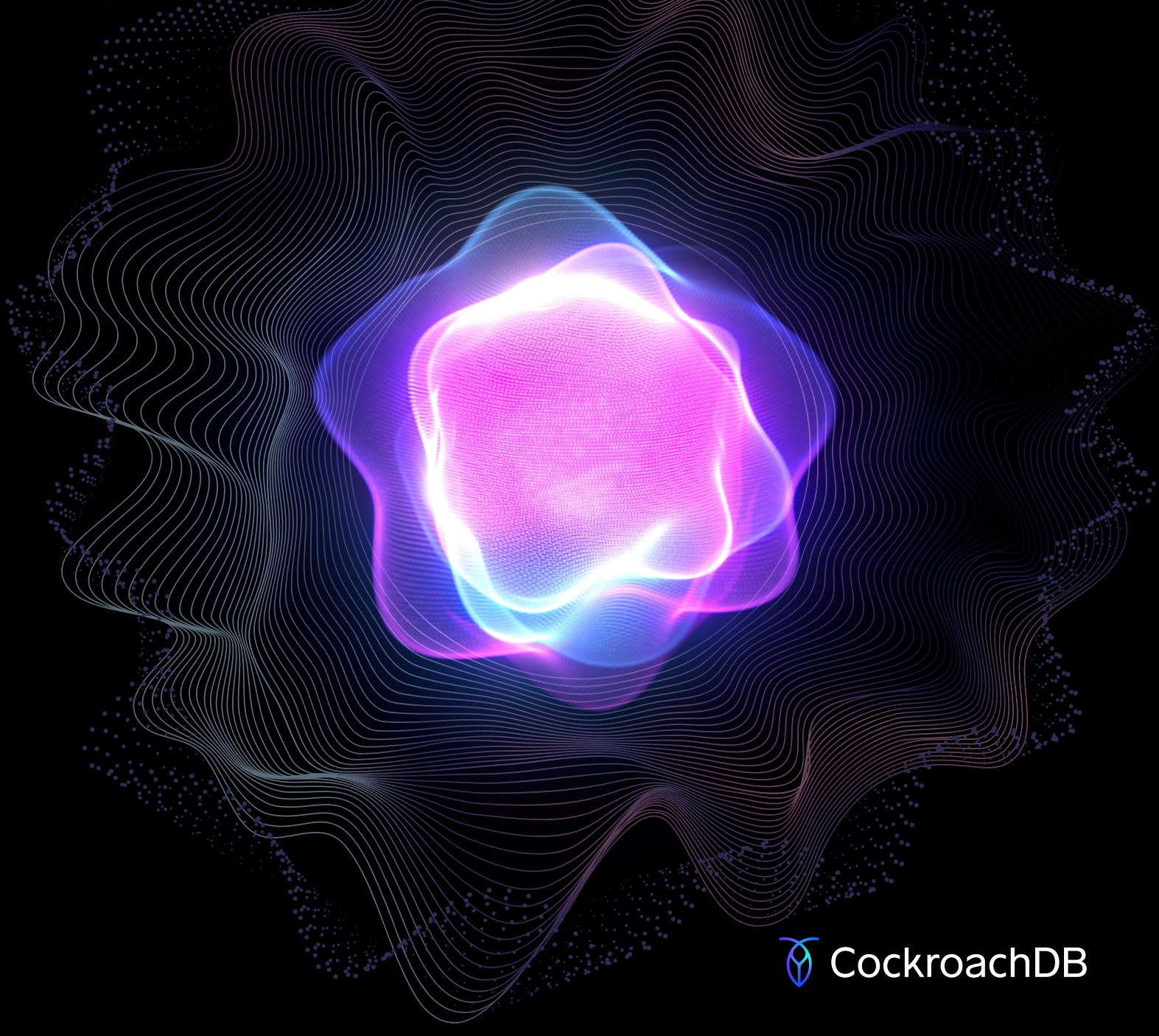
The State of AI Infrastructure 2026:

Can Systems Withstand AI Scale?



CockroachDB

Executive summary

What 1,000+ senior cloud architects and technology executives around the world are saying about their organization's preparedness for AI workloads, and their strategies for the next year

2025 may be remembered as the year of outages and the year AI hit production at scale. But there is a link between these two phenomena and their relationship needs to be investigated.

The *State of AI Infrastructure 2026* report explores a critical tension shaping today's enterprise architecture: the rapid expansion of AI workloads is outpacing the systems meant to support them. Based on a global survey of 1,125 senior cloud, engineering, and infrastructure leaders, the report reveals exactly how invested companies are in AI, the hidden cost of AI success and what top-performing teams are doing to survive and scale.

1. AI growth is no longer optional–it's guaranteed.
100% of respondents expect AI workloads to grow in the next year. More than 60% predict increases of 20% or more. At this point, AI adoption is an inevitability. The only question is how existing systems will respond to the volume and velocity of what's coming next. The data reveals an unsettling answer.

2. Infrastructure failure is expected very soon.
AI is shifting from an innovation initiative to a systems-level risk. 83% of leaders believe their data infrastructure will fail without major upgrades in the next 24 months. The fragility of current infrastructure won't be solved with routine maintenance, it requires an overhaul of how organizations approach the architectural foundation on which their systems rely.

3. For one-third of companies, the breaking point is less than a year away.
34% expect their infrastructure to fail within the next 11 months. For a significant share of enterprises, infrastructure failure related to AI scale is viewed as an imminent event, not a distant possibility.

4. The database layer is emerging as a critical point of failure.
30% of respondents identified the database as the first point of failure in an AI-overload scenario, second only to the cloud infrastructure itself. The problem is not running on the cloud, the problem is how design decisions are made with regard to how data is ingested, processed, stored, and moved.

5. AI is expected to drive a meaningful share of outages.
AI-related reliability issues are no longer hypothetical. 77% expect AI to drive at least 10% of all service disruptions in the year ahead.

6. The financial impact of AI-related downtime is already substantial and rising.
As AI systems become embedded in core business operations, outages carry immediate and material financial consequences. 98% of companies say an hour of AI-related downtime would cost at least $10,000; nearly two-thirds say it would cost over $100,000.

7. Leadership misalignment is accelerating the risk. Nearly two-thirds of respondents (63%) say their leadership teams underestimate how quickly AI demands will outpace existing data infrastructure. This suggests that while companies have been investing in AI, the investments have been too reactive, and may not truly prevent disaster.

Today, resilience is not just a best practice, it's the battleground for the next phase of AI adoption. The infrastructure powering today's enterprises wasn't built for AI-native scale, and it's showing. Recursive agents, real-time inference, and always-on automation are overwhelming brittle backends across every industry. As AI adoption continues to accelerate, the winners will be those whose infrastructure can handle unprecedented scale. Organizations that can't absorb constant, compounding demand will encounter outages, cost volatility, and degraded performance long before AI ambitions are realized. The path forward starts with distributed, resilient systems built to withstand continuous scale and operate through success, not just recover from failure.

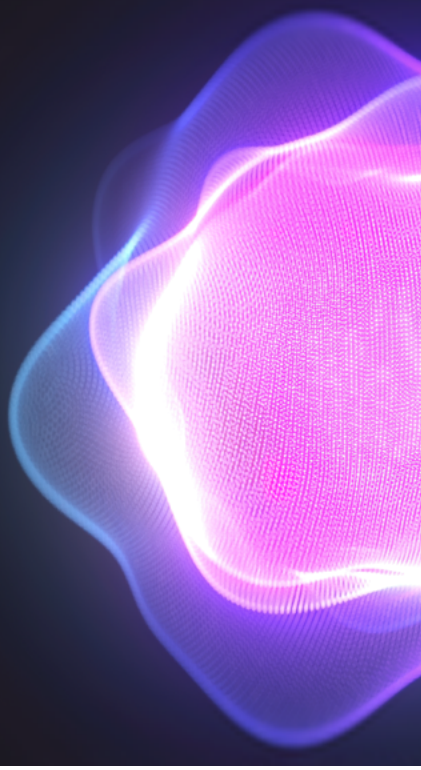
Survey Methodology

The State of AI Scale & Resilience Survey was conducted by Cockroach Labs and Wakefield Research among 1,125 Senior Cloud Architects, Engineering, & Technology Executives, with a minimum seniority of Director in 3 regions across 11 markets: North America (U.S., Canada), EMEA (Germany, Italy, France, UK, Israel), and APAC (India, Australia, Singapore, Japan), between December 5th and December 16th, 2025, using an email invitation and an online survey.

