conduktor

Five Unexpected Kafka Challenges at Scale

Introduction

In a digital economy that prioritizes the speed, scale, and quality of real-time data, Apache Kafka is an indispensable component of many application architectures. Organizations across many different industries rely on Kafka to power key business functions, from customer-facing applications to analytics to AI.

While Kafka usage varies by use case, it's since proven itself to be a versatile tool capable of meeting many needs. For instance, an online retailer might use Kafka to ingest event streams for a real-time recommendation engine, while a manufacturer may rely on Kafka to break down data silos by connecting both legacy and up-to-date technologies across a complex environment.

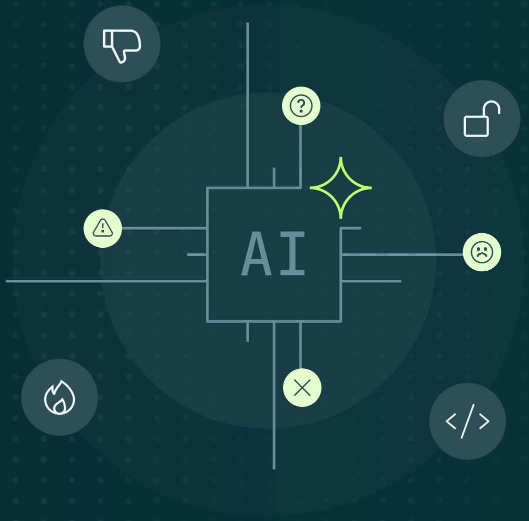
But for all its potential, operating Kafka, especially at enterprise scale, can be fraught with pitfalls. Through our talks with both current customers and prospects, Conduktor has uncovered five key challenges that architects, product managers, tech leads, and executives struggle with. Despite their different sectors (which range from logistics to finance), these leaders and practitioners encounter the same obstacles over and over again, including:

1. POOR QUALITY DATA, WHICH LEADS TO LOW-QUALITY OUTPUTS
2. A LACK OF SELF-SERVICE ABILITIES
3. THE DISCONNECT BETWEEN THE OPERATIONAL AND ANALYTICAL DATA LAYERS
4. THE GROWTH IN ZOMBIE INFRASTRUCTURE — AND THEIR ASSOCIATED COSTS
5. THE GAP BETWEEN LEGACY AND CLOUD TECHNOLOGIES

These challenges emphasize an important point: technology is only as good as the people and processes involved. Without the right practices and governance, teams can struggle with solutions like Kafka, failing to tap its full power and perhaps even leaving their organizations worse off than before.

CHALLENGE 1

Poor quality data



For organizations, every major initiative, including AI, analytics, compliance, and digital transformation, requires trustworthy, accurate data to execute. Unfortunately, many Kafka pipelines ingest poor quality data at scale, introducing inconsistent, misformatted, and missing data into downstream systems — with disastrous results. This produces low-quality outputs, reduces confidence among both internal and external users, impacts mission-critical applications, and can lead to outages or disruptions.

One example comes from a major European postal service, where "blurry borders" around team responsibilities for data quality, schema governance, and data retention left key duties unattended. Owners were responsible for ensuring that their data lived up to specific standards — but due to reasons such as a lack of accountability or ill-defined requirements, this task was neglected. As a result, downstream teams (like application developers) did not trust their data.

This organization was not alone. One technical lead at a logistics company mentioned that when producers misformatted data fields for ingestion, this would immediately break something on the consumer side — usually something powering a business-critical function. Without a way to block the intake of this data, developers and end users saw only broken applications and blamed Kafka.

Even organizations that utilized schema registries, the traditional solution for data quality, encountered major issues, specifically a lack of visibility and validation. Data scientists at one retailer were only able to identify issues after they appeared in applications and impacted KPIs, as they had no way to monitor messages within Kafka itself.

By implementing clear ownership, in-stream observability, and automatic enforcement, teams can stop low-quality data from entering applications. This will lead to:

Improved trust

High quality data creates high quality outputs, which is true for everything from AI to analytics. In turn, this increases trust among colleagues and customers alike, as they can use results and applications with confidence.

