raw object data they represent and all metadata about the embedding process itself are then stored in a vector-capable database where the database itself can then use one or more semantic search algorithms (Cosine Similarity, Euclidean Distance, etc.) to locate and retrieve similar information. These vector-capable databases will

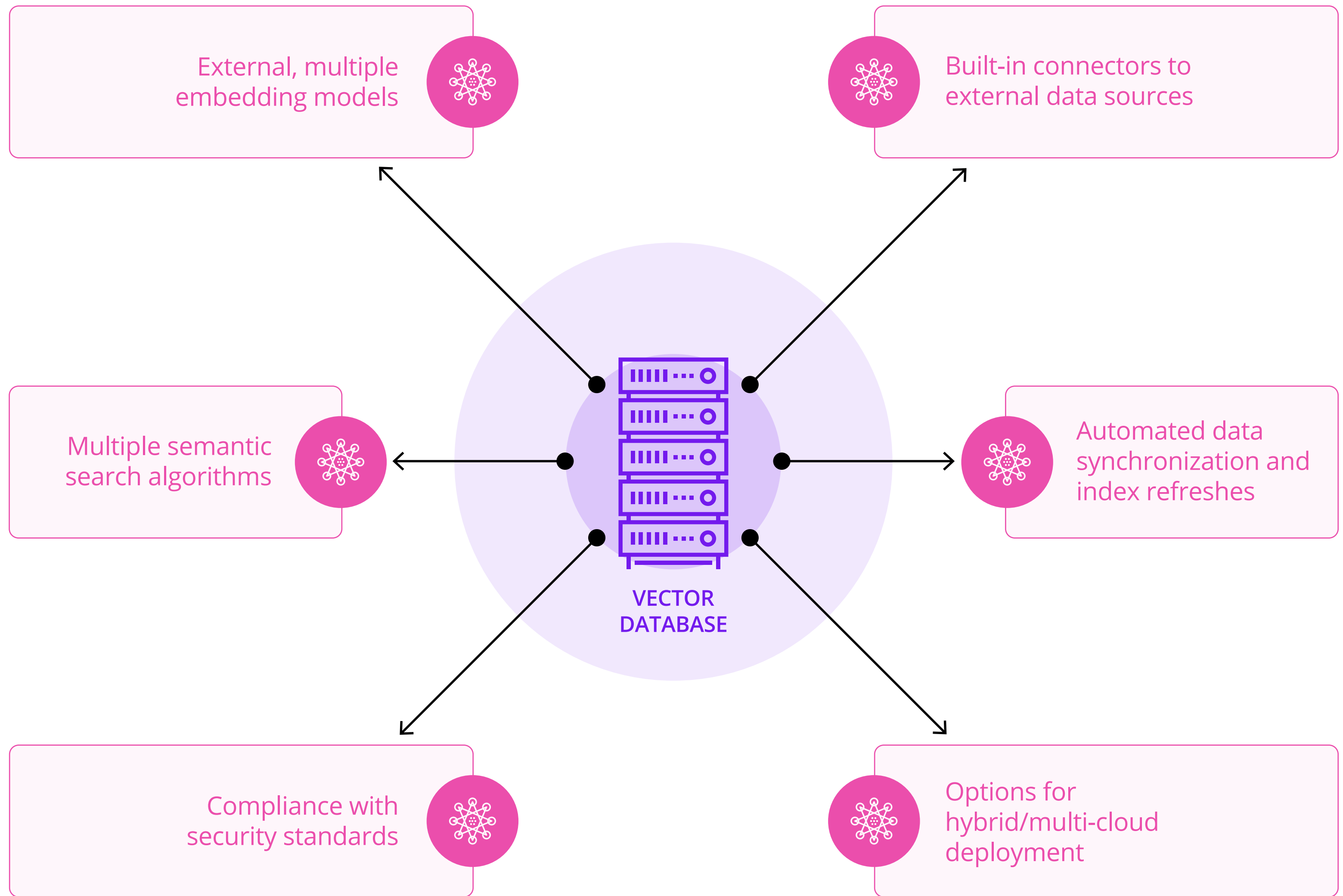


often employ unique indexing techniques algorithms to speed the searches, with single-purpose vector databases often using their own vector-specific querying languages or APIs rather than traditional SQL to initiate searches.