Results of any sample are subject to sampling variation. The magnitude of the variation is measurable and is affected by the number of interviews and the level of the percentages expressing the results. For the interviews conducted in this particular study, the chances are 95 in 100 that a survey result does not vary, plus or minus, by more than 3.1 percentage points in the global sample, 6.9 percentage points in the United States, 19.6 percentage points in Israel, and 9.8 percentage points in each of the remaining markets from the result that would be obtained if interviews had been conducted with all persons in the universe represented by the sample.

INTRODUCTION

In the wake of failures, AI looms large



From Google and AWS to Microsoft and Cloudflare, 2025 was defined by disruption. But it wasn't just these major service providers that went down, so did their clients. As a customer, you remember how disruptive these outages were to your daily life. Your banking apps froze, flights were grounded, retail checkouts crashed. What once felt like isolated incidents of failure, became everyday events. These failures are symptomatic of widespread structural strain. The systems behind our most essential services are buckling under the pressure of AI-scale demand. In a market where 52% of consumers will abandon a brand after a single bad experience, the stakes have never been higher.

What happens when AI success begins to outpace the systems enterprises rely on to stay online?

We conducted this report this year because AI isn't just advancing, it's accelerating. In 2025, AI workloads surged into production across nearly every industry, from customer-facing chatbots to mission-critical autonomous systems. But as AI-usage moves from experimental and episodic into always-on, production-grade systems, the question is no longer what AI can do, but whether the infrastructure behind it can keep up.

The answer is increasingly clear: it can't. A cascade of high-profile outages—from hyperscalers to the applications built on top of them—has exposed how fragile traditional systems are under sustained, autonomous demand. AI agents introduce continuous, machine-driven activity that breaks the assumptions underlying most enterprise architectures, which were built for human-paced interactions.

If 2025 marked the arrival of production-grade AI, 2026 will be defined by its scale. And that scale is already overwhelming the systems meant to support it. As this report shows, leaders are bracing for impact. The next era of infrastructure won't be shaped by tools, but by architecture built to survive success.

The state of AI scale and resilience 2026

AI is upon us, and it's scaling faster than prior technological phenomena. This report reveals a striking disconnect: while awareness of AI's infrastructure demands is nearly universal, most organizations remain unprepared for what's coming.

Unlike the internet, which gave enterprises years to evolve alongside it, AI's growth curve has been sharp and unrelenting. That speed has challenged the foundation on which all critical workloads are built on. What's different this time isn't just volume, but continuity: AI systems operate persistently, generating demand that doesn't pause, peak, or recede with human behavior.

Many companies are at risk of not just total outages, but to a range of real-world operational breakdowns: latency degradation, cascading service issues, partial blackouts, gray failures. These all threaten business continuity, brand reputation, and the bottom line. As AI workloads move deeper into production, these failure modes are becoming more frequent—and more costly—long before systems reach complete collapse.

This report unpacks this tension across five parts:

Part 1: Scaling to meet the demands of AI focuses on how AI has changed the way businesses operate, from how they are spending their money to what kinds of AI workloads they anticipate in the coming year.

Part 2: The state of today's enterprise infrastructure reveals how companies are currently deploying AI workloads, the challenges the current system faces, what are the critical points of failure in their infrastructure, and when (not if) they anticipate their infrastructure will collapse under the weight of AI.

Part 3: The cost of downtime in the age of AI uncovers exactly how expensive AI-related outages are for companies. How likely do companies think AI-related failures are? How are leading enterprises trying to bolster their systems in preparation for surges in traffic due to AI workloads?

Part 4: Why distributed SQL is uniquely positioned to handle the AI era leverages key takeaways and data points from this report to show how a distributed SQL database provides resilience and critical capabilities necessary for any enterprise, preventing disaster and withstanding the strain of AI.

Part 5: How to meet the future of agentic AI completes the report by summarizing the report's findings. What are the challenges leaders are facing today? How do they predict AI will shape businesses in 2026 and beyond? What are leaders investing in now to prepare for the future of AI?

PART 1

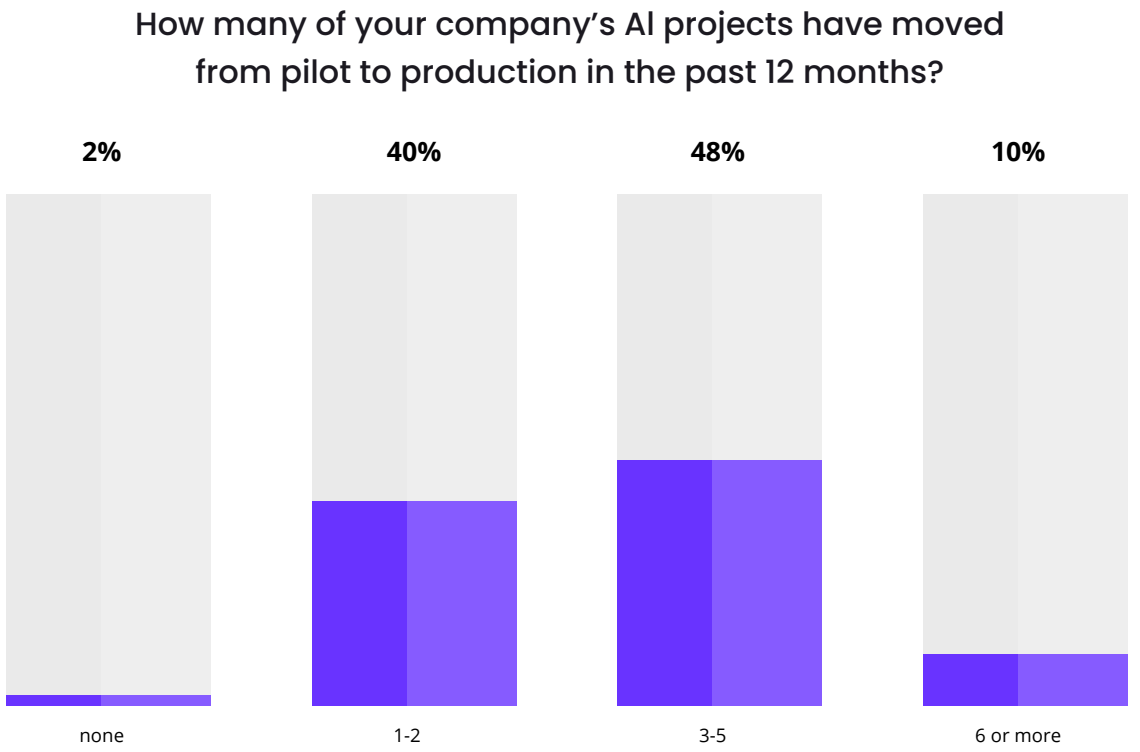
Scaling to meet the demands of AI

How AI has changed businesses, by the numbers

100% of respondents expect their company's AI workloads to increase in the next 12 months

1.1 Meeting the call for AI

In the past 12 months, 81% of all companies of all sizes have scaled AI adoption, with 26% scaling fast with major new deployments. In addition 98% of all respondents report that at least one AI project has moved from pilot to production in the last year. 57% of respondents indicated that 3 or more AI projects had made their way to the hands of customers.



These numbers reflect the demand not just externally from users but also internally from leaders to keep pace with what AI has to offer and what customers have come to expect.

This past year has shown some of the potential of AI, whether that be chatbots and virtual assistants for customer service, content generation for marketing, or AI copilots for engineering. AI has increasingly become a part of everyday life.

Two in five companies (40%) have moved one or two AI projects from pilot to production in the past 12 months. Another 48% have done so with three to five projects.

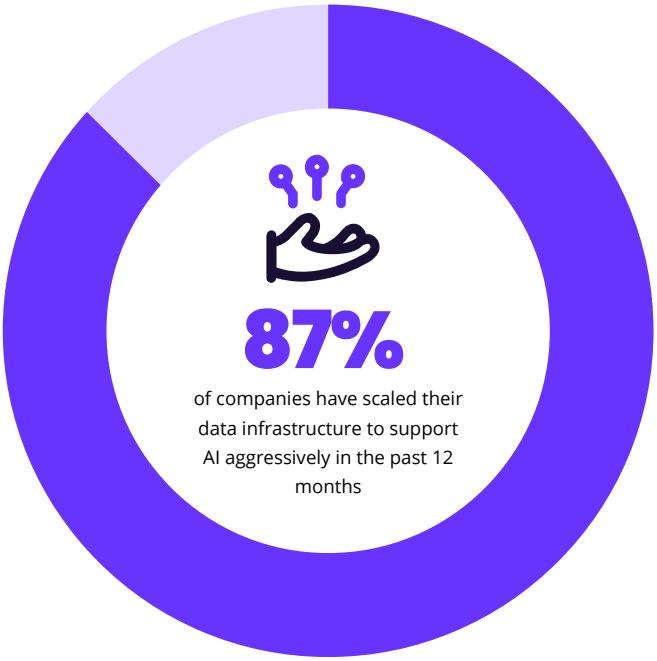
1.2 Building infrastructure that supports AI

87% of companies have scaled their data infrastructure to support AI aggressively in the past 12 months

Scale today, for tomorrow

New workloads, however, have also required new infrastructure to support them. With most companies recognizing the importance of data infrastructure in supporting AI workloads, 21% are scaling very aggressively to get ahead of the work to come.

