2026

Confluent Predictions

CONFLUENT

In 2025, the promise of AI services with the agency to reason, make decisions, and take actions has taken over data discussions.

Successfully bringing AI to production currently requires large-scale human investment and has kept organizations glued to their AI strategy. Despite this, we saw a sobering number of failures with AI projects.

One lesson continues to remain true with every new AI generation: **all AI problems are data problems.** It is clear that the data community is at the center of one of the most profound social revolutions since the Industrial Revolution. CIOs and CFOs are looking for strong returns on AI investments, but this won't come to fruition until the agentic AI ecosystem matures. If 2025 was focused on racing to bring agentic AI to production, 2026 will be about addressing many of the barriers that made it hard to do so.

Many of Confluent's customers are seeing the impact of data challenges on AI readiness—and this is validated by the broader market. In fact, 68% of IT leaders recently cited data silos as a major roadblock for AI success.[1] With input from our Technology Strategy Group, experts in the field and our partner ecosystem, as well as insights from industry analytics, we've put together this 2026 Predictions Report to answer burning questions about this rapidly evolving landscape. The predictions here are focused on what's coming next and how 2026 will be the year to get ahead of the curve as the agentic AI ecosystem continues to mature.

[1] 2025 Data Streaming Report

> "Data streaming is going to be critical for financial services organizations in the future—especially when it comes to AI/ML use cases where the ability to stream data, implement workflow automation, and make real-time predictions are the key to success."
>
> **SHUJAHAT BASHIR**
> **DIRECTOR OF ENGINEERING**

thrivent

Contents

PREDICTION #1

The Rise of Agentic Commerce: Machines Are Your New Customers

Machine-to-machine transactions have long powered industries like financial services[2] and logistics, but in 2026, entirely new sectors will need to learn how to effectively support and sell to AI agents acting as customers.

Consider how much time it can take to find the best deal on something as mundane as a plunger or something more important, like choosing the right insurance policy. Most people don't like to spend hours researching these decisions—and increasingly, they don't have to. These are examples of consumer transactions that will start to transform in 2026—with AI agents handling more of these routine purchases optimizing across price, quality, and convenience.

Importantly, this shift won't stop at the consumer level. Enterprises will see the same transformation across supply chains and B2B relationships. Companies will need to figure out how to optimize sales and marketing for the machines that will increasingly do the decision making, and in some cases, ultimately the buying. If you think human customers are fickle, machine customers can be ruthless: they have zero patience for latency, no brand loyalty, and can switch vendors mid-transaction whenever a better offer appears. If your systems can't interface with them and respond instantly and reliably, you might as well be invisible to them.

[2] Automating Finance: The Rise and Impact of Machine-to-Machine Payments

WHAT THIS MEANS FOR YOU:

To prepare, businesses need to start opening their commerce ecosystems. That means exposing agent-safe APIs, adopting tokenized payment protocols, and making real-time product data available. With the right real-time data pipelines in place, you can build an always-synchronized commerce layer—one where autonomous agents can search, decide, and purchase instantly. The result? Commerce that evolves into a dynamic, real-time ecosystem of machine-mediated transactions.

Recommended Reading

CONFLUENT BLOG

From Reactive to Orchestrated: Building Real-Time Multi-Agent AI With Confluent

PREDICTION #2

Leading Platforms Will Offer Model Context Protocol

One thing that the agentic AI revolution has taught us in the past year is that it takes hard work to do it well. One of the biggest challenges remains in ensuring that these autonomous AI systems have access to the right data at the right time and the ability to take appropriate actions for them to successfully deliver value.

To address this, the major large language model (LLM) providers have each introduced their own approaches to external tools and function calling that AI application developers can use. But this doesn't hold up in a market where AI applications and developers need to easily swap between LLMs and providers.

In 2025, we saw Model Context Protocol (MCP),[3] an open source standard for connecting AI applications to external systems, rapidly became the standard "context-oriented protocol" with several early adopter data and technology platforms adding MCP support to enable access to AI applications, regardless of the underlying LLM.