But a vector database isn't just about storing embeddings. It should support different embedding models so one could, for instance, reduce the complexity of their IT environment by using the same software to store embeddings generated from text, images, audio, and video. It should also be able to connect to other databases where source data may be stored, and make it easy to combine the results of vector searches with traditional database queries. And it should provide a mechanism for keeping the source database, vector database, and vectors synchronized with any changes to the underlying data that created them (see Figure 6).

Yet because there are so many moving parts with vector search it can be difficult for companies to provide the near-instant results users demand for anything beyond a very rudimentary proof of concept—especially at scale.

Figure 6: The heavy burden shouldered by vector databases



The trouble with single-purpose vector databases

Vector searches must deal with massive amounts of underlying data whose processing can easily take longer than traditional online transaction processing (OLTP) updates. While it may seem obvious that searching vectors with one thousand or more dimensions will take more time, it's still critical to complete them quickly since the individuals that initiated them have limited patience. Just like relational databases with business data, vector databases use indexes to speed up the searching of massive amounts of data quickly. Most vector databases support indexes, but their ease of use as well as their speed of creation is where some single-purpose vector databases fall short.

The devil hiding in the details concerns the use of a single-purpose, pure-play vector database like Chroma, Weaviate, Pinecone, or FAISS to perform vector searches of data that's housed in a typical OLTP or analytical database such as a data warehouse or data lake. In other words, the problem arises when using two databases to perform multiple operations on the same underlying data.

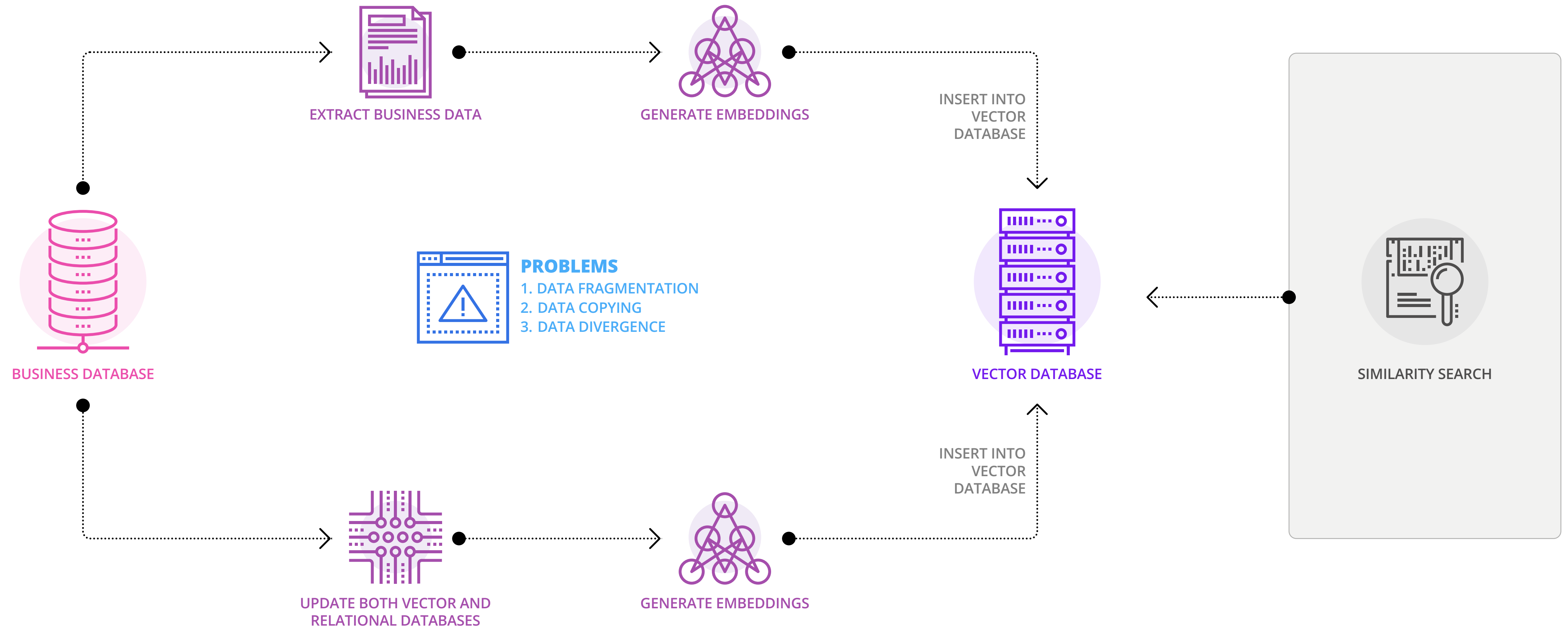
This "one more database" approach creates a number of issues that can increase management and maintenance costs, introduce numerous data- and code-related security risks, place upper limits on performance, elevate solution latency, and lead to costly integration errors.

