VAST

The Database Renaissance

AI's New Center of Gravity and the Changing Economics of Enterprise Data

FOLLOW US:

vastdata.com

AI Is Rewriting the Rules of the Data Layer

AI applications and agents are changing what enterprises need from their data infrastructure. They must work across structured and unstructured data, combine multiple forms of retrieval, incorporate information as it changes, and maintain governance throughout the data lifecycle, all while operating at production scale.

VAST Data is pleased to present the Everest Group report, ***The Database Renaissance: AI's New Center of Gravity and the Changing Economics of Enterprise Data***. The report examines how AI is reshaping enterprise data architectures and why capabilities that have traditionally lived in separate systems – including vector retrieval, streaming, transactional processing, analytics, and multimodal data – are increasingly moving closer together.

Everest Group identifies several important shifts already underway. Vector search is moving closer to the data it represents. Retrieval is becoming increasingly hybrid and multimodal. The traditional separation between transactional and analytical systems is narrowing. And data ecosystems are becoming more real-time as enterprises seek to reduce the hand-offs that introduce latency, complexity, and stale information.

The report also makes an important distinction: convergence is not about putting every workload into a single engine. Instead, Everest Group points toward convergence at the data layer, where different processing resources can operate against a shared, governed data foundation.

The VAST AI OS: A Shared Data Foundation for AI

This vision aligns closely with the architecture of the VAST AI Operating System.

VAST brings together storage, database, native vector search, real-time event streaming, and data processing on a common data foundation. Structured and unstructured data, metadata, and vectors can remain connected on a unified platform rather than being copied across separate infrastructure, while applications can access the different data types and retrieval methods they need.

For AI and agentic applications, that means organizations can:

- Keep vectors close to enterprise data. Native vector capabilities in VAST DataBase allow embeddings, metadata, and structured data to be accessed together without requiring a separate vector database.
- Combine retrieval methods. Vector similarity can be used alongside SQL, metadata filtering, and other enterprise data to provide richer context for AI applications.
- Work with continuously changing data. VAST Event Broker and the broader AI OS support real-time data ingestion and processing alongside persistent enterprise data.
- Maintain governance across the AI data lifecycle. Keeping source data and derived representations on a shared platform reduces the copies and synchronization points across which permissions, metadata, and governance must otherwise be reconciled.
- Scale the data layer for production AI. VAST's disaggregated shared everything (DASE) architecture allows compute and storage resources to scale independently as data volumes, retrieval demand, and AI workloads grow.

Everest Group argues that the database renaissance will ultimately be shaped not by the platform with the longest AI feature list, but by architectures capable of keeping enterprise data accurate, fresh, governed, and economically usable as AI moves into production.

The VAST AI OS was designed around that same challenge: giving AI applications and agents a scalable, unified foundation for accessing and acting on enterprise data.

Everest Group®

October 2026

The Database Renaissance

AI's New Center of Gravity and the Changing Economics of Enterprise Data

VAST This document has been licensed to VAST Data

Contents

03 Introduction

04 Expanding role of databases beyond systems of record

06 Enterprise AI databases reality check

07 Database realignment is already underway

11 The economics of convergence

15 Finding the right convergence point

18 Future outlook

Copyright © 2026, Everest Global, Inc. All rights reserved.

Introduction

Enterprise AI is increasingly moving out of the experimental phases and into core business processes. While models have dominated much of the AI conversation so far, the move from pilots to production is shifting attention further down the technology stack. Enterprises are realizing that model intelligence is one piece of the puzzle, but AI success depends on whether the underlying data can be found, understood, updated, and governed quickly enough and on a sufficient scale for applications and agents to act on it.

This is changing the database's job description. It has to continue supporting traditional applications and predictable transactional workloads, while also serving AI workloads that search repeatedly, combine multiple types of data and information together, and respond to changing realities. A single business request can fan out into several retrievals, incorporate new information as the task progresses, and result in more downstream queries or actions. The databases of today and tomorrow will have to ensure that data is accurate, relevant, current, and that permissions are checked as and when AI needs it.

Providers are responding to these changing demands by bringing in newer capabilities such as vector retrieval, hybrid search, multimodal support, streaming, and AI-ready data processing to their platforms. Providers that once occupied distinct parts of the data ecosystem are now entering each other's territory, creating a broader convergence across the market. The promise is an architecture with fewer copies, hand-offs, and synchronization points. However, convergence can also introduce questions around performance at scale, workload economics, governance, concentration risk, and the continuing value of specialist technologies.

This Viewpoint examines some of the key questions enterprise leaders are facing as database architectures evolve:

- How is AI changing the role and workload of enterprise databases?
- Why are traditional data architectures becoming harder to sustain at scale?
- Where can greater convergence improve economics, freshness, and operability?
- How can enterprises determine the right convergence point for different workloads?

www.everestgrp.com | this document has been licensed to **VAST Data**

Expanding role of databases beyond systems of record

For most of database history, data demands have come from applications built around established business processes. If a customer placed an order, an employee updated the details and an analyst would create a report. The activities could be larger and more complex in certain situations, but the sequence of steps was largely known in advance. The role of databases was clear: maintain the record, provide information at each step, and do not lose anything.