While significant hurdles remain for MCP, particularly around security, the gravitational pull towards a single, easy open protocol that reduces friction and developer overhead will prove irresistible in 2026. While other competing standards like Agent2Agent (A2A) and Agent Communication Protocol (ACP) continue to vie for relevance in agent-to-agent communication, MCP will become table stakes for any platform serious about participating in the AI ecosystem.

[3] Model Context Protocol

WHAT THIS MEANS FOR YOU:

Prioritize adoption of MCP-enabled platforms like Confluent to ensure AI application portability and avoid LLM vendor lock-in. Ultimately, this will future-proof your AI strategies and standardize tool-calling integration across diverse LLM environments.

Recommended Reading

CONFLUENT BLOG
Powering AI Agents with Real-Time Data Using Anthropic's MCP and Confluent

THE NEW STACK ARTICLE
A2A, MCP, Kafka and Flink: The New Stack for AI Agents

THE NEW STACK ARTICLE
The Precision Engine: Why Agentic RAG Is GenAI's Next Leap

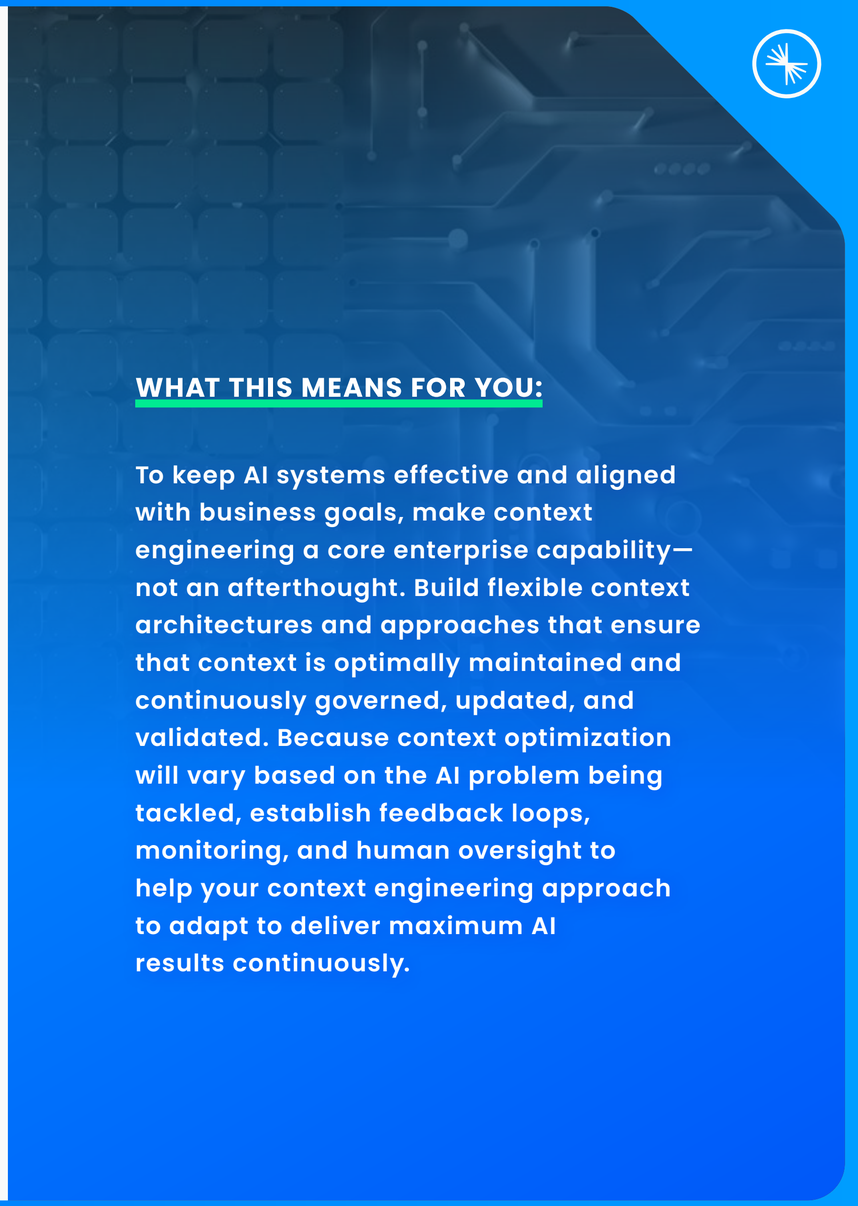
PREDICTION #3

Context Engineering Is the Next AI Unlock

AI in 2024 was about retrieval-augmented generation (RAG), 2025 was the year of agentic AI,[4] and 2026 will see the focus shift to context engineering. We have discovered that even if we can get access to all the right data and context to complete a task, there have been a number of challenges around managing all that needs to be overcome when working with LLMs.

For example, how does one fit everything required for a complex or long sequence of interactions in the limited context window? How do you ensure that you don't overload the LLM with context, slow down its processing time, and decrease the accuracy of its responses? As the amount of context grows, how do you ensure the most important items are not lost in the haystack?

Context engineering will be central to overcoming these problems and making AI work better in 2026. If software engineering is about writing pre-determined rules that are executed perfectly every time you run the code, when you build on pre-existing foundation models you guide the model with data and context. You must evaluate continuously, replay history to refine logic, and adjust live through prompts, rules, and context. Traditional software is about iterating on code while context engineering is about iterating on data.



WHAT THIS MEANS FOR YOU:

To keep AI systems effective and aligned with business goals, make context engineering a core enterprise capability—not an afterthought. Build flexible context architectures and approaches that ensure that context is optimally maintained and continuously governed, updated, and validated. Because context optimization will vary based on the AI problem being tackled, establish feedback loops, monitoring, and human oversight to help your context engineering approach to adapt to deliver maximum AI results continuously.