Beyond the technical foundation for these workloads, companies are also actively investing in their AI future.



AI is today's major IT investment

85% of companies are spending 10% or more of their total IT budget on supporting AI initiatives that place significant demands on data infrastructure, such as compute scaling, model training environments, real-time data processing, and database optimizations for AI workloads.

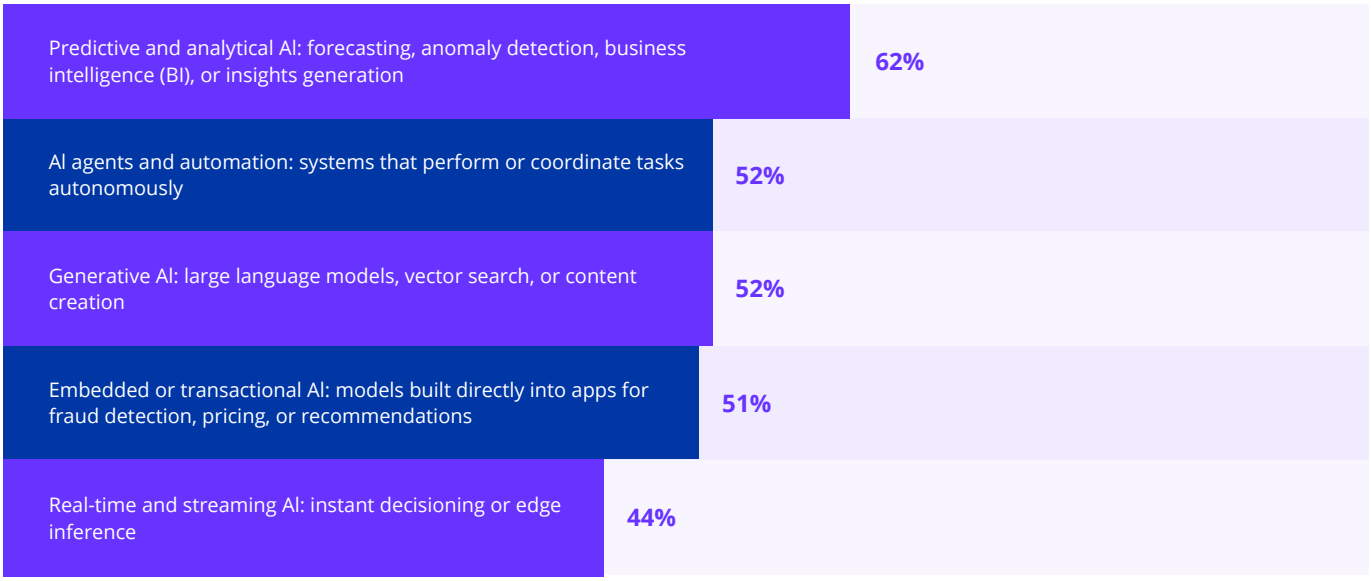
24% of companies are spending more than 25% of their total IT budget on these kinds of efforts. The increase in IT spend isn't just about the scale of workloads, it's the uniqueness of AI that forces new requirements on underlying systems. Unlike traditional application workloads, AI workloads, especially those involving real-time inference and agentic automation, place sustained, unpredictable pressure on databases. The database is at the heart of AI workloads because it must ingest more data than ever before and process more transactions.

Databases that weren't purpose-built for scale under chaos and pressure simply aren't equipped to handle AI. Instead of being triggered by known human behavioral patterns, like Black Friday, Super Bowl Sunday, and other busy seasons, AI workloads generate continuous queries, writes, and coordination tasks. This breaks conventional wisdom about usage spikes and traffic behavior.

1.3 The future of AI workloads

As companies look to the next year, all of them expect their AI workloads to increase even more. 63% of companies expect the number of AI workloads to increase by more than 20%.

Perhaps surprisingly, more established companies with 20+ years in business expect their AI workloads to double as opposed to younger companies (19% of the more established companies expect their AI workloads to increase by at least 51%, while only 9% of the less established companies expected the same.)



When asked what kinds of AI workloads they expected would most influence their company's data-infrastructure strategy in the next two years, answers varied:

- **Tried and true AI use cases:** 62% of respondents surveyed cited predictive and analytical AI such as forecasting, anomaly detection, business intelligence (BI), and insights generation as a key driver
- **Agentic AI is here:** 52% of respondents surveyed selected AI agents and automation as a critical driver of data-infrastructure strategy
- **GenAI uses continue to swell:** 52% of respondents anticipate generative AI workloads will carry heavy influence in the next two years

While we can only wait to see what the next couple of years hold, it's clear that leaders and companies across the world and across industries know they have to create environments that can not only support AI workloads but allow them to thrive, even under immense pressure.