For example, with very basic applications where vectorized information doesn't change (an embedding of archived help desk transcripts, for instance), the overhead involved in first extracting that data from a data lake, running that data through an embedding model, and then inserting that data into a vector database as a one-time event will not create a tremendous amount of technical debt or impose any latency for solution users.

However, many vector databases take hours to be refreshed from current data, which can lead to a form of double vision with rapidly changing data, where the underlying data behind the vectors being searched are out of sync with live data in the source database (see Figure 7). If you add any form of immediacy and scale, such as performing a vector search on new stock market trade information, then a simple vector search application or RAG pipeline will be unable to keep up.

Unfortunately, such issues are only going to worsen as companies continue to build more mature and interesting vector search implementations, moving beyond basic RAG techniques to explore advanced topics such as search result re-ranking, user query transformations, hierarchical indices, and concepts that haven't crossed our perception threshold yet.

Figure 7: Single-purpose vector databases present numerous data challenges



SOURCE: OMDIA

Architecting the future: unifying vector databases with corporate data

Realizing how hard it is to create, maintain, store, and update vector databases, many enterprise database practitioners are actively investing in converged systems and accelerators for AI (see Figure 8), using resource consolidation and other approaches to reduce technical debt. Increasingly, this entails doing away with unnecessary points of integration by adopting a highly unified, multi model database, such as Oracle Database, that can manage vector embeddings natively alongside enterprise data.

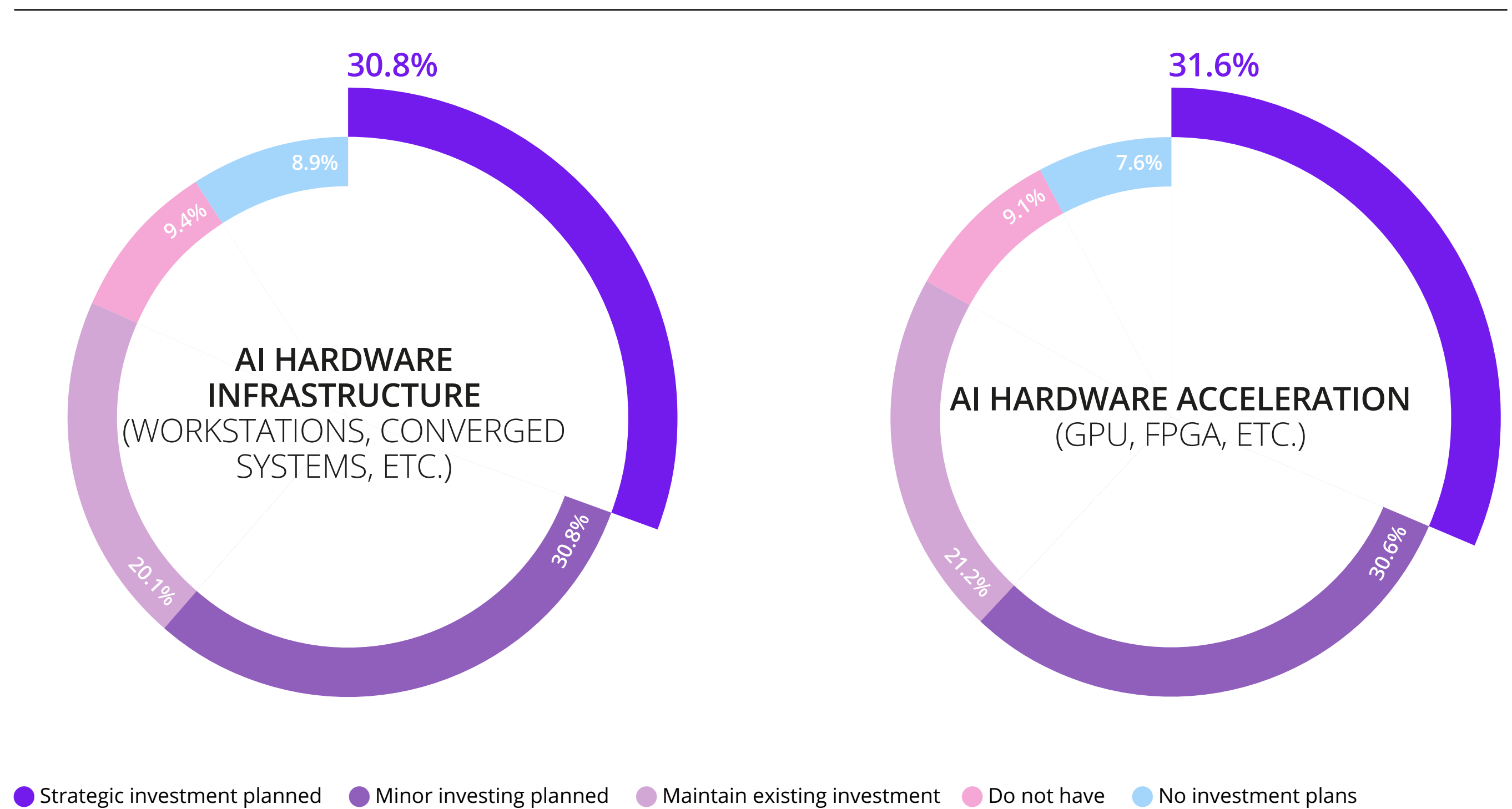
Instead of adding a single-purpose vector database to work in parallel to their existing data stores, cost- and resource-conscious enterprise practitioners are increasingly looking for databases that support standardized access based on open SQL interfaces. These data platforms are typically able to house and process many forms of data; structured, semi-structured, and unstructured.