[4] 2025: The Year of AI Agents

Recommended Reading

DIGINOMICA ARTICLE
Confluent – 'AI Agents Are As Rocks' Without Real-Time Data Streaming

INDEPENDENT BLOG
The Rise of Context Engineering and the End of Static Software

CONFLUENT BLOG
Introducing Real-Time Context Engine: Simplified Context Engineering With Real-Time, Processed Data for AI

PREDICTION #4

AI Will Apply Increased Pressure to Existing Databases

As noted in our previous prediction, adoption of the MCP standard has gained significant momentum. MCP is proving to be an effective way to provide agents access to the data they need to deliver meaningful business outcomes. At first glance one might think that agents are just replacing certain human tasks and therefore will be making similar queries to those currently being run against enterprise databases.

In reality, agents will be far more demanding in their data requirements and will eventually be performing tasks at scales not practical for humans to attempt. What's concerning is that many enterprise operational systems of record are already struggling to meet the demands of modern business. Agentic AI is going to multiply data query volumes manyfold,[5] intensifying existing pressures. The ongoing efforts to offload queries to caches or supplemental databases is going to become even more critical. Now is the time to implement change data capture (CDC) pipelines and ensure data is flowing in near real time to systems capable of satisfying the insatiable data appetite we will see by the end of 2026.

[5] Why Enterprises Need AI Query Engines to Fuel Agentic AI

WHAT THIS MEANS FOR YOU:

Prioritize implementing CDC to stream operational changes from core systems into real-time serving layers. This will ensure that data remains fresh and readily available for agentic workloads with minimal impact to systems of record.

Recommended Reading

CONFLUENT BLOG
Data Products, Data Contracts, and Change Data Capture

USE CASE ARCHITECTURE
Building CDC Pipelines With Apache Flink®

PREDICTION #5

AI Will Drive Cyber Crime to Unprecedented Levels

As criminal organizations adopt generative AI for phishing, deepfake fraud, and automated malware creation, the economic consequences will be staggering. While today's projections estimate global cyber crime losses at roughly $12 trillion,[6] widespread use of AI by threat actors could push that figure as high as $18 trillion annually—a 71% increase[7] over conservative estimates.