In today's enterprises, and more so for future enterprises, most of these demands come from AI applications, and increasingly agents, bringing in a very different request pattern. Unlike traditional applications that tend to follow predefined workflows, AI applications can draw on a wider range of information dynamically, while agents take this further by pursuing an outcome rather than a fixed list of steps. What an agent needs next will depend on what it learns as the task gets executed. A single request can fan out in multiple directions and loop around until the action is completed. While on one hand, these tasks have to be achieved in almost real time, they are also becoming less sequential and predetermined.

Consider an AI customer service application or agent in a large enterprise given the task to resolve a customer query. The information it needs to address that query can be spread across customer records, product manuals, recent transactions, purchase contracts, and even similar past cases. A conventional AI application may retrieve and synthesize this information to support a response, while an agent may go further by beginning with a recent transaction, discovering a contractual term, consulting the relevant policy document, and then updating the customer record. Some of this information can be neatly stored in rows and columns, but there is also an increasing dependence on information that is unstructured. The eventual answer may not sit in a single system and the path to finding it may also be unknown.

During pilot and Proof of Concept (PoC) stages, a limited number of such interactions might not seem consequential. But in production, as AI applications spread across business processes and agentic use cases begin to scale, these interactions get multiplied across customers, processes, and the number of agents deployed, increasing both the volume and complexity of the data requests. This is leading to a structural shift in database workloads.

- **AI draws on a much larger data landscape:** As in the example above, the pool of potentially useful information can span multiple formats and modalities, including documents, tables, images, or events. AI also creates derived representations such as metadata, embeddings, and vectors that should remain connected to the source data. Databases are thus required to expand beyond the structured records to support a broader range of data types and representations

- **Retrieval is becoming a core workload:** Traditional databases have been successful in querying data when the requests are schema-based. However, AI applications increasingly need to determine what information is relevant, while agents may need to do this repeatedly as the task evolves. This can require hybrid retrieval across structured filters, metadata, keyword search, and vector similarity while preserving freshness and permissions
- **Interaction with databases is becoming more continuous:** A traditional application would run a query and return the result. But an AI application or agent may run retrievals multiple times, incorporate newer data and existing context, capture outcomes, and then repeat the process again depending on the task. Agent loops can also branch into several lines of inquiry or retry failed steps, multiplying what began as a single user request into many database interactions. At the same time, the underlying data itself is continuously changing, which means derived representations such as embeddings and vector indexes may need to be refreshed or regenerated as source information evolves. AI applications are therefore increasing both the frequency and continuity of database access
- **The acceptance bar is continuously rising as AI applications and agents mature:** At production scale, databases might be expected to manage multiple interactions across modalities in real time, while avoiding delays that can compound across successive steps. They also need to preserve consistency in governance as data moves from raw sources through transformation, embedding, indexing, and retrieval, so that permissions, lineage, and policy controls remain aligned throughout the lifecycle. Depending on the requirement, scale and speed will have to be balanced with quality and compliance across the lifecycle

Together, these shifts are pulling capabilities such as unstructured data support, vector retrieval, continuous ingestion, and policy-aware access closer to the database. They are also increasing the importance of keeping source data, derived representations, and governance controls synchronized as information changes. The shift is not only in what databases need to manage, but also in who is consuming them, at what frequency, and with what level of autonomy.

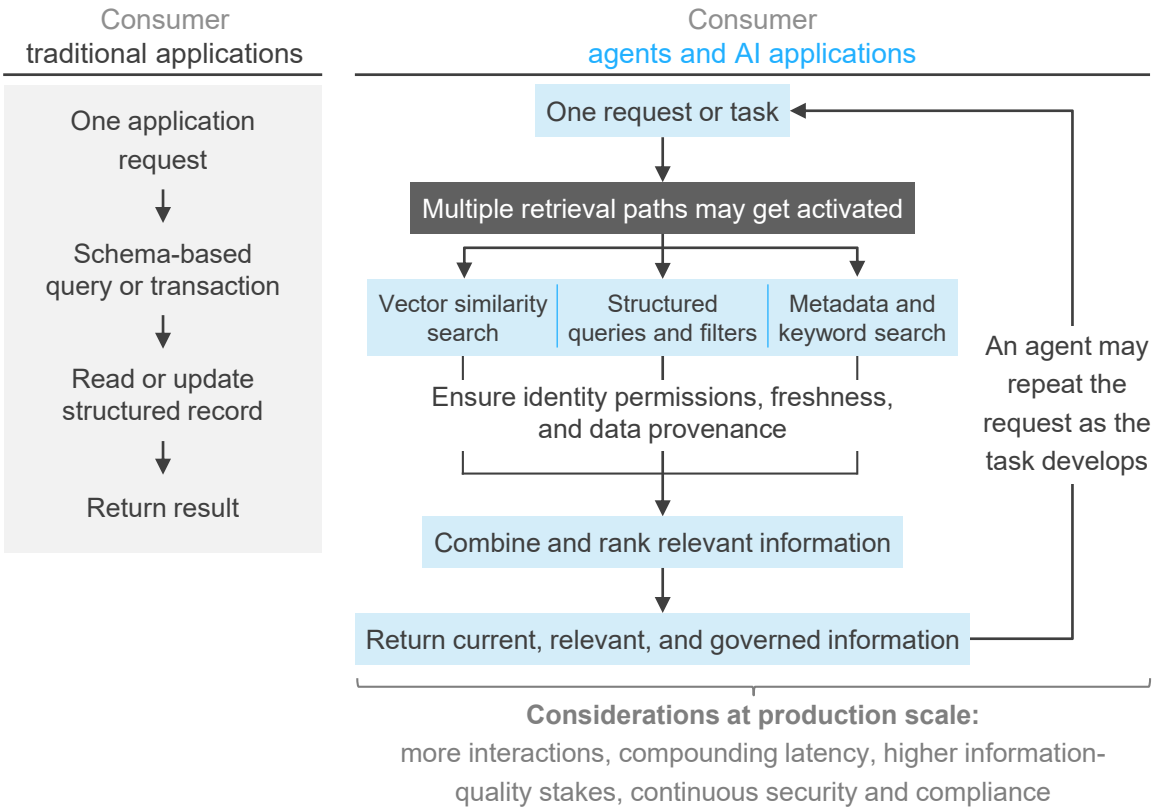
While AI is exponentially expanding the scope of database workloads, most enterprise data environments have not evolved at the same pace. This gap might seem manageable when testing pilots, but it can become a major bottleneck as AI moves into production.