But not all multi-model databases are created equal. Omdia is tracking a burgeoning trend among both analytical and transactional database players to add vector support natively within those databases. Adding such support has proven to be an effective way for established database players to attract front-end developers interested in building LLM-based GenAI outcomes. It's critical that a multi-model or "converged" database supports several key capabilities.

Figure 8: IT buyers are investing in AI

AI REQUIRES MULTIPLE TYPES OF INFRASTRUCTURE

WHAT ARE YOUR AI TECHNOLOGY INVESTMENT PLANS FOR THE ABOVE DURING THE NEXT 18 MONTHS?



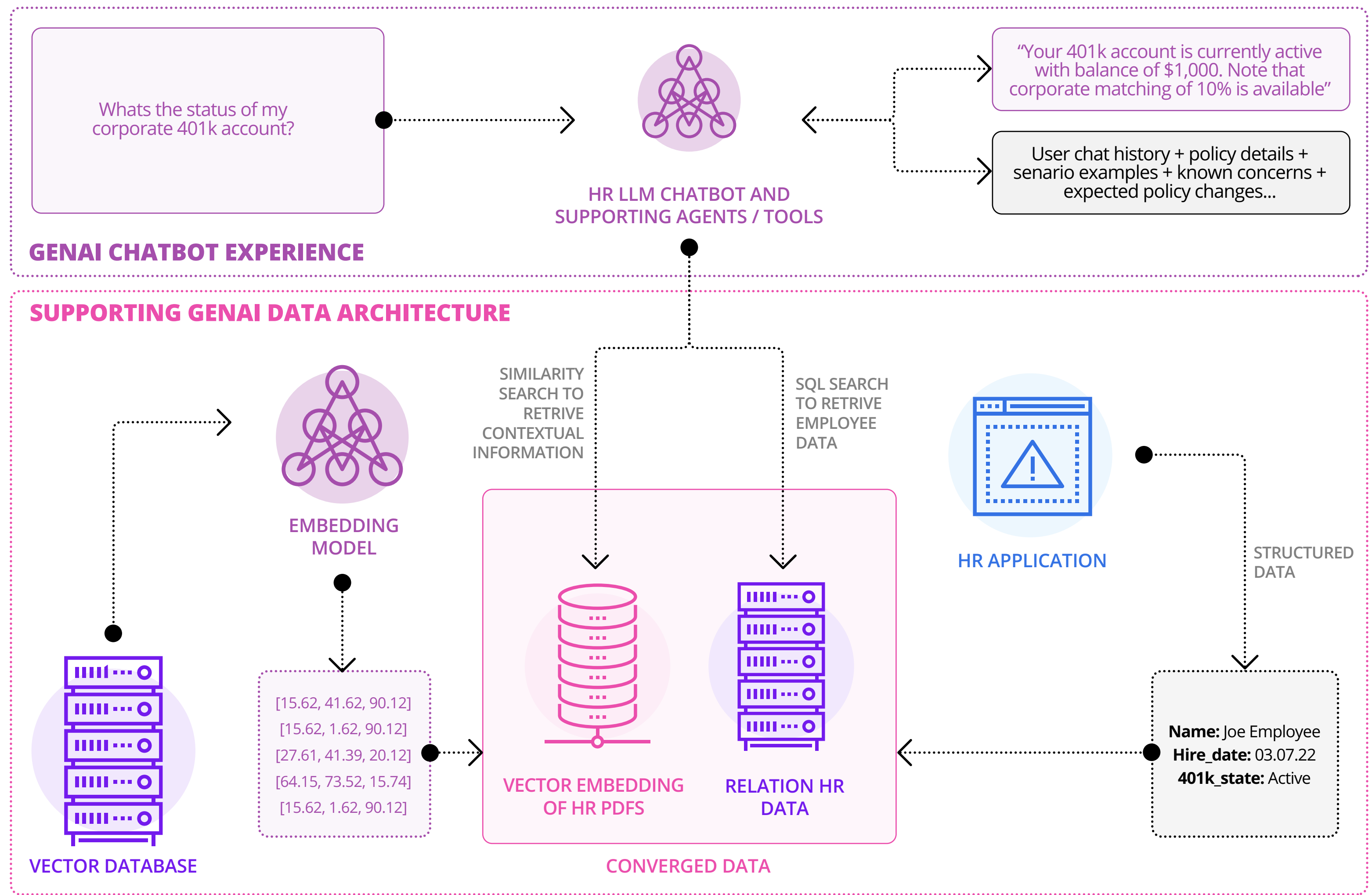
SOURCE: OMDIA, IT ENTERPRISE INSIGHTS: IOT, CLOUD, AI, 5G, AND SUSTAINABILITY - 2024