Reduce and reallocate time and money

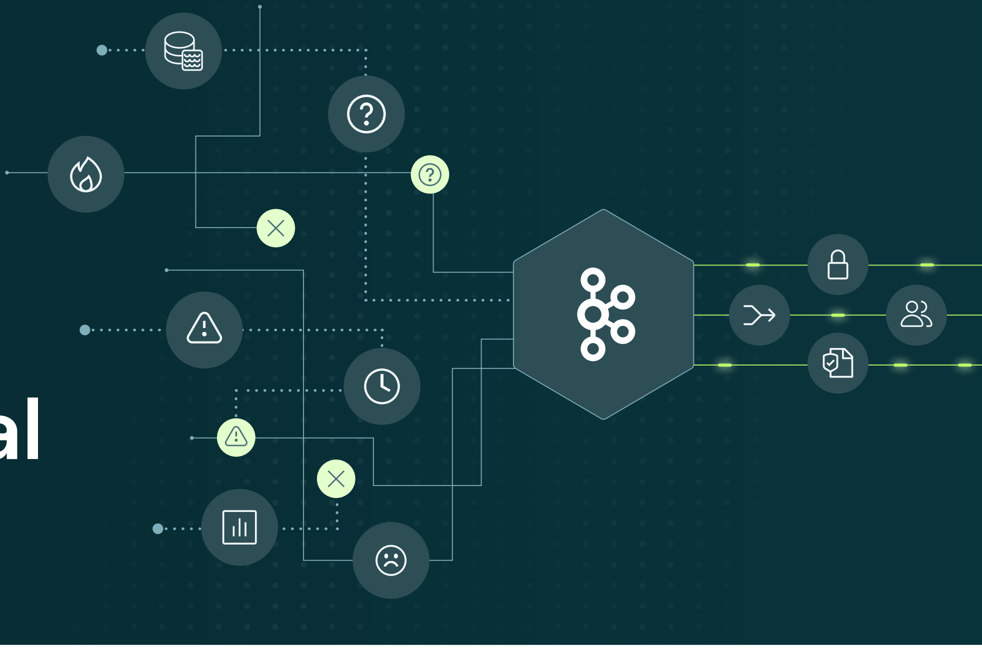
By preempting bad data, teams can also save hours spent on reactive approaches, such as a migration or a lift-and-shift. Employees can now pivot to other work, increasing productivity and innovation.

Meeting budgets, KPIs, and SLAs

Uptime and accuracy are two important SLAs and KPIs — both of which are threatened by low-quality data. By blocking this data from entering your environment in the first place, your teams can improve application reliability, meet key goals, adhere to SLAs (and avoid any associated financial penalties.

CHALLENGE 2

Slow, overly manual workflows



Developers, analysts, and product teams need data — now. If they have to wait, blocked by overly manual processes and unclear governance, then the business suffers.

Unfortunately, many organizations lack the systems, standards, and processes to help these internal users access data rapidly and safely. Too often, platform engineers, responsible for performance, stability, and security, build in guardrails to reduce risk — but without automation or self-service, these procedures can become bottlenecks. If developers and analysts wait too long, their productivity is impacted and inter-team tensions may build.

At the European postal service, the process of requesting operational data to be persisted for consumption by developers, data scientists, and analysts, required weeks. First, tickets were filed with the platform team, who would then create and configure connectors and S3 buckets. Finally, platform teams have to validate the permissions for connectors before internal users can finally consume data — an overly manual and frustrating process for all involved.

In contrast, one platform team at a multinational retailer built a one-click system to move data to end users for analysis. With this system, they were able to deploy 700 connectors and sync 2,000 topics across them.

By implementing automation, standardization, and self-service, an organization can help analysts and developers move faster while reducing the workload for platform teams. These benefits include:

Faster time-to-value and time-to-market

Whether it’s applications or analytics, helping analysts and developers move faster means more rapid results. Products and features can be created, tested, and released more quickly, and decision makers can access insights to stay ahead of the competition.

Scale operations without chaos

Growing data utilization is vital to the success of your business. Implementing self-service with guardrails ensures that data usage grows — without throwing operations into disarray.

Improved business agility

Reducing platform friction increases developer throughput — and lets your business move more quickly. Pivot faster to capitalize on market trends, competitor weaknesses, and changing economic conditions.

Reduce overhead and employee turnover

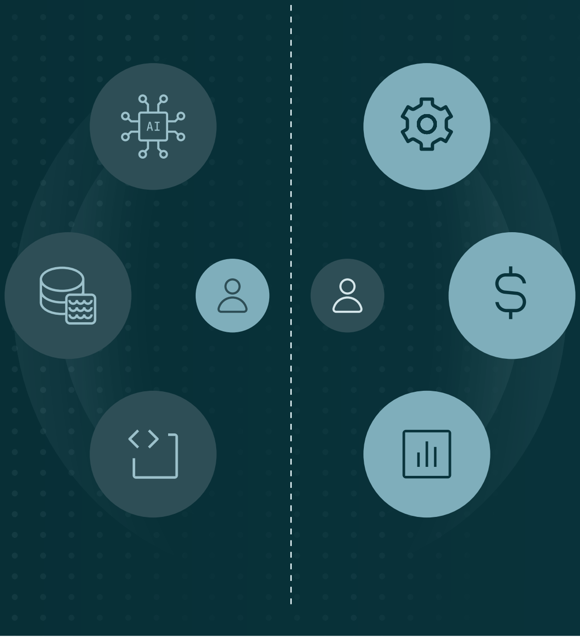
Stressed, overworked employees leave. By automating self-service with guardrails, organizations can help reduce administrative overhead, so that platform engineers can focus on more fulfilling tasks — and stay longer.

Clear ownership and procedures = more efficiency

Creating consistent, repeatable workflows saves time and money. Employees in different teams no longer have to waste time reinventing the wheel, and can instead focus on the high-value responsibilities they were hired to complete.

CHALLENGE 3

A disconnect between teams



To this day, organizations continue to struggle with silos, specifically the separation between operational and analytical systems — and the teams that run them. Operational data, such as credit card payments, delivery vehicle movements, or IoT sensor data, remains locked within technologies such as Kafka or application databases, while analytics teams use data lakes such as Snowflake, BigQuery, or Databricks.

To complicate matters, developers and analysts may not share the same governance, procedures, or even understanding of data — leading to a breakdown in trust. For instance, analysts may lack a way to explore and access streaming data, as well as knowledge of how it is ingested in the first place, leading them to blame developers for not creating better data pipelines with more consistent quality. Conversely, developers may resent analysts for constant requests for access, quality enforcement, or discoverability.

As a result, analysts can't find or get timely data, forcing them to work with stale or incomplete data and risk skewing their results. Without the right tools, analysts and data scientists may not even know about the missing data in the first place — an unknown unknown that can further exacerbate flaws in analytics, AI, and decision making.

To overcome these barriers, organizations need a common domain for different data systems and products — essentially a shared data language and structure that all teams and business units are familiar with and participate in. This also clearly defines owners for specific components (such as pipelines or topics) and controls (such as data quality or encryption).

At an organizational level, implementing shared technologies and bridging the breakdown between operational and analytical systems (and teams) brings key benefits.

Faster, better insights

The company that can use more of their data, whether it's for analytics, AI, or decision making, can generate high-volume, high-quality insights more rapidly than their competitors. This has positive implications for innovation, organizational agility, and revenue, enabling businesses to move faster and with more confidence.

Higher trust and cohesion

Removing the obstacles between operational and data teams can also improve morale and cooperation. By ensuring that teams buy into new systems and creating clear accountability, leaders can not only build more efficient workflows, but also remove old cycles of recrimination and blame.

Better governance for AI — and more accurate results

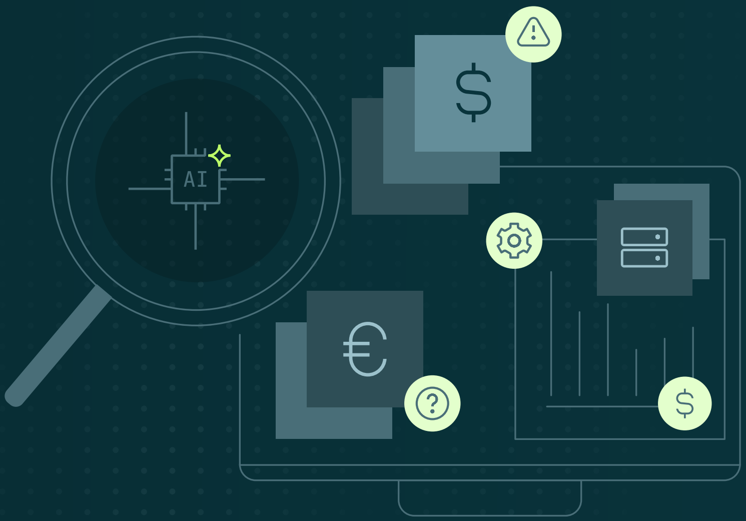
As mentioned earlier, AI is only as good as its inputs. By ensuring governance and access, teams can maintain the integrity, security, and relevance of their training data. This leads to more reliable, precise outputs and builds trust — among both internal and external users.

Decreased risk

Improving compliance and auditing, in the form of automatic policy enforcement, RBAC, and user logs, also reduces the possibility of violating key legislation such as HIPAA or GDPR. By embedding governance and policy management into operational data infrastructure, teams can move faster and more confidently, leading to more efficiency and productivity.

CHALLENGE 4

Mystery costs and zombie infrastructure



Kafka usage rarely grows in a logical, planned manner. Instead, it grows organically, with topics and schemas being created, partitioned, used, and abandoned, often without documentation or visibility. When migrations or re-factoring initiatives occur, these excess topics, schemas, and partitions might be forgotten entirely. This problem becomes even more entrenched if teams fail to establish clear ownership and naming standards at the time of creation.

While some topics, partitions, and schema may be revived, many will remain dormant — seeing minimal usage even as it continues to generate costs. These unused assets will often accumulate in legacy or non-production environments, where they can remain unnoticed by busy teams. Without any way to identify and remove zombie infrastructure, teams risk wasting valuable resources and driving up expenditures — without even knowing why.

Automation and observability can go a long way in resolving this issue. One team created a seven-day cleanup policy in their dev environment, which was neither production nor customer facing, deleting all unused assets after this time period. They also used improved monitoring to tag usage by environments, projects, and teams to ensure better visibility for financial planning.

It can also be helpful to standardize and implement policies across an environment. In order to keep expenses down, one organization limited self-service users to 10 partitions each, with any additional partitions triggering a request that had to be manually approved.

Ultimately, these solutions result in the same thing for organizations: realizing operational and financial efficiencies, including:

Decreased cloud and platform spend

Providing more detailed visibility into financial data enables leaders to better understand where they are wasting money, enabling them to operate more efficiently. In addition, teams can now improve budgeting and forecasting, allowing them to assess the impact of resources and spend — and allocate them more effectively.

Increased reliability and responsiveness for business-critical applications

Cutting away unused, excessive Kafka infrastructure also boosts performance, such as lower latency and fewer incidents arising from clutter or misconfigured assets. In turn, this has positive effects on customer experiences, as applications and products become faster, more reliable, and more responsive.

Lowered risk

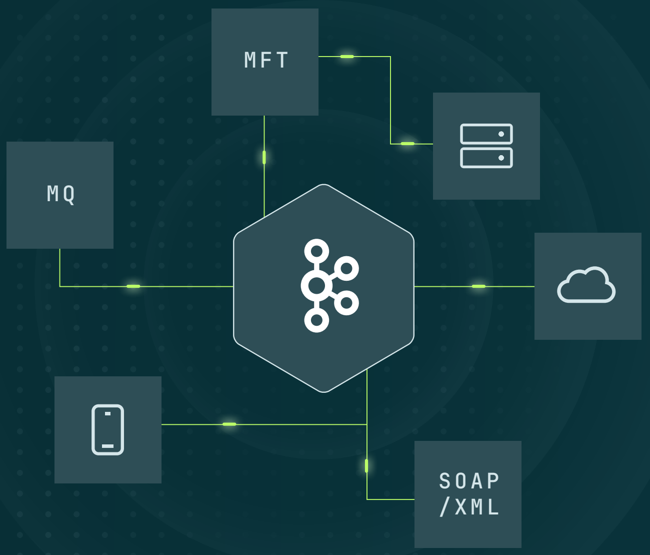
Zombie infrastructure may contain sensitive data and legacy configurations, which makes it doubly vulnerable to hacks and leaks. Removing this infrastructure (or even updating retention policies) can reduce organizational exposure and improve compliance postures.

Higher team velocity and shorter time-to-market

By streamlining Kafka architectures, developers deal with fewer irrelevant assets, enabling teams to onboard and work faster and more efficiently. This cuts down development cycles, helping organizations build, iterate, and release products more rapidly than