Exhibit 1 illustrates how a single AI request can span a wider set of database workloads compared to a traditional application request.

www.everestgrp.com | this document has been licensed to **VAST Data**

Exhibit 1: Expanding role of databases in the AI era

Source: Everest Group (2026)



Enterprise AI databases reality check

AI pilots are built on their best behavior. The data is curated, the scope is defined, and the monitoring team can step in when something goes wrong. Production is not this forgiving. AI agents have to work with live data, changing permissions, more business users, and processes that cannot pause while the underlying data issue is being fixed. For most enterprises, the difficulty is not that the required information has never been captured. The challenge lies in information being scattered and ungoverned across the enterprise.

While there have been multiple attempts to consolidate data, enterprises still struggle with persistent data silos. Over time, organizations have adopted different databases and platforms for different needs such as transaction processing, analytics, storing documents, search, streaming, and recently, for managing vectors. Each system works well for its intended purpose, but AI often needs information from several of them at the same time.

www.everestgrp.com | this document has been licensed to **VAST Data**

According to an Everest Group survey of 123 enterprises, 77% of the enterprises stated that their current data foundations cannot reliably support AI use cases.[1] A customer-service agent (like in the example mentioned in the previous section) may need transactional data from an operational database, contracts from a document store, recent events from a streaming platform, and semantically similar cases from a vector index. In a fragmented landscape, the agent can run into challenges such as missing context, conflicting answers, and latency delays. Nearly 67% of the surveyed enterprises stated latency as a limitation to their AI scaling efforts.

These challenges become even more difficult when additional representations of the same underlying data are created. A single document can exist as the original file, extracted text, chunks, embeddings, vector-index entities, and metadata. These representations are logically connected but are often stored and refreshed separately. If the source changes, its downstream representations may not change at the same time. If a user's permission is revoked, the retrieval layer may not immediately reflect it. This can result in agents picking up outdated and restricted information.

Enterprises are compensating for these gaps with an increasingly complex layer of pipelines and integrations. A lot of the connective tissue for the data layer is being built across separate products, often with custom logic to manage hand-offs between them. These measures might work well for more standalone use cases but can become increasingly brittle as data volumes increase, sources change, and more teams build AI use cases. Agent failures can become harder to trace and changes can become more difficult to implement with ownership scattered across teams.

Database realignment is already underway

The database challenges have not emerged because enterprises chose their stacks poorly. Most traditional systems were implemented to solve a specific problem, and they were often solving it well in the pre-generative AI era. Challenges being faced today are largely because AI requires data to move continuously across systems. Hence, the specialized systems of yesterday now seem like an increasing coordination burden.

The market has already started responding to these changing requirements. Not by creating an entirely new database category, but by stretching existing categories into adjacent territory. Capabilities that previously sat in separate layers of the data stack, such as vector retrieval, full-text search, streaming, transactional processing, and multimodal storage, are increasingly appearing within the same provider portfolios. Instead of being isolated feature additions, these movements help providers to support a larger share of the data lifecycle while reducing the synchronization required between different systems.

1 Everest Group's Building a Trusted Data Foundation for Scalable AI report

As a result, the previously established boundaries between operational databases, analytical platforms, vector stores, and streaming systems are becoming more porous.

Exhibit 2 illustrates the expanding capability stack of modern databases, distinguishing foundational table stakes from emerging capabilities required for increasingly complex AI workloads.

Exhibit 2: The expanding database capability stack for enterprise AI

Source: Everest Group (2026)

	Table stakes capabilities	Emerging capabilities
Data storage and management 	Durable data persistence Consistent read/write operations Support for at least one data model Schema and metadata management	Native vector data types and indexes Structured and unstructured data managed through a common foundation
Retrieval and relevance 	Structured and model-specific querying Exact-match and range retrieval	Hybrid retrieval across vector, keyword, and structured queries Multimodal retrieval Re-ranking and routing across retrieval methods
Workload support 	Operational or analytical workload support Change-data capture Batch ingestion and integration	Transaction ingestion alongside analytical and AI consumption Native streaming and event-driven processing AI-assisted enrichment and embedding generation
Enterprise governance and operations 	Authentication and access controls Audit logs and monitoring Backup and point-in-time recovery High availability and capacity scaling	Consistent permissions across different data representations End-to-end lineage Automated resource management

These changes in the capability map are not a checklist that every database provider will eventually offer. Providers have different starting positions and are unlikely to converge at a single identical architecture. However, the direction of movement is the same, wherein capabilities that previously required separate systems are now moving closer to the core data platform. We are witnessing four such trends in the market.