First, they must deliver critical performance, security, privacy, and governance capabilities to all of their data types—including vector embeddings. Capabilities that are “added-on” to the side of a database may not inherit all of the core capabilities of the underlying database.

Second, they must be able to store and use vector information alongside other forms and representations of enterprise data. Not only should they handle the basics (objects, strings, numbers, etc.), they also should provide a unified means to combine vector search results with other important data types such as graph, spatial, time series, streaming, etc. This ability is key (see Figure 9).

Users of a truly unified database, can create highly performant solutions that combine both structured, unstructured, and vector data without having to move data or incorporate disparate API calls or query languages into applications. Oracle is the first of the established database players to natively integrate vector capabilities into its converged database and not just add pgvector capability onto the side.

Figure 9: Enhancing GenAI question and answer through the inclusion of corporate data



SOURCE: OMDIA

Maximizing the impact of vector search infrastructure

If the GenAI revolution has taught enterprise buyers anything, it is that there's never enough computing power to do everything—with performance and at scale. As a result, success often depends on efficiently using resources that allow companies to get the best value from their data.

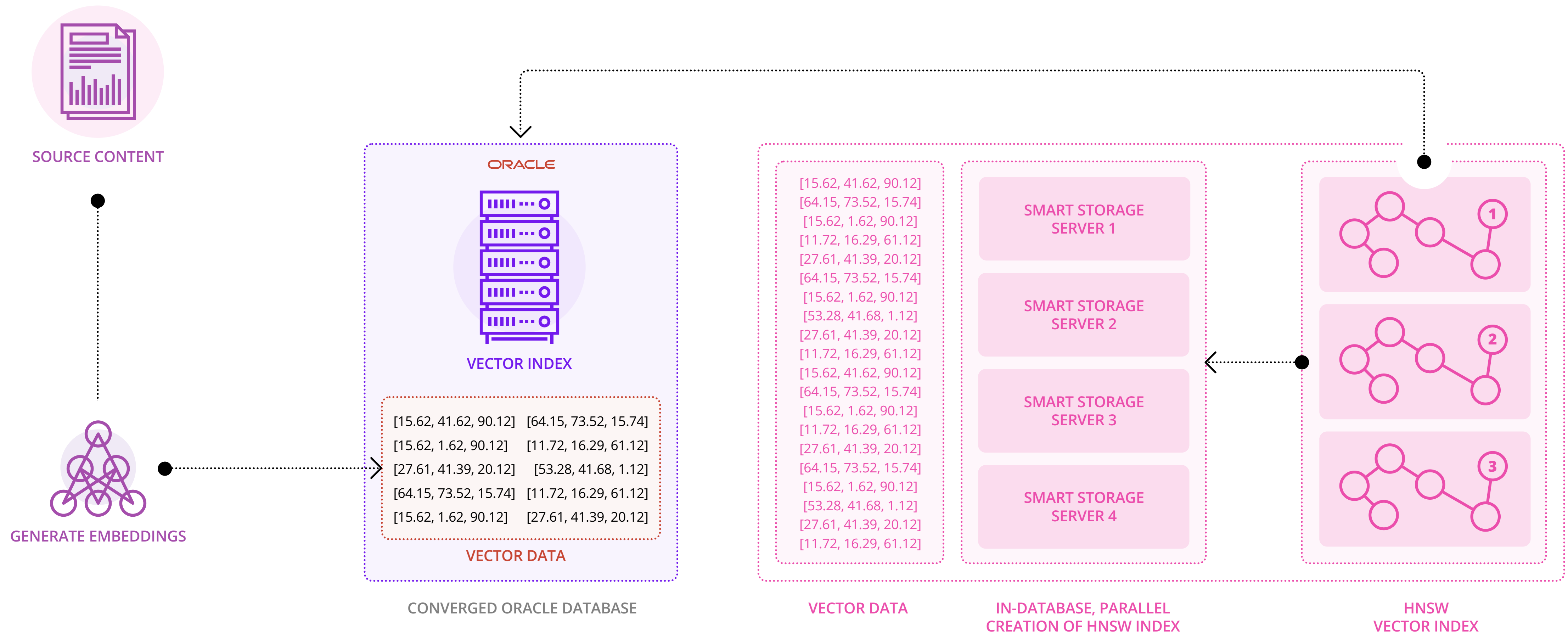
Infrastructure leaders like Oracle and AMD, for example, are working together to enhance the customer's ability to expand the scale and performance of vector search on corporate data by vertically integrating the entire hardware and software stack. This approach is most evident in the Oracle Exadata platform which uses AMD EPYC™ processors with up to 96 cores per processor to increase parallelism in database servers and offload data intensive workloads such as building vector search indexes from the database servers to the storage servers where the data resides. Furthermore, since Oracle Database is a converged database, these same platforms

enable organizations to operate more efficiently by using the same resources to manage vector and corporate data while eliminating the need to move and synchronize data across different databases and infrastructure.