The volume, variety, and sophistication of attacks are already rising rapidly, but 2026 will mark a major turning point.[8] By the end of the year, organizations may face nearly double the daily attack volume expected under normal growth trends. The average cost of a data breach will also surge as AI enables criminals to orchestrate coordinated, multi-vector attack campaigns at industrial scale, while dramatically reducing the technical expertise required to execute them.

Across industries, many technical leaders are already revealing that while developing an enterprise AI strategy is critical, mitigating AI-enabled cyberattacks is the most urgent priority.[9] In 2026, we will see large scale investments in defensive AI, real-time data infrastructure, and continuous threat detection to keep pace with the evolving cyber crime landscape.

[6] Cybercrime Statistics 2025: Cost, Threats & Trends

[7] Cybercrime To Cost The World $10.5 Trillion Annually By 2025

[8] 72% of cyber leaders say risks are rising. Here's how states and businesses are responding

[9] 90% of Large Organizations Unprepared for AI-Enabled Threats

WHAT THIS MEANS FOR YOU:

Security architects need to pivot from static detection to defensive AI capabilities powered by real-time stream processing to detect complex cyberattacks at industrial scale. They must urgently adopt a tiered data storage model to optimize costs. Technical leaders should allocate significantly more budget to building high-volume, low-latency data infrastructure for instant analysis and automated response.

Recommended Reading

SOC PRIME BLOG
AI Malware and LLM Abuse: The Next Wave of Cyber Threats

CONFLUENT BLOG
How Does Real-Time Streaming Prevent Fraud in Banking and Payments?

VENTUREBEAT ARTICLE
Software commands 40% of cybersecurity budgets as gen AI attacks execute in milliseconds

PREDICTION #6

AI Will Accelerate Enterprise Investment in Data Governance

In 2026, we predict a sharp rise in investment and focus on enterprise-wide data governance across all industries—not just those traditionally focused on tight regulation. In particular, cross-system data lineage will become a top priority as organizations seek to ensure that data feeding AI models is traceable and trustworthy.

For years, many industries have poured resources into improving data governance. In fact, 84% of technical leaders recently cited data management and governance as a top-tier technology priority.[10] Yet many organizations still struggle to answer key questions, like: **Where does this data come from? Can I trust it? What data was used to make this decision?** In highly regulated industries like financial services and healthcare, being able to answer these questions has always been critical. But even in these industries, executives still tell us that enterprise data governance is very hard to get right, particularly when it comes to data lineage and handling sensitive data in a way that makes data reuse easy.

Having governance capabilities built directly into modern data platforms has helped, but all organizations still operate complex systems. It's not uncommon to see scenarios where data flows from mainframes through message queues, SQL databases, APIs, and multiple integration tools before reaching downstream systems. With this level of complexity, tracing data accurately from source to consumption becomes incredibly difficult.

The rise of AI is now making this challenge impossible to ignore across all industries. For AI applications—especially those handling personally identifiable information, financial records, or other sensitive data—organizations must answer the same core questions about data trust, quality, and lineage, but with even higher stakes. Institutional knowledge or fragmented documentation of processes can no longer fill the gaps.

[10] 2025 Data Streaming Report

WHAT THIS MEANS FOR YOU:

Architects must prioritize implementing cross-system data lineage to provide an undeniable, auditable record of data used in AI decision-making, which is essential for regulatory compliance. Executives should mandate that all new data infrastructure investments include native, end-to-end stream governance to establish trusted data products for enterprise-wide AI consumption.

Recommended Reading

FORRESTER BLOG
Predictions 2026: The Race to Trust and Value

SILICONANGLE ARTICLE
Confluent Cloud Brings Real-Time Data to AI Agents

PREDICTION #7

Apache Iceberg™ Will Become the Standard for Cost-Effective Cold Data Management

Today, massive amounts of cold data are infrequently accessed. This data at rest will transition from being a liability to becoming a valuable asset for long-term trend analysis, auditing, security forensics, AI model training, and more. By the end of 2026, Apache Iceberg™ will solidify its position as the go-to open table format for efficient, long-term data retention.[11]

This shift is about building a tiered data strategy with the ability to support an "active archive" that balances cost-effectiveness with managed accessibility. Organizations will be able to retain more data and more effectively manage large volumes of historical data required by compliance standards—which have traditionally been left dormant in deep archives. As a result, teams will be able to more easily preserve and analyze high-volume data that was previously too expensive to keep or query.

Iceberg is poised to lead this transformation, with continued maturity in its Puffin metadata format,[12] advancements in data compaction and sorting, and emerging row-level lineage capabilities.[13] Compared to other open table formats, like Databricks Delta Lake (which is optimized for "hot" read performance), Iceberg's merge-on-read and partitioning design make it particularly well suited for cold data workloads. Finally, the industry adoption for Iceberg is massive; your data will be usable by all relevant platforms—including Databricks.

[11] How Apache Iceberg Won the Open Table Wars

[12] Puffin File Format

[13] Apache Iceberg v3: Unifying the Data Lakehouse Ecosystem

WHAT THIS MEANS FOR YOU:

Enterprise architects should immediately implement a tiered data strategy using Iceberg for the "active archive" layer, leveraging its native features to more effectively retain data affordably while ensuring it remains efficiently queryable for auditing and AI model training. High volume security and observability data should specifically be considered as an early target for data too expensive to keep in SIEM and observability tools.

Technical leaders should encourage Iceberg adoption across teams to transform historical data from a compliance cost into a long-term strategic asset, enabling future AI initiatives with deep historical context and simplifying regulatory compliance through cost-effective data retention.

Recommended Reading

DZONE ARTICLE
Building an AI/ML Data Lake With Apache Iceberg

DEV COMMUNITY BLOG
Dremio, Apache Iceberg and their role in AI-Ready Data

PREDICTION #8

Your AI Strategy Will Need an Independent Data Plane to Avoid Overcommitting

After a whirlwind rush to adopt generative AI, 2026 will be the year businesses realize they need to future-proof their usage. Thus far, the ability to switch between LLM vendors has been quite easy from a technical perspective. Major vendors like OpenAI, Anthropic, Gemini, and xAI have been focused on building differentiated models. While the competition will no doubt continue, the ability to easily switch or use multiple AI vendors represents lost revenue they want to protect.

That's why the battle between these vendors is rapidly shifting from differentiating the models themselves to bolstering the ecosystems around them. It's no longer just the major cloud providers; model makers are now aggressively building their own platforms, using compelling agent frameworks and trying to leverage the age-old trap of data gravity. Once an organization's enterprise context and real-time operational data are deeply embedded in a vendor's specific ecosystem, and agentic applications are built on their specific platforms, the resulting operational dependency makes the cost and complexity of migrating become monumental. In 2026, smart companies will deliberately build their AI strategy on an independent data plane and agentic framework to future-proof AI investments. This strategic decoupling will ensure organizations can switch AI partners without a costly divorce.

[14] Generative AI Market Report 2025

[15] AI Data Gravity: Why Model Training Is Moving Closer to Colocation Sites

WHAT THIS MEANS FOR YOU:

Business leaders must encourage a strategic separation between their core enterprise data and their chosen LLM platforms, requiring that data residency remains on a neutral, independent data plane. Architects should build AI strategy on a data platform that connects to any LLM ecosystem, ensuring portability and avoiding the operational dependency and lock-in trap of vendor-specific agent frameworks.

Recommended Reading

CONFLUENT BLOG
Data Streaming: The Key to Tackling Data Challenges for AI Success

CONFLUENT REPORT
The 2025 Data Streaming Report: Real-Time Data, Real Business Results

CONFLUENT BLOG
The AI Silo Problem: How Data Streaming Can Unify Enterprise AI Agents

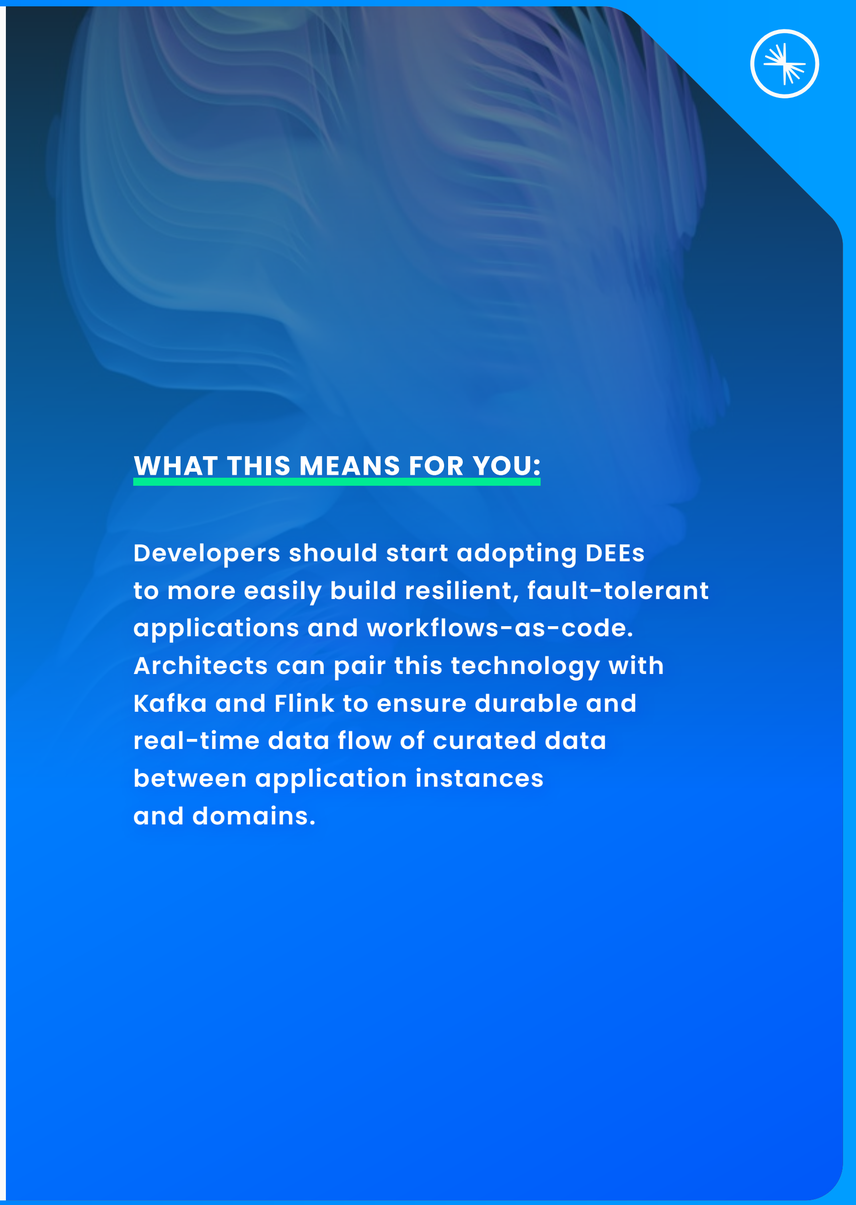
PREDICTION #9

Early Adopters of Durable Execution Engines Will Gain a Competitive AI Edge

Interest in durable execution engines (DEEs) has been slowly growing since Uber open sourced Cadence in 2017. In 2026, awareness and adoption of these fault-tolerant frameworks will increase significantly driven by agentic AI demands. Frameworks like LangGraph and Pydantic AI, for instance, have started investing in some aspects of durable execution. DEEs make event-driven architecture much easier and pair well with the established standards like Apache Kafka® and Apache Flink®. They turn reliability into a built-in primitive rather than something that has to be engineered into your code.

DEE platforms like Temporal and Restate persist workflow state, virtualizing execution across crashes and making retries, timers, and compensation first-class citizens. This can help shift the need for complex patterns, like Saga and CQRS, out of custom microservice code and into workflows-as-code. At the same time, Kafka will remain the scalable event backbone and Flink a high-throughput low latency processing and real-time analytics engine.

Aside from making event-driven applications much easier to build, this combination will prove valuable in the construction of AI agents where multisystem interaction, local state management, and multi-step processes are the norm. And as a result, early adopters will be able to bring durable scalable applications to market faster while ultimately enabling better inter-domain data reuse through universal data products built with real-time data streaming.



WHAT THIS MEANS FOR YOU:

Developers should start adopting DEEs to more easily build resilient, fault-tolerant applications and workflows-as-code. Architects can pair this technology with Kafka and Flink to ensure durable and real-time data flow of curated data between application instances and domains.

Recommended Reading

INDEPENDENT BLOG
The Rise of the Durable Execution Engine (Temporal, Restate) in an Event-Driven Architecture (Apache Kafka)

TEMPORAL BLOG
The Definitive Guide to Durable Execution

PREDICTION #10

Improvements in Generative AI Will Help Businesses Finally Address Legacy Tech Debt

In 2026, GenAI will help crack one of the most stubborn problems in enterprise technology: legacy system modernization. Since the beginning of enterprise software, modernizing legacy systems has been a difficult problem due to cost, lack of human skills, risk, and data gravity. The prevailing approach has been "leave and layer"—existing systems remain untouched while modern functionalities are added on top to minimize risk and shorten time to market. This approach has worked well for dealing with data gravity and reducing risk, but it has left organizations with a very expensive problem that has gotten worse with time.

Maintaining and understanding legacy systems has only gotten harder as the age of systems has increased, leaving modern organizations hostage to vendors, with little negotiating power. And in the meantime, valuable employees work on building new capabilities instead of refactoring hard-to-rewrite applications. Although GenAI is certainly not a push-button solution for these challenges (and LLMs benchmark poorly with parochial languages like Cobol), the viability of rewriting legacy code with GenAI is getting better at a rapid pace.

In the hands of system integrators who now specialize in solving this age-old problem with GenAI, we have turned a major corner in the cost-risk calculation. In 2025, for instance, across partner and customer use cases, we have seen a number of legacy JMS applications rebuilt into modern event-driven applications and moved into production while delivering significant business benefits. 2026 will be an inflection point for legacy application migration and things will only improve.

WHAT THIS MEANS FOR YOU:

Executives should begin to re-analyze the cost-risk for legacy system modernization with GenAI, leveraging specialized integrators for generative code translation and understanding. Developers should focus on migrating legacy messaging systems (like JMS) to modern event-driven architectures, using AI tools to accelerate the process and transform hard-to-maintain legacy code into new, valuable capabilities.

Recommended Reading

THE NEW STACK ARTICLE
The Precision Engine: Why Agentic RAG Is GenAI's Next Leap

CONFLUENT BLOG
How Developers Can Use Generative AI to Improve Data Quality

Looking Ahead to 2026

Although AI feels like a raging wild fire, in the broader data ecosystem we are seeing convergence around a key set of data platforms. Businesses have been moving fast to bring agentic AI to production in 2025, and 2026 will bring solutions to many of the barriers that made it difficult.

What we now need are modern platforms and architectures that build the base for streaming data that can intelligently support AI aspirations and provide the foundation for scalable, intelligent data flows that support advanced AI use cases. As CIOs look to grapple with all the new AI technologies, it will be more important than ever to contain tool sprawl. The cost to switch between AI vendors will inevitably decrease, and powerful new standards will help prevent vendor lock-in by crushing data gravity.

Platforms like Confluent, Databricks, and Snowflake—as well as offerings from hyperscalers—provide the built-in integration, processing, governance, and storage capabilities needed for AI to succeed at scale. The real opportunity lies in simplifying the stack so data, not tools, becomes the true driver of AI advantage.

Register for our Predictions Webinar

Sign up for a **FREE** Confluent Cloud trial

EBOOK

Turning Data Into Business Value in the Age of AI

EBOOK

Data Streaming Platform

The Key to an Evolutionary, High-Velocity Organization