Vector search is getting integrated with where data originally lives

Early generative AI architectures extracted unstructured data, converted it into embeddings, and stored these embeddings in a dedicated vector store. This offered a quick fix for enterprises to test PoCs and pilots, but when the agents move into production, the dedicated vector stores become yet another version of enterprise data that needs to be refreshed, maintained, governed, and secured. Increasingly, database providers are allowing vectors to sit alongside the business data they represent rather than as an external layer. Across the major database providers, vector capabilities such as storage, indexing, and search are being embedded directly within the same system. This in turn reduces the extra synchronization effort needed to maintain separate stores and keep them current.

Over time, we expect vector support to become a table stakes capability rather than a differentiator by itself. The edge for providers will come from how well they are able to support demanding use cases which require massive scalability, superior retrieval quality, strong operational resilience, and high-performance metrics.

Hybrid retrieval is becoming the new baseline for search

Vector-native retrieval architectures leverage semantic similarity where the focus is on finding information closest in meaning. But in critical agentic applications, relevance can be more complicated. For example, a factory maintenance agent investigating an equipment failure will need error-code mappings, service records, sensor events, and technician manuals. The agent will not only have to retrieve these differently structured information pieces but also re-rank the combined results based on the situational circumstances.

Retrieval is thus becoming more hybrid and multimodal. Platforms are combining vector similarity with keyword searches, structured queries, metadata filtering, and re-ranking. They are also enabling retrieval across multiple types of data such as documents, images, audio, and events. Combining these capabilities can help produce a result that is not only similar to the query but also better aligned to the situation.

Going forward, as providers add more retrieval methods, they will also have to enable the process of choosing between these methods based on efficiency. Instead of developers manually choosing the right retrieval mix for each use case, retrieval systems could add capabilities for dynamic selection based on weights and available data. Hence, the focus will not be on the providers who can return the nearest result fast but who can combine multiple signals to produce holistic and relevant results.

www.everestgrp.com | this document has been licensed to **VAST Data**

The gap between transactional and analytical systems is narrowing
Enterprises have traditionally separated transactional and analytical data processing. Operational databases handled fast-moving transactions, while data was copied into warehouses or lakes for analysis. This architecture becomes a bottleneck when an AI application needs both historical analysis as well as data from an event that happened a few seconds ago. Analytical platforms are now adding support for low-latency queries, while operational databases are expanding into analytics and retrieval. Emerging architectures are bringing together capabilities to support both transactional and analytical data in the same platform. While there are multiple approaches for this, the common objective is to shorten the journey between data being created and becoming available for downstream AI applications.

This trend does not necessarily point toward one engine capable of running both types of workloads together. Rather, the more likely scenario would be a convergence at the data layer with different processing resources that can operate on a shared and governed foundation. This will also shift the differentiation toward how quickly and consistently each workload can use the same underlying data.

Data ecosystems are expected to respond in real time
Traditional databases were designed primarily to store records and respond when an application queried or updated them. More extensive activities such as streaming updates, enriching data, or refreshing indexes typically relied on more specialized tools. However, every hand-off in this lifecycle added latency, more infrastructure to manage, and yet another place where the process could potentially break.

Recently, capabilities such as streaming ingestion, change-data capture, event triggers, and pipeline automation have been brought closer to the databases. This helps in keeping data and indexes fresher while reducing some of the plumbing required to move data between separate processing layers.

The databases are unlikely to absorb every form of pipeline execution and heavy processing might still reside in a separate layer. However, time-sensitive data preparation, enrichment, and event-driven actions can benefit from moving closer to where data is captured and governed. The differentiation will depend on how processing can occur efficiently with the least delay, movement, and loss of control.

Convergence and scale will be part of the same equation

All the above trends are trying to bring more of the data lifecycle closer together. While convergence might look very convincing on an architectural blueprint, it is not sufficient for enterprises to simply question whether the platform provides all the required capabilities. The real test is whether these capabilities continue to work well together as the workload grows. A platform that performs well with a few million vectors can still break when scale reaches the magnitude of tens of billions of vectors with multiple agents trying to work concurrently.

www.everestgrp.com | this document has been licensed to **VAST Data**

Enterprises will have to ask questions that test platform performance at larger scales. Can a large indexing job continue without slowing down another retrieval? Can query volumes increase without affecting other applications? Will concurrency impact the performance negatively? It is critical for enterprises to consider platform capabilities against expected data workloads in the future instead of siloed demos of each capability.

If data ecosystems start breaking as workloads increase or become complex, convergence has not delivered the value it was supposed to and the earlier challenges of fragmentation have just moved inside the platform. Therefore, for convergence to be a viable enterprise strategy, its economic value has to hold up at scale.

The economics of convergence

As database capabilities converge, enterprises should choose providers not on feature checklists but on whether bringing these capabilities together reduces the cost of keeping data accurate, fresh, and governed at scale. To determine this, enterprises will have to assess AI economics from the angle of the data layer.

While there have been many debates and research about AI economics focusing on model training, inference, and GPU consumption, enterprises often overlook the cost that AI creates on the data layer. This includes ingestion, movement, preparation, and operating effort required to serve data to the AI workloads.

Where data layer costs accumulate

Every AI workload requires an essential set of data operations including ingestion, transformation, embedding, indexing, and retrieval. All these activities consume resources, leading to an essential data operation cost. While different providers might have different approaches to accomplish these tasks, leading to variations in this cost across platforms, the underlying work has to be performed regardless of the architecture.

However, when this work gets distributed across multiple systems and tools, a new cost item appears because now enterprises will have to coordinate across the system boundaries to ensure that the information provided to AI is relevant and complete. This cost accumulates in the form of a data coordination burden. Essential costs and coordination burden can sometimes seem interrelated. For example, creating an embedding and building its index is critical for vector retrieval. However, keeping copies of those embeddings across multiple disconnected systems, moving them frequently, or updating data through multiple pipelines can add to the coordination burden.

Exhibit 3 illustrates the components that make up the total data layer cost.

www.everestgrp.com | this document has been licensed to **VAST Data**

Exhibit 3: Components of total data-layer cost

Source: Everest Group (2026)



Essential data operations

Core activities needed for AI such as ingestion, preparation, storage, governance, embedding, and retrieval

Data coordination burden

Costs incurred due to cross-system data movement, duplicated capacity, synchronization, integration and operating effort, and repeated governance and control

Total data layer cost

Convergence tilts the equation

Convergence can meaningfully improve the economic value of AI use cases across four dimensions: more efficient infrastructure utilization, higher data freshness, increased productivity, and simplified governance and control.

More efficient infrastructure utilization

One of the most immediate and direct benefits of convergence is reducing the infrastructure needed to support the workload. If the same dataset no longer needs to be maintained across multiple platforms, enterprises will be able to reduce duplicate storage, memory, indexing, and data-transfer requirements. This becomes increasingly more important at higher scales. Additionally, since memory, flash capacity, and power are becoming more constrained, the economic benefit of convergence is also increasing.

Higher data freshness

In a fragmented architecture, a change may need to travel through multiple hoops before it becomes visible to an AI application. Every additional hand-off adds another opportunity for lag. Bringing some of these functions onto a common data or processing layer can shorten that path. Convergence can increase the enterprise value by reducing the time between a change in reality and the point at which AI can act on that change.

Increased productivity

A best-of-breed stack can offer excellent component-level performance but requires teams to continuously integrate and monitor the entire process flow. Fewer integration

boundaries can reduce the engineering work needed to move from a successful PoC to production. This benefit can also be reflected through lower troubleshooting effort, faster release cycles, and fewer specialist skills required to operate the environment. While all of these benefits cannot be directly seen on cloud bills, they can bring down the overall AI TCO.

Simplified governance and control

When source data and its derived representations are spread across several systems, enterprises have to ensure that all the relevant permissions, policies, lineage, retention, and audit controls are consistently applied and reconciled for each environment. Convergence reduces the number of copies and control points that enterprises need to govern. Therefore, it can make it easier to implement permissions throughout the lifecycle, maintain a consistent audit trail, and avoid gaps caused by outdated policies. While it does not remove governance measures altogether, it can significantly reduce the effort required to meet them consistently.

Convergence has a price tag

Enterprises should not consider convergence-related savings to be automatic. The final economic value can depend on multiple factors.

Cost of transition

Migrating a workload can involve multiple effort-intensive steps, including moving source data and metadata, rebuilding embeddings and indexes, changing applications, and retesting governance controls. This effort can increase significantly if the workload is large, complex, deeply embedded in business processes, or subject to strict compliance requirements. Enterprises also have to ensure reskilling of existing employees and change management for higher adoption. All of these put together can create significant transition costs before any savings from convergence can be realized.

Platform economics and concentration risk

Bringing more workloads to one platform can also alter the relationship dynamics with the provider. Integrated capabilities can carry a premium, although some of them might get offset by lower coordination requirements. As commercial dependence on one provider increases, negotiation flexibility might also be reduced. If the platform relies on proprietary capabilities, it can also increase switching costs and lock-in concerns. While convergence will reduce the number of individual failure points, a provider outage can impact a larger number of workloads at once.

www.everestgrp.com | this document has been licensed to **VAST Data**

Cost of giving up specialist advantage

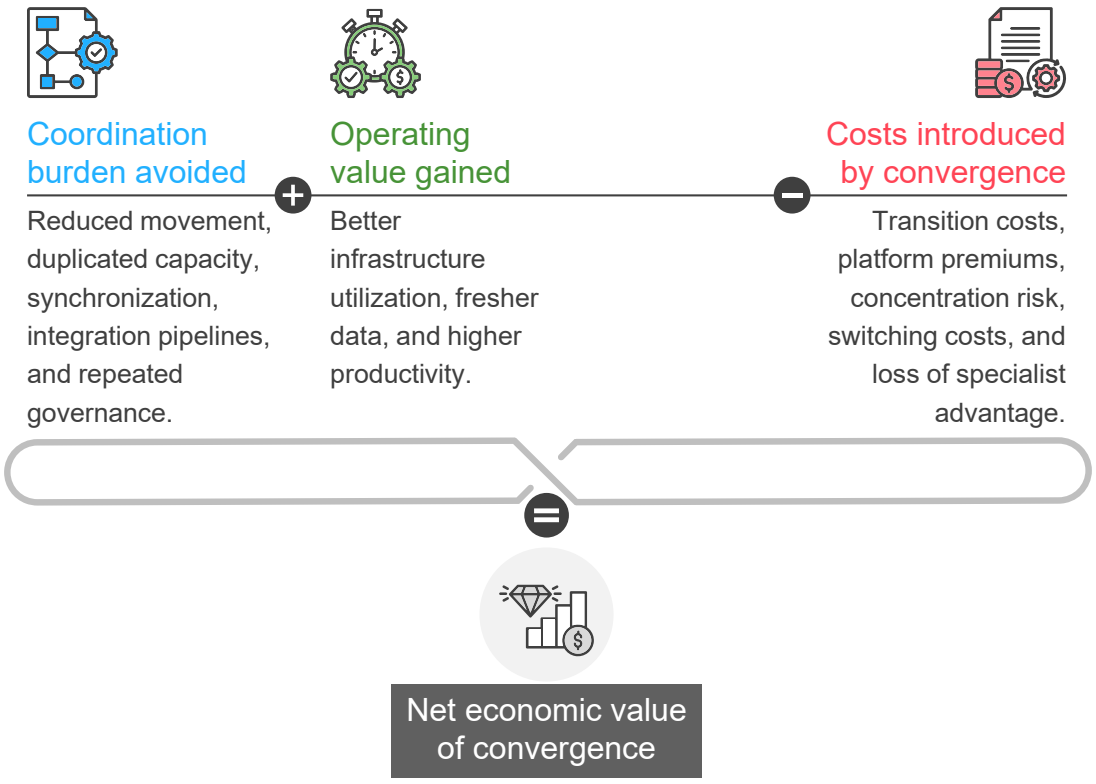
Specialist tools and platforms can usually provide better functionality for the relevant workloads, whereas a converged platform might need additional compute to achieve the same quality and latency. Thus, enterprises will have to consider convergence gains against potential performance trade-offs.

These factors discuss some of the hurdles enterprises should consider when evaluating the economic value of convergence because platform pricing alone does not provide the complete picture.

Exhibit 4 enumerates the economic trade-offs that enterprises should consider while moving toward convergence.

Exhibit 4: Balancing key trade-offs to determine the economic value of convergence

Source: Everest Group (2026)



This equation points enterprises toward a spectrum of convergence, rather than the blanket conclusion that everything should converge.

Finding the right convergence point

After making the case for convergence in the previous section, the obvious next question could be whether to converge or not. However, there is no straightforward answer. The degree of convergence is a spectrum, and for different enterprises, the right point on this spectrum can look very different. To add to the challenge, the same enterprise might occupy multiple points on the spectrum.

Let us consider a typical enterprise and the nature of workloads they can be dealing with. A knowledge repository might use only semantic retrieval on a relatively stable list of policy documents when a query comes in. An AI-enabled fraud detection system will have to continuously process live signals and compare them with recent transaction data to flag an activity. A customer service agent will have to combine data from Customer Relationship Management (CRM), Master Data Management (MDM), past queries, and transactions to respond efficiently. All these workloads place very different demands on the underlying data layer, and an enterprise would want all of these processes to run in parallel.

Therefore, the enterprise objective to find that one golden convergence point for its entire data estate supporting a variety of AI workloads can be very challenging. Some workloads will benefit from a converged ecosystem, and some will continue to benefit from specialist systems. A cookie-cutter approach can overlook such nuances. Enterprises will have to evaluate critical workloads individually and decide where each of them should belong on the spectrum instead of trying to bring all workloads toward one common point.

Key considerations when deciding the right convergence point

Finding the right point of convergence can be a convoluted exercise due to the presence of multiple contributing factors and how the same factor can pull the enterprise in opposite directions. For example, common governance can simplify control but can complicate sovereignty or isolation requirements. Large vector scale will make data movement across systems more expensive, but such enterprises might also need the higher performance of a specialist platform.

Exhibit 5 presents the IMPACT framework, which brings together some of these considerations that enterprises can assess when deciding the convergence point for their workloads. It includes factors related to the workload such as interdependencies, scale, pace of change, and need for specialist capabilities, along with the broader organizational factors such as governance requirements and operating readiness. The primary purpose of this framework is to show which factors support closer integration, which justify retaining boundaries, and where the trade-offs need to be evaluated more deeply.

www.everestgrp.com | this document has been licensed to **VAST Data**

Exhibit 5: The IMPACT framework for determining the right convergence point

Source: Everest Group (2026)

Decision parameters	Key questions to ask	Impact on the convergence decision
Interdependency 	How many systems and datasets does an AI task need to touch before the final response? Can these systems be accessed independently, or does the result from one determine the next retrieval? Will data and context from different sources have to be synchronized for the final result?	Multiple dependent hand-offs across systems can make a distributed ecosystem harder to sustain in the long run because latency and synchronization gaps can compound across the task.
Magnitude of demand 	How much data and how large a vector index has to be actively searchable in production? As workloads move from PoC to production, how many users, applications, and agents will be querying simultaneously? How many retrievals is each request expected to trigger? Can indexing, retrieval, and transaction workloads scale independently without competing for resources or compromising on quality?	PoCs often mask limitations related to scale. In production, larger data, concurrency, and repeated retrievals can degrade latency and quality. Convergence can help by reducing copies and hand-offs, but it has to allow for different workloads to scale independently.
Pace of change 	How frequently does the underlying information change? Do different data sources change at different rates? How quickly should a change become available to the agent? Does each change require data to be changed or reindexed across several systems?	Frequently changing data can strengthen the case for convergence, especially when synced updates have to reach the agent quickly. However, this can be more complicated when sources are changing at different rates or continuous updates disrupt retrieval efficiency.
Advantage of specialization 	Does the workload depend on a certain performance or capability parameter that a broad platform cannot match? Can the specialist system be connected to the data foundation without a significant coordination burden?	A critical performance or capability advantage justifies retaining a specialist. However, when this advantage is marginal compared to the coordination burden, convergence can be the more efficient option.
Control and governance 	Will the permissions and controls be required to carry over from source to agent action? Do sovereignty or regulatory requirements need certain data or systems to remain separate? Can the enterprise maintain one audit trail for the agent across the systems?	Convergence can help with keeping policies aligned across the interaction and maintaining a consistent audit trail. However, the convergence should be compliant with the regulations.
Technology and operating readiness 	Does the workload depend on a system which is deeply entrenched in the business processes? Can converged capabilities be introduced gradually while existing systems run in parallel? Do ownership and skills support a shared foundation?	Readiness will decide the degree of convergence that is practical today. Convergence can be easier to implement for newer workloads, but if the systems are embedded and ownership is siloed, gradual integration might work better.

www.everestgrp.com | this document has been licensed to **VAST Data**

In real enterprises, it is unlikely that all IMPACT factors will point in the same direction. Let us consider the example of the customer service agent again, which is dependent on information from policy documents, CRM records, product information, previous interactions, and recent transactions. Most routine requests might need the CRM record and relevant policy document, whereas transaction records might be needed for issues related to disputed payments. All these sources move at very different speeds. Policies can change periodically, permissions will have to take effect immediately, and transactions continue to happen in real time.

If the biggest delay comes from moving updated documents into a separate vector index and keeping its metadata and permissions synchronized, bringing document storage, vector indexing, and access controls might make sense. However, this will not justify migrating the CRM or the transactions platform, just because they are part of the same workflow and are performing well otherwise. In fact, migrating the CRM and the transaction platform can add to the cost and risk factors without solving the delay issues. If, after some time, an enterprise realizes that they need a vector platform with specialist performance that a broad platform cannot match, the decision might change again. It is essential for enterprises to examine actual request paths before defining which boundaries to move.

The IMPACT framework is the first step in the decision process. It helps enterprises evaluate workloads and their convergence boundaries. However, this proposed scope also needs to pass the economic test from the previous section. Enterprises will have to test whether the potential benefits like lower coordination effort, improved freshness, better utilization, and reduced operational complexity are large enough to justify the costs of change. This implies comparing the expected coordination savings and operating gains against the cost of transition, switching costs, and any loss of specialist advantage. If the economics do not hold, the enterprise might choose to retain the current architecture. If they do hold, the workload becomes an ideal candidate for movement along the spectrum.

Moving across the convergence spectrum

Once the workloads that can benefit from closer integration have been identified, enterprises will also have to chart the course to achieving it while ensuring minimal disruption to the business.

Focus on the boundary creating the biggest problem instead of trying to consolidate the entire architecture. The best starting point is the boundary where an AI workload repeatedly loses time, money, and resources. Pinpointing that source of friction can provide a clearer scope and a more credible RoI for architectural change than an open-ended effort to consolidate everything.

Use production conditions as the baseline rather than PoC stages. Testing with representative data volumes, update patterns, and concurrent demand shows whether the proposed architecture can continue to sustain at scale.

www.everestgrp.com | this document has been licensed to **VAST Data**

Plan for operational change while you are planning for the architectural shift. Convergence can change ownership across teams and might also create newer skills and support requirements. Addressing and planning for these changes early can reduce disruption and help with better effectiveness.

Treat convergence as a continuing architecture decision rather than a one-time exercise. Workload requirements, enterprise priorities, regulations, and data technologies are ever evolving. Regularly revisiting the spectrum can help determine if the contributing factors still hold or if the workload moves on the spectrum.

Future outlook

Tech boundaries between different providers are blurring as more and more providers invest in AI database capabilities. There is no single architecture that can be marked as the golden destination for all enterprises, and it is unlikely that the rapid pace of innovation will slow down anytime soon. Therefore, enterprises will have to make their database and convergence choices as the market continues to evolve.

Databases will be expected to continue supporting traditional applications alongside the AI applications and agents of tomorrow. As these capabilities become more widely available, today's differentiating feature can quickly become tomorrow's table stakes. Enterprises will have to base their decisions less on feature lists and more on actual production evidence. Does retrieval quality hold as the vector scale reaches petabytes? Does new information become available quickly enough? Can traditional and AI workloads run together without disrupting each other? And will the economics be viable as agents become the permanent consumer of the data ecosystem?

Data architectures also have their own version of the butterfly effect. A choice that looks small during a PoC (where a vector index is captured, how permissions are synced, or where agent state is stored) can create much larger consequences when the workload reaches the production stage. As technologies continue to develop, enterprises will need flexibility to revisit these convergence choices rather than becoming locked into assumptions at an earlier stage.

The database renaissance will not be won by the platform with the longest AI capability checklist. It will be shaped by architectures that keep enterprise data accurate, fresh, governed, and economically usable. AI's center of gravity is moving toward the data layer because every useful AI action ultimately depends on it.

www.everestgrp.com | this document has been licensed to **VAST Data**

Everest Group®
With you on the journey

Everest Group is a leading research firm helping business leaders make confident decisions. We guide clients through today's market challenges and strengthen their strategies by applying contextualized problem-solving to their unique situations. This drives maximized operational and financial performance and transformative experiences. Our deep expertise and tenacious research focused on technology, business processes, and engineering through the lenses of talent, sustainability, and sourcing deliver precise and action-oriented guidance. Find further details and in-depth content at www.everestgrp.com.

This study was funded, in part, by VAST Data

For more information about Everest Group, please contact:

+1-214-451-3000

info@everestgrp.com

For more information about this topic please contact the author(s):

Amardeep Modi, Vice President
amardeep.modi@everestgrp.com

Mansi Gupta, Practice Director
mansi.gupta@everestgrp.com

Snigdha Uppala, Senior Analyst
snigdha.uppala@everestgrp.com

Ishi Thakur, Senior Analyst

Notice and Disclaimers

Important information. Please read this notice carefully and in its entirety. By accessing Everest Group materials, products or services, you agree to Everest Group's Terms of Use.

Everest Group's Terms of Use, available at www.everestgrp.com/terms-of-use, is hereby incorporated by reference as if fully reproduced herein. Parts of the Terms of Use are shown below for convenience only. Please refer to the link above for the full and official version of the Terms of Use.

Everest Group is not registered as an investment adviser or research analyst with the U.S. Securities and Exchange Commission, the Financial Industry Regulatory Authority (FINRA), or any state or foreign (non-U.S.) securities regulatory authority. For the avoidance of doubt, Everest Group is not providing any advice concerning securities as defined by the law or any regulatory entity or an analysis of equity securities as defined by the law or any regulatory entity. All properties, assets, materials, products and/or services (including in relation to gen AI) of Everest Group are provided or made available for access on the basis such is for informational purposes only and provided "AS IS" without any warranty of any kind, whether express, implied, or otherwise, including warranties of completeness, accuracy, reliability, noninfringement, adequacy, merchantability or fitness for a particular purpose. All implied warranties are disclaimed to the extent permitted by law. You understand and expressly agree that you assume the entire risk as to your use and any reliance upon such.

Everest Group is not a legal, tax, financial, or investment adviser, and nothing provided by Everest Group is legal, tax, financial, or investment advice. Nothing Everest Group provides is an offer to sell or a solicitation of an offer to purchase any securities or instruments from any entity. Nothing from Everest Group may be used or relied upon in evaluating the merits of any investment. Do not base any investment decisions, in whole or part, on anything provided by Everest Group.

Everest Group materials, products and/or services represent research opinions or viewpoints, not representations or statements of fact. Accessing, using, or receiving a grant of access to Everest Group materials, products and/or services does not constitute any recommendation by Everest Group to (1) take any action or refrain from taking any action or (2) enter into a particular transaction. Nothing from Everest Group will be relied upon or interpreted as a promise or representation as to past, present, or future performance of a business or a market. The information contained in any Everest Group material, product and/or service is as of the date prepared and Everest Group has no duty or obligation to update or revise the information or documentation.

Everest Group collects data and information from sources it, in its sole discretion, considers reliable. Everest Group may have obtained data or information that appears in its materials, products and/or services from the parties mentioned therein, public sources, or third-party sources, including data and information related to financials, estimates, and/or forecasts. Everest Group is not a certified public accounting firm or an accredited auditor and has not audited financials. Everest Group assumes no responsibility for independently verifying such information.

Companies mentioned in Everest Group materials, products and/or services may be customers of Everest Group or have interacted with Everest Group in some other way, including, without limitation, participating in Everest Group research activities.



About VAST Data

VAST Data is the AI Operating System company – powering the next generation of intelligent systems with a unified software infrastructure stack that was purpose-built to unlock the full potential of AI. The VAST AI OS consolidates foundational data and compute services and agentic execution into one scalable platform, enabling organizations to deploy and facilitate communication between AI agents, reason over real-time data, and automate complex workflows at global scale. Built on VAST's breakthrough DASE architecture – the world's first true parallel distributed system architecture that eliminates tradeoffs between performance, scale, simplicity, and resilience – VAST has transformed its modern infrastructure into a global fabric for reasoning AI. Learn more at vastdata.com and follow VAST Data on LinkedIn, YouTube and X.