Exadata platforms are not just about scaling data storage or consolidating compute resources. They are fine-tuned to securely handle mixed OLTP, analytics, and AI workloads together as one and to do so at extreme scale. They use high performance, energy efficient AMD EPYC processors in database servers to provide hundreds to thousands of database processing cores, terabytes of high-speed memory, and lightning-fast IO that accelerate all aspects of vector search and traditional database operations. Exadata platforms also take advantage of highly optimized capabilities built into AMD EPYC processors to increase database security by decrypting and encrypting data in real time.

The distinctive scalability features built into Exadata let organizations consolidate their database operations on fewer systems and customize their configurations to meet their aggregate database needs. In addition, Exadata automatically offloads data intensive SQL processing and tasks such as vector index creation to intelligent storage servers. By offloading data intensive vector search algorithms to storage servers, Exadata frees up database server resources so they can process more queries and support more concurrent users (see Figure 10).

Figure 10: Optimizing vector database index creation with Oracle Exadata



SOURCE: OMDIA

Conclusion and advice

Given the relentless pace of innovation within the AI marketplace, companies looking to do more than jump onto the exploding GenAI bandwagon must move quickly and not generate technical debt by cobbling together disparate tools as a minimum viable product. This is particularly true when it comes to building out enterprise-grade vector search functionality with evergreen data.

Creating applications that directly use semantic search capabilities and building RAG pipelines using vector search capabilities is a new, but manageable challenge that can be addressed by using Oracle's converged database that includes vector capabilities. This approach helps prevent data movement, security issues, and management fragmentation that can occur when using multiple single-purpose databases. Companies should therefore follow these best practices:

- Adopt Gen AI architectures and technologies that help maximize accuracy and relevance and are flexible enough to meet changing needs.
- Prioritize unified data platforms capable of natively supporting vector data types and searches as first-class citizens alongside corporate data models (relational, graph, geospatial, streaming, etc.).
- Choose vector-capable databases can deliver OLTP-like performance that meets users' needs for low latency.
- Select a database with native, in-database facilities to run ML workloads, be those embedding generation or advanced RAG semantic search results re-ranking.
- Explore opportunities to directly leverage underlying hardware infrastructure as a means of accelerating vector search processes.