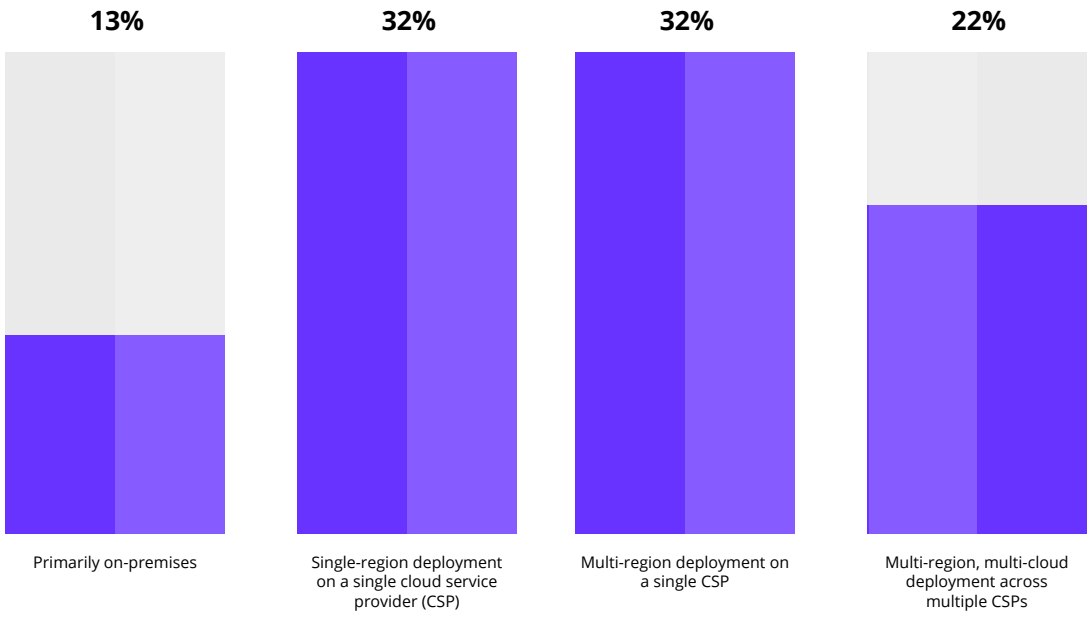
PART 2

The state of today's enterprise infrastructure

The cost of keeping up with AI is ever-present and growing

2.1 The data infrastructure landscape

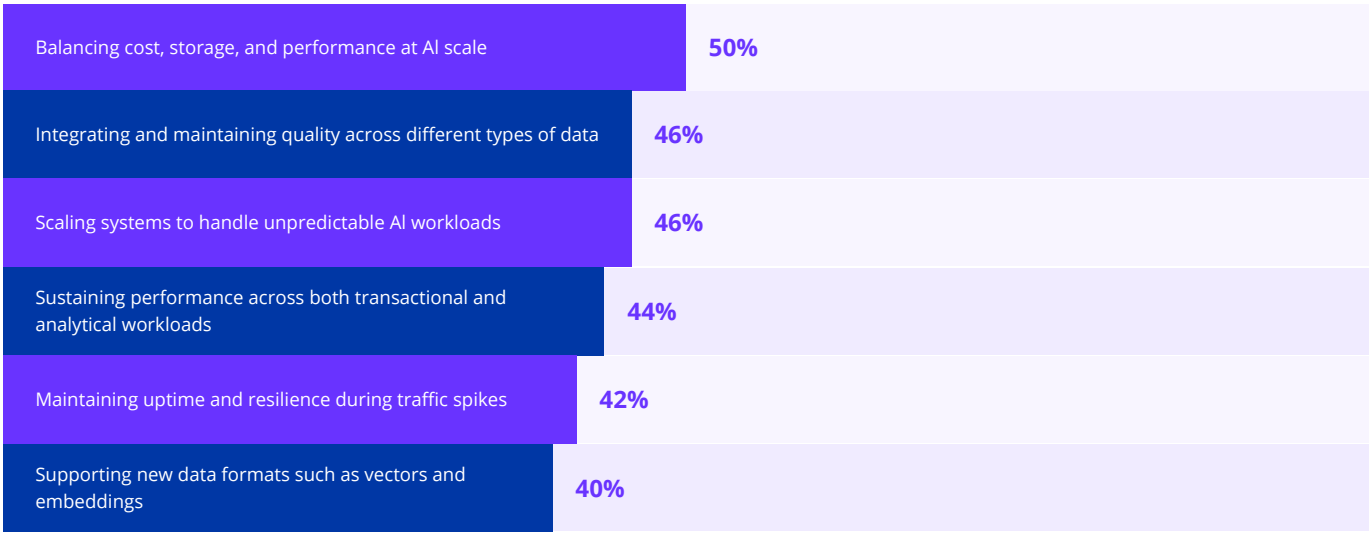
87% of organizations are currently using cloud service providers (CSPs) to deploy their AI workloads



Enterprises run on the cloud, but architecture-at-scale is the problem

When asked how their organizations are deploying AI workloads, a vast majority (87%) indicated they were running on cloud service providers (CSPs), with a fairly even split between single-region on a single CSP (32%) and multi-region on a single CSP (32%), with a smaller portion running multi-region, multi-cloud deployments across multiple CSPs (22%).

Over recent years, we've seen the rise of cloud deployments, and how companies have chosen to run AI workloads only confirms the dominance of the cloud in today's enterprises. However, just because you're running in the cloud does not mean you have an inherently brittle or resilient system in place. While cloud infrastructure provides the raw materials, it's the data architecture layered on top that determines whether systems thrive or fail at AI scale. Your architecture is a culmination of your design decisions–how is your data collected, stored, accessed, moved, and secured across systems, whether in the cloud, on-prem, or hybrid?



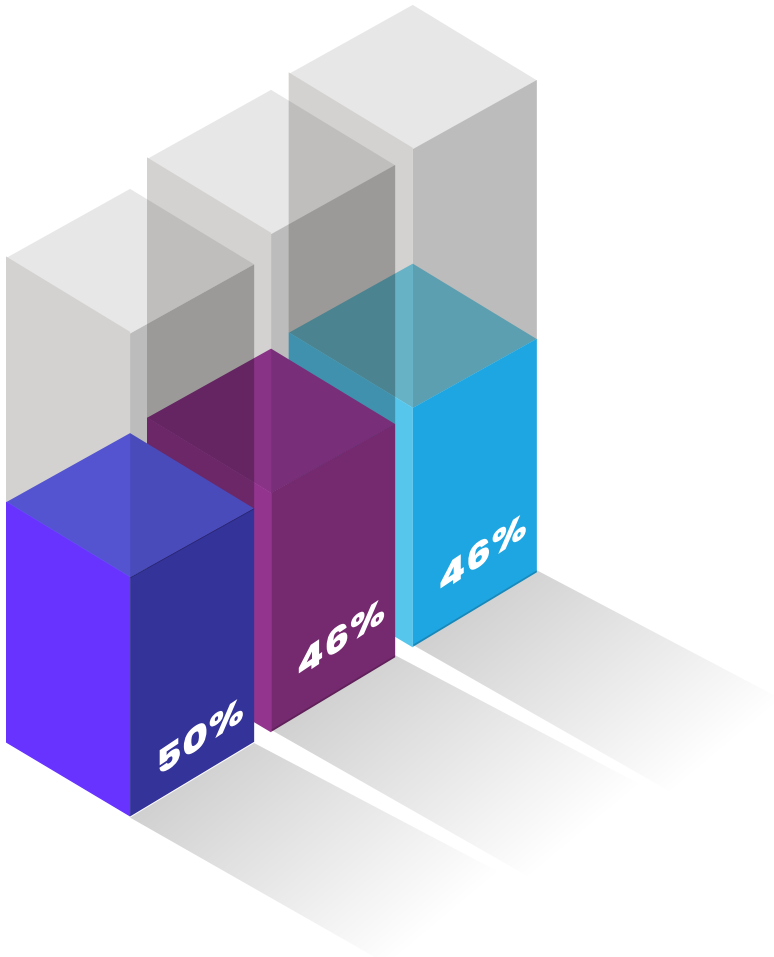
The challenges of AI and agent-driven systems

Given the tectonic shifts that AI has caused over the past several years, it's unsurprising that companies are facing a number of challenges. But exactly what each organization is tackling varies across the board. The percentages do not vary much across company revenue, size, years in business, or location.

The three most cited challenges were:

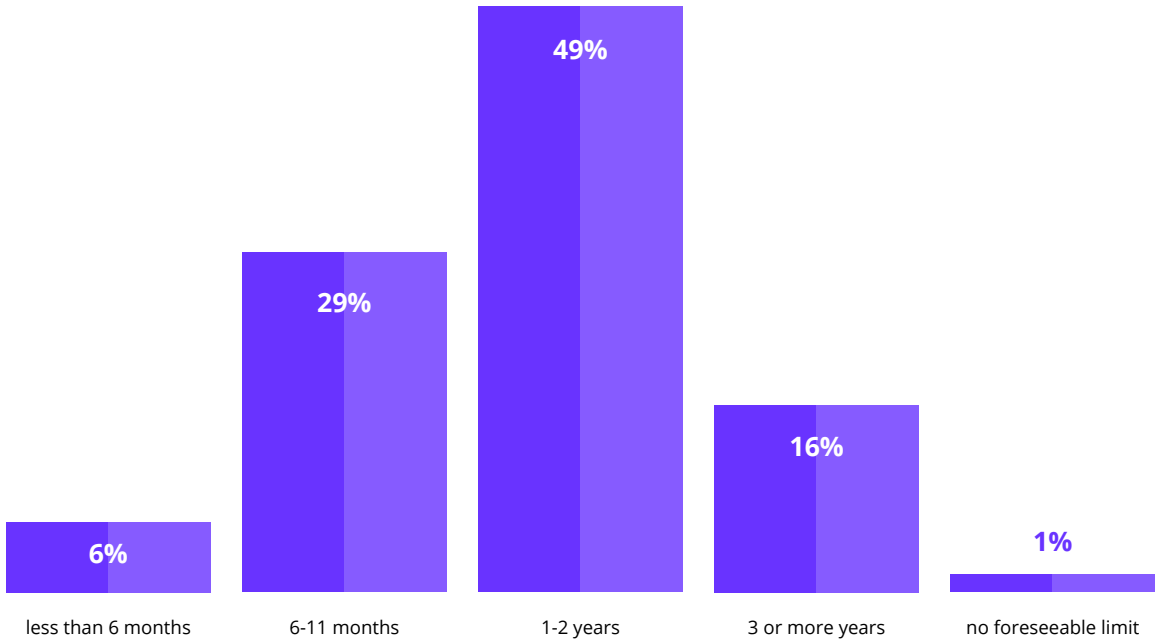
- Balancing cost, storage, and performance at AI scale (50%)
- Scaling systems to handle unpredictable AI workloads (46%)
- Integrating and maintaining quality across different types of data (46%)

The consistency across companies suggests that similar tactics may be used to solve the problem of AI scale and resilience.



2.2 When and how leaders believe traditional infrastructure will hit its limits

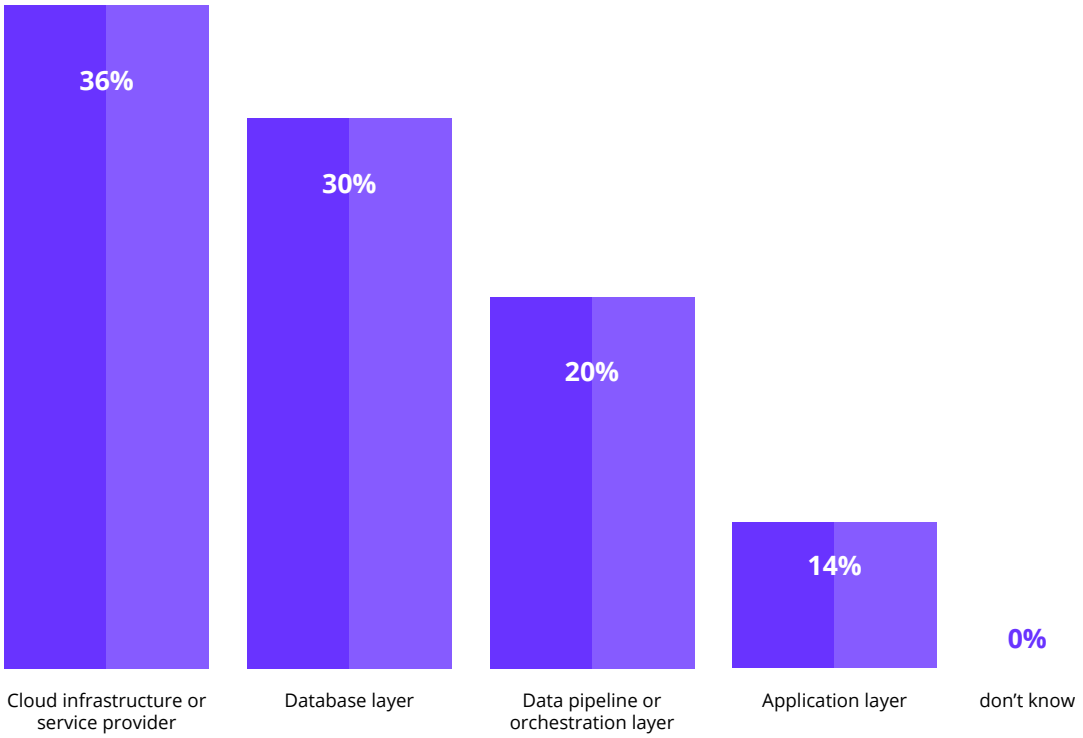
83% of companies expect their data infrastructure to hit its limit for supporting AI growth without major upgrades in the next 2 years



Recently, it seems that the thirst for leveraging AI only grows more every day. Unfortunately, traditional infrastructure cannot keep up with that growth without serious maintenance and upgrades.

With the majority of respondents believing their data infrastructure will crumble under the weight of AI in the next two years (83%) and over one-third of companies believing their infrastructure won't last the next 11 months (34%), leaders expect operational breakdowns under AI scale. The urgency of the issue goes beyond routine maintenance and demands a strategic pivot as soon as possible.

For more established companies with over 20 years in business, the threat is even larger, with 40% of respondents believing their infrastructure will fail in the next 11 months.

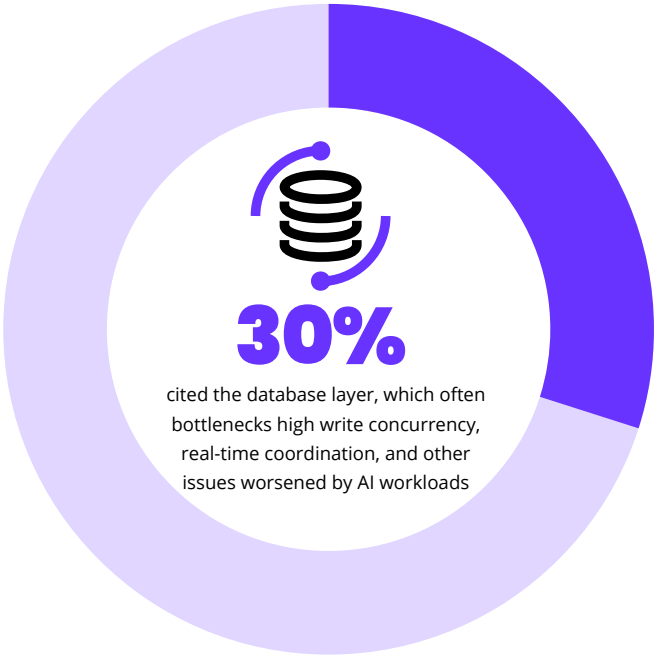


In gauging what would first fail if AI workloads exceeded system capacity,

- **36%** believed it would be their cloud infrastructure or service provider
- **30%** cited the database layer, which often becomes the bottleneck with regard to high write concurrency, real-time coordination, and other issues that will only be worsened with AI workloads

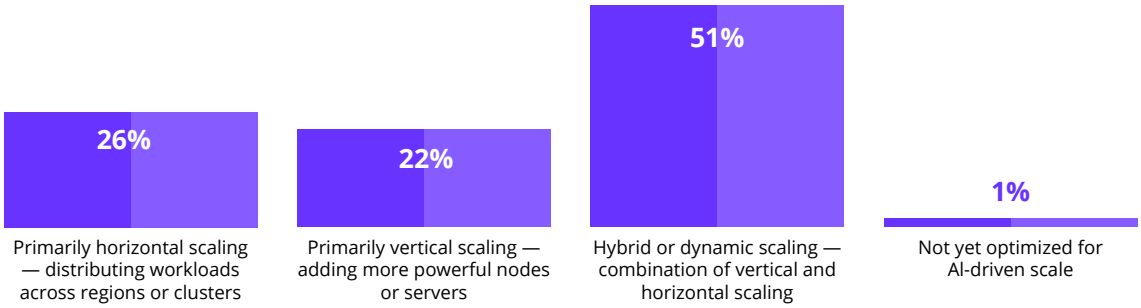
Cloud infrastructure provides the platform, but it's the data architecture, especially at the database layer, that determines whether a system can survive scale. As AI workloads move from sporadic to continuous, the coordination burden on databases increases exponentially, exposing weaknesses that cloud elasticity alone can't solve.

Compounding the issue is a gap in executive awareness. **63% of respondents agreed that leaders at their company severely underestimate how quickly AI needs could outpace their data infrastructure's capacity.** When separated by level, 65% of respondents who were C-suite, owner, or partner agreed. Even among leadership, the sentiment persists that AI needs are truly growing at an unprecedented rate. This underestimation is not only a strategic blind spot, but suggests that systems may fail even sooner than the compressed timelines articulated in this study.



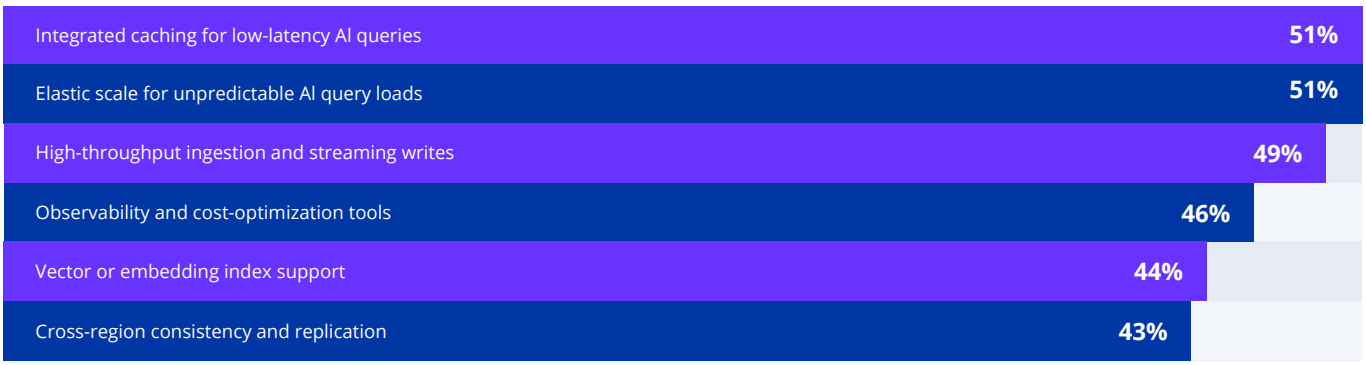
2.3 What leaders are betting on to succeed under the pressures of the AI hype

Given the number of unknowns with regard to the future of AI, companies and leaders are taking a multi-pronged approach to improving AI scalability and database performance.



As AI data and compute demands continue to rise, companies are slightly favoring horizontal scaling (distributing workloads across regions or clusters) (26%) over vertical scaling (adding more powerful nodes or servers) (22%). 51% of companies have focused on hybrid or dynamic scaling, combining both horizontal and vertical scaling.

Leaders Are All In on AI Infrastructure—But Divided on Where to Focus



Of all respondents, 99.6% are prioritizing investments to improve AI scalability and database performance in the next year. Given how spread out the focus of investment is, we can see that leaders are hedging their bets on exactly how to spend their resources.

Whatever ultimately is the focus for each individual company, it is clear that a vast majority (83%) agree that within the next 12-18 months, AI demand will exceed the capacity of most organizations' data infrastructure and the database layer could become a critical point of failure.

With the database layer emerging as a key point of failure, the urgency isn't just about spending more, it's about spending smarter. Building for resilience at AI scale requires a shift from short-term stopgaps to long-term architectural change. These so-far reactive investments explain why confidence in infrastructure readiness remains low, despite high levels of spending.

Ultimately, AI doesn't introduce entirely new risks, but amplifies existing ones—compressing planning horizons and exposing architectural assumptions faster than organizations expect.

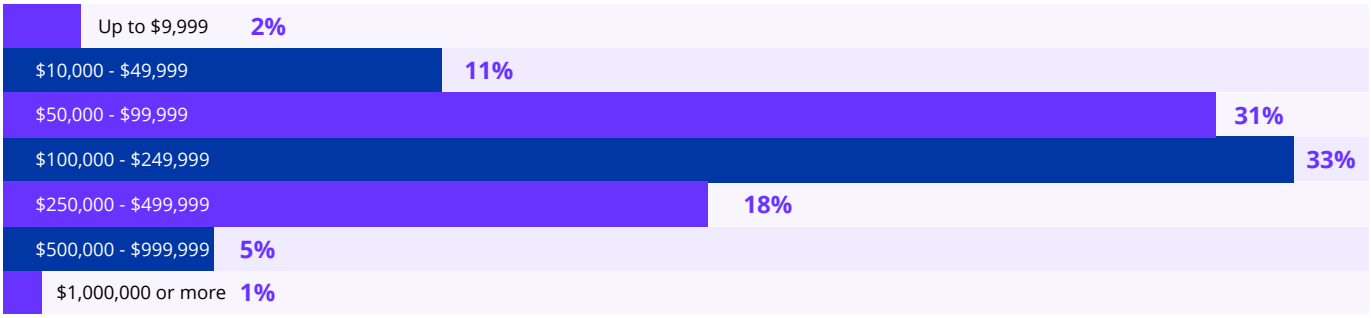
PART 3

The cost of downtime in the age of AI

After a year of outages in 2025, the cost of downtime is only scaling up with AI use cases

3.1 The true cost of AI-related downtime

98% of companies reported that just one hour of AI-related downtime would cost $10,000 or more

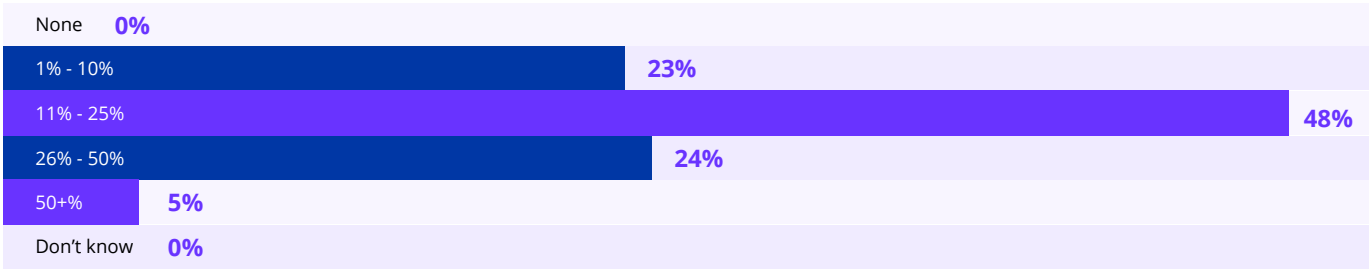


If there was an issue that could cost well over $10,000 per hour wouldn't you want to fix it as soon as possible? That's just how important it is in 2026 to ensure your entire system can handle AI workloads. What's more, a majority of respondents (57%) reported that an hour of AI-related downtime would cost $100,000 or more.

The bigger the company, the bigger the cost: Larger enterprises often have more tightly coupled systems and higher customer expectations. The data supports the hypothesis that these circumstances will only amplify the impact of AI-driven disruptions. Almost 10% of respondents representing companies making $500M in revenue or more per year projected a cost of over $500,000 per hour of AI-related downtime (8%). 63% reported a cost of $100,000 or more per hour of AI-related downtime, while 48% of those working at companies making less than $500M in revenue per year reported the same.

When 99.9% availability is almost 9 hours of downtime per year, $100,000 or more per hour of downtime adds up quite quickly.

Companies Are Bracing for AI-Induced Disruptions



While there are other causes of downtime, AI workloads are quickly creeping up as a major cause and reshaping resilience risk:

- **77%** of people surveyed predicted that over 10% of their company's outages or service disruptions in the next year would be related to AI demand or workloads.
- **Almost one-third (29%)** of respondents predicted that over 25% of their company's outages in the next 12 months would be related to AI demand or workloads.

To be clear, downtime is not just about lost revenue, it's also about the erosion of trust that is absolutely inevitable in the wake of failures. When combined with the increased cost of AI-related downtime, these numbers predict bleak outcomes for companies unprepared for AI workloads and potential outages.

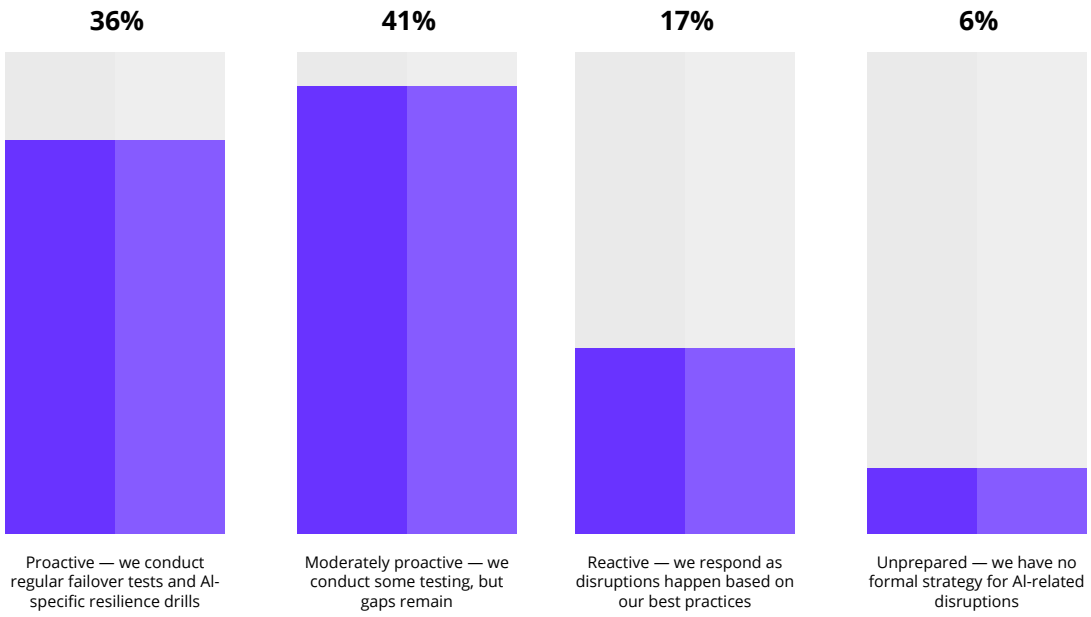
3.2 Preparing for outages

98% of companies have modeled the potential cost of downtime and/or stress-tested their systems in preparation for AI-driven traffic surges

Looking ahead to 2026, companies are gearing up for the inevitable stress that will be placed on their infrastructure:

- About one-quarter are preparing for the oncoming AI-driven traffic surges by modeling the potential cost of downtime (24%)
- Another quarter is stress-testing their systems (22%)
- More than half of companies have done both (52%)
- Only 2% of companies have done neither

Given the signals from peer companies, it seems the best path forward is to both model the cost of downtime and stress-test your system to ensure your company is staying ahead of the curve and providing the best possible outcome for your customers.



Despite modeling and testing, there are still significant areas of improvement with regard to AI-related resilience. When asked about their organizations' preparedness to respond to significant service disruptions caused by AI workloads:

- The category most selected by respondents was "moderately proactive," such that they knew of gaps in their company's testing (41%).
- Almost a quarter of respondents felt their company was reactive or unprepared for AI-related disruptions (23%).

With almost two-thirds of people (64%) surveyed noting at least some gaps in their company's current strategy surrounding AI-related disruptions, companies need to be looking more proactively at how to combat what could be a disastrous amount of strain placed on their infrastructure.

AI-related downtime compounds quickly: financially, operationally, and reputationally. Today's enterprises are feeling the pressure to rethink their infrastructure assumptions in order to succeed in the near-term and long-term.

PART 4

Why distributed SQL is uniquely positioned to handle the AI era

Based on the findings of this report, it is clear that organizations have an urgent need to modernize their data infrastructure. A distributed SQL database that unifies operational and AI data provides the scalable foundation required to keep pace with the demands of the AI era.

4.1 From findings to implications

The data examined above reveals that AI is changing how systems fail. AI workloads are scaling faster than enterprise infrastructure can adapt, compressing failure timelines across the board.

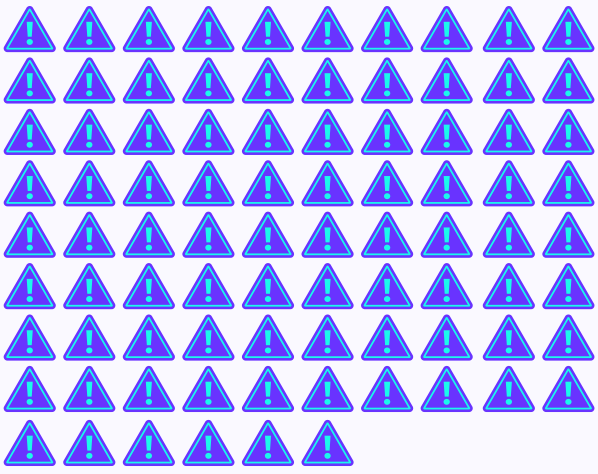
But the pattern is deeper than outages alone. The failure modes themselves are evolving. Unprecedented agentic behavior will only stress systems further in ways that traditional architectures have never been tested before.

The takeaway? A new kind of architecture, one that is built for resilience at scale is needed more than ever.

4.2 Scale-driven failure modes require a different data architecture

For cloud systems at scale, failure has always been a question of "when" not "if." Our *State of Resilience 2025* Report surveyed 1,000 senior technology leaders, and found that on average, organizations experience 86 outages each year. But with AI being brought to production, how systems are failing has also been transformed.

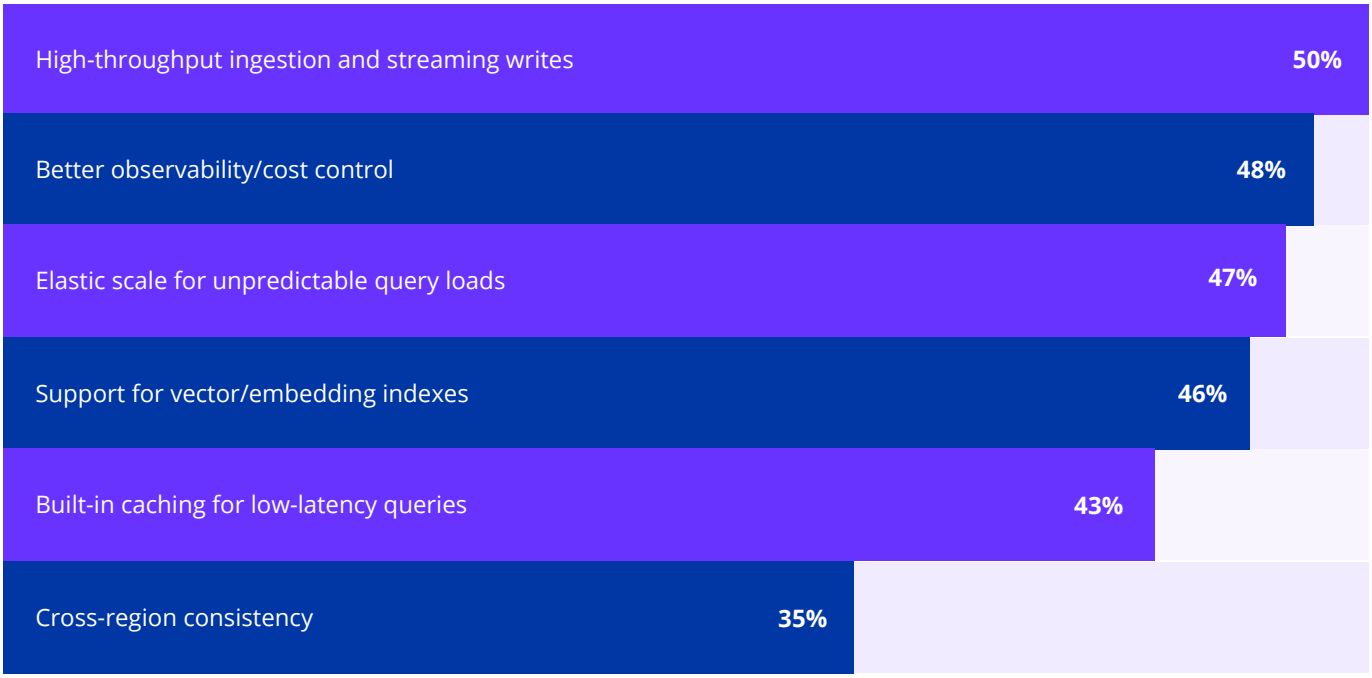
- **Continuous load:** AI agents don't sleep. Unlike human-initiated traffic spikes (Black Friday, viral campaigns), AI workloads are always-on and intensifying.
- **Concurrency under stress:** Real-time systems like recommendation engines and AI copilots issue simultaneous reads and writes, demanding database engines that can handle high-throughput with minimal contention.
- **Coordination at scale:** Agentic AI chains actions across systems and services—introducing interdependent transactions that strain consistency models, especially across regions.
- **Degraded performance, not just outages:** The database layer (identified as the second most likely point of failure) often degrades quietly via latency, contention, and deadlocks, long before a total collapse.



86 **On average, organizations experience 86 outages each year**

4.3 What “resilience at AI scale” actually requires

When asked about new database capabilities most critical to supporting future AI workloads, responses highlighted the need for higher-throughput ingestion and a way to manage rising costs and unpredictable loads due to the sheer scale of AI.



Surviving this new class of failure requires more than cloud capacity, it demands a system built for resilience in motion. That means architecture with:

- Global distribution by default
- Multi-active availability that lets any node serve reads and writes
- Built-in fault isolation to prevent cascading failure from a single node or region
- Transactional consistency under load, with strong guarantees at global scale
- Automated rerouting in the event of failure, recovery, and data movement, without human intervention
- Elastic scaling both up and down that adapts in real time to unpredictable demand

These aren’t theoretical features—they’re architectural imperatives in the age of agentic AI.

4.4 Why distributed SQL meets those requirements

Enter distributed SQL: a relatively new approach to relational databases built from the ground up to solve precisely these failure modes.

CockroachDB, the first commercially available distributed SQL database, embodies this shift:

- **Horizontal and seamless scale without sharding:** CockroachDB's nodes can each read and write. Therefore, the core architecture provides horizontal scalability by adding nodes and automatically rebalancing data to incorporate the node into the cluster. This enables seamless and elastic scaling on-demand without manual sharding as AI workloads evolve. Benchmarks show near-linear increases in throughput with added nodes, maintaining stable latencies as the cluster grows.
- **Multi-active availability:** CockroachDB will automatically service queries from healthy nodes in the event it detects a failure, all in real-time, without any manual intervention. Like traditional active-active designs, all replicas can handle both reads and writes, but CockroachDB was also designed to ensure transactional consistency across replicas. As a cloud-native database, replicas can be stored across cloud providers, ensuring that even a CSP-outage is a survivable event.
- **Strong consistency with ACID transactions:** CockroachDB implements globally consistent ACID transactions using Raft consensus. Each write is replicated to a quorum of replicas before committing, ensuring strong consistency even across regions and high concurrency.
- **Range/replica locality & query routing:** CockroachDB's replication and data locality features help reduce query latency by keeping data close to users.
- **Enhanced vector indexing:** CockroachDB built out and supports a novel approach to vector indexing designed for billions of vectors, enabling high-performance similarity search embedded directly in the database. This is intended for LLM/AI applications and provides distributed, update-efficient vector indexes.
- **Built-in metrics & dashboards:** CockroachDB exposes extensive performance metrics that can be viewed in the DB Console or exported to third-party tools. This includes SQL performance, replication, storage, latency, contention, and more. The CockroachDB Console offers real-time and historical insights into workloads, SQL activity, latency, retries, and more, helping pinpoint performance bottlenecks and optimization opportunities.

Even with all of the focus on disaster prevention, CockroachDB also provides critical disaster recovery tools in the event that the worst happens:

- **Backup and Restore:** Options include full and incremental backups, scheduled backups, point-in-time restore (PITR) and more to ensure your data is protected no matter what happens externally.
- **Physical Cluster Replication (PCR):** Continuously and asynchronously replicates data from a primary cluster to a passive standby cluster. PCR ensures transactional consistency while replicating, and the standby cluster is able to serve traffic within minutes with only seconds of data loss from replication lag.
- **Logical Data Replication (LDR):** Continuously replicates tables between a source CockroachDB cluster and a destination CockroachDB cluster. Both clusters can receive reads and writes, and the active-active setup protects against cluster, data center, or region failures.

Where legacy databases were designed for predictable traffic and single-region deployments, distributed SQL systems like CockroachDB are purpose-built for a world where AI is the driver, not the edge case. By fulfilling the promises of truly distributed, resilient systems, distributed SQL databases are poised to be the backbone of AI use cases in 2026 and beyond. With unmatched scale, strong consistency, resilience, vector and transactional data capabilities, as well as cost efficiency, companies can rely on distributed SQL to weather the growing AI demand.

4.5 Performance under Adversity: Proof, not promise

Looking to 2026, it will not be good enough to simply recover from disaster. Global enterprises who had outages in 2025 suffered from negative media coverage, consumer outrage, and lost revenue. It's in your company's best interest to invest in solutions that approach disasters from a preventative lens.

Most databases, which 30% of respondents identified as the first point of failure in the event of AI workloads overtaxing their system, are being measured with an outdated benchmark: TPC-C. First approved more than 30 years ago, TPC-C measures performance under perfect conditions, not within the chaotic environment of the real world. Designed under the premise of single data-center deployments without measuring performance during failures or maintenance, companies and customers deserve much more rigor in 2026.

CockroachDB is leading the charge with its novel benchmark, Performance under Adversity, which not only measures throughput under normal conditions, but puts this distributed SQL database to the test under increasingly stressful conditions, including disk stalls, network failures and even regional outages.

Through each phase of failure, throughput (tpmC), latency (90th and 95th percentile), and recovery time to baseline are measured. The results of the benchmark showcasing CockroachDB's steady throughput, low latency, and lack of downtime can be seen via an interactive dashboard.

While most database companies are not testing their products to this degree (or are at least not publicizing the results), you can trust that at least one distributed SQL database is providing the necessary proof of performance under stress.

4.6 What distributed SQL enables for AI-native systems

Distributed SQL, and specifically CockroachDB, closes the widening gap between AI ambition and operational reality. It provides a single, resilient, and scalable foundation that enables AI-native systems not just to function, but to thrive under pressure.

CockroachDB empowers organizations to move from deterministic SQL queries to long-lived, autonomous agentic workflows without rearchitecting or losing context. It's not just a new kind of database, it's a new kind of coordination layer for modern AI systems.

Unify structured and semantic context: CockroachDB natively combines transactional data, vector embeddings, and agent memory in a single system. This enables agents and applications to reason across historical records, relational facts, and contextual semantics without relying on glue code or cross-system syncs.

Persist and scale agent memory: As agents shift from prompt-response cycles to long-lived, autonomous tasks, they require durable, consistent state. CockroachDB supports agent memory as first-class data, ensuring context is preserved across sessions, failures, and evolutions in behavior.

Operate at machine pace, not human pace: Agentic systems don't take breaks. They act continuously and concurrently. CockroachDB delivers strong consistency at scale, allowing thousands of agents to read and write safely in parallel without race conditions or logic drift.

Eliminate fragmentation and migrations: Most AI systems today rely on piecemeal architectures, mixing PostgreSQL, vector stores, and other tools. CockroachDB replaces this sprawl with a single operational platform, so teams can start small, iterate quickly, and scale AI workloads without rewrites or loss of control.

Enable full AI lifecycle maturity, from prototype to autonomy: Whether you're storing feature metadata, powering production-grade semantic search, or running autonomous agents across regions, CockroachDB provides the same foundation. It evolves with your AI maturity.

PART 5

How to meet the future of agentic AI

Priorities in meeting the call for AI and preparing for what's to come

Operational resilience has become increasingly critical over the years. With the age of AI finally upon us, the stress on our systems has reached unprecedented levels. But what's changing in 2026 isn't just the volume of AI workloads, it's the nature of them.

The rise of agentic AI in production marks a distinct turning point. AI agents are autonomous, persistent, and self-coordinating, triggering continuous load, making real-time decisions, and chaining tasks across services without waiting for human input.

This survey of 1,125 global enterprise technology leaders highlights existing oversights for organizations today, as well as upcoming trends as the world continues to turn to AI for answers. While all respondents expect AI workloads to increase in the next year, almost 90% also expect their company's data infrastructure to fail in the next 2 years without major upgrades.

AI is accelerating faster than organizations have ever had to adapt before. Looking ahead, agentic AI only promises to keep rewriting the rules. The future will belong to companies whose infrastructure is designed not just to scale, but to stay online in the face of unpredictable, relentless demand.

As AI-usage accelerates, from generative models to autonomous agents, so too does the pressure on the systems meant to support it. What was once a gradual adoption curve has become a tidal wave of scale, catching many organizations off guard. The gap between ambition and infrastructure readiness is quantifiable, and it's shrinking fast.

Cost of AI-related downtime: With just one hour of AI-related downtime costing over $100,000 for nearly two-thirds of large enterprises (66% of companies with 1000+ employees, 63% of companies with $500M+ ARR) being unprepared is no longer an option. Furthermore, more than three-quarters of survey respondents predicted that over 10% of their company's outages or disruptions in the next year would be AI-related (77%). Without specifically preparing for AI workloads, companies are risking more than just dollars, but also customer trust. As it becomes easier for customers to find other service providers, brand reputation may be worth just as much as the revenue.

Today's data infrastructure isn't ready–and companies are throwing money at the problem: Over one-third of respondents believe their data infrastructure will fail under the weight of AI in the next 11 months (34%). 83% agree that AI demand will exceed the capacity of most organizations' data infrastructure in the next 12-18 months. With regard to modernizing data infrastructure, this is an incredibly tight timeline. It's now or never and companies are investing: 99.6% are prioritizing investments to improve AI scalability and database performance for 2026.

AI agents and generative AI will shape 2026: Over half of respondents (52%) expect AI agents and automation will be most influential on their company's data infrastructure strategy in 2026. Another 52% expect generative AI will be most influential. While 2025 saw AI being brought to production at scale, agentic AI and generative AI will become more and more strategically significant in the new year. These technologies will allow the autonomous coordination of tasks, mass content creation, and more.

Preparedness in 2026 means resilience, compliance, and AI-readiness

As a response to this complicated landscape, enterprises are looking at a range of strategic initiatives to arm themselves against the weight of AI scale:

- **Building an AI-ready foundation:** Whether it's investing in the underlying infrastructure on which AI relies or exploring new AI use cases, AI-driven initiatives are at the forefront of most companies. Being AI-ready in 2026 means harnessing infrastructure that can absorb constant growth and volatility without requiring human intervention.
- **Preparing for unpredictable scale (and cost):** With AI workloads, and especially AI agents increasing the demand placed on companies, scale is becoming more volatile. Choosing the most cost effective solutions will become paramount as scale reaches unseen heights.
- **Investing in resilience:** As the cost, financial and otherwise, of AI-related downtime only goes up, resilience is increasingly about preventing degradation at scale. Not all data infrastructure is built equal. Relying on infrastructure that puts resilience first is critical to ensure not just disaster recovery but also disaster prevention.

Resilience has long been the most difficult and pressing challenge for mission-critical applications. But AI is about to make scalability the defining problem of the next era. Unlike previous technology shifts, AI (particularly agentic AI) does not simply increase usage. It introduces autonomous, persistent, machine-driven activity that compounds continuously. As agents interact with APIs, tools, and each other, they generate sustained operational demand that places unprecedented stress on the data architectures behind modern applications.

The scale of this shift is without historical precedent. While agentic activity remains small relative to human activity today, that balance is poised to flip rapidly. As autonomous systems proliferate, machine-generated traffic will vastly exceed human-driven workloads, compressing timelines and exposing architectural limits faster than organizations have ever experienced. Applications not designed for this level of scale will be forced to modernize–or fail under the weight of their own success.

This is why the findings in this report matter now. AI is scaling faster than any platform shift that came before it, including the rise of the internet, and the infrastructure assumptions that carried enterprises through past waves of growth will not hold. In the age of agentic AI, scalability is no longer a future concern. It is the constraint that will determine which systems endure and which do not.

cockroachlabs.com