Oracle's full-stack approach to GenAI directly addresses these best practices and can be best viewed in the company's rapidly maturing set of vector search services that are currently available across several Oracle Cloud Infrastructure (OCI) offerings. These include OCI OpenSearch, Oracle Database AI Vector Search, and MySQL HeatWave Vector Store as well as GenAI integrated in Oracle Fusion Cloud Applications, Oracle NetSuite, and industry applications. Whether supporting traditional content mining operations with OCI OpenSearch or enabling app developers to rapidly build complex RAG workflows at scale with AI Vector Search, Oracle and its partners stand apart from most rivals with the ability to help companies go well beyond basic Proofs of Concept to power up vector search in production and in so doing capture the full potential of GenAI in the enterprise.

Appendix

About

Oracle Corporation

Worldwide Headquarters

2300 Oracle Way, Austin, TX 78741 USA

Worldwide Inquiries

T +1.650.506.7000

T +1.800.392.2999

W [oracle.com](https://www.oracle.com)

Connect with us

Call +1.800.392.2999 or visit [oracle.com](https://www.oracle.com)

Outside North America, find your local office at [oracle.com/contact](https://www.oracle.com/contact).

Connect on social

 blogs.oracle.com

 [facebook.com/oracle](https://www.facebook.com/oracle)

 twitter.com/oracle

Integrated Cloud Applications & Platform Services

Omdia

Omdia is a global technology research powerhouse, established following the merger of the research division of Informa Tech (Ovum, Heavy Reading, and Tractica) and the acquired IHS Markit technology research portfolio*.

We combine the expertise of more than 400 analysts across the entire technology spectrum, covering 150 markets. We publish over 3,000 research reports annually, reaching more than 14,000 subscribers, and cover thousands of technology, media, and telecommunications companies.

Our exhaustive intelligence and deep technology expertise enable us to uncover actionable insights that help our customers connect the dots in today's constantly evolving technology environment and empower them to improve their businesses – today and tomorrow.

*The majority of IHS Markit technology research products and solutions were acquired by Informa in August 2019 and are now part of Omdia.



The Omdia team of 400+ analysts and consultants are located across the globe

Americas

Argentina
Brazil
Canada
United States

Asia-Pacific



Australia
China
India
Japan
Malaysia
Singapore
South Korea
Taiwan

Europe, Middle East, Africa

Denmark
France
Germany
Italy
Kenya
Netherlands
South Africa
Spain
Sweden
United Arab Emirates
United Kingdom

Omdia

E insights@omdia.com
E consulting@omdia.com
W omdia.com

 [OmdiaHQ](#)
 [Omdia](#)

Citation Policy

Request external citation and usage of Omdia research and data via citations@omdia.com

COPYRIGHT NOTICE AND DISCLAIMER

Omdia is a registered trademark of Informa PLC and/or its affiliates. All other company and product names may be trademarks of their respective owners. Informa PLC registered in England & Wales with number 8860726, registered office and head office 5 Howick Place, London, SW1P 1WG, UK. Copyright © 2024 Omdia. All rights reserved. The Omdia research, data and information referenced herein (the "Omdia Materials") are the copyrighted property of Informa Tech and its subsidiaries or affiliates (together "Informa Tech") and represent data, research, opinions or viewpoints published by Informa Tech, and are not representations of fact. The Omdia Materials reflect information and opinions from the original publication date and not from the date of this document. The information and opinions expressed in the Omdia Materials are subject to change without notice and Informa Tech does not have any duty or responsibility to update the Omdia Materials or this publication as a result. Omdia Materials are delivered on an "as-is" and "as-available" basis. No representation or warranty, express or implied, is made as to the fairness, accuracy, completeness or correctness of the information, opinions and conclusions contained in Omdia Materials. To the maximum extent permitted by law, Informa Tech and its affiliates, officers, directors, employees and agents, disclaim any liability (including, without limitation, any liability arising from fault or negligence) as to the accuracy or completeness or use of the Omdia Materials. Informa Tech will not, under any circumstance whatsoever, be liable for any trading, investment, commercial or other decisions based on or made in reliance of the Omdia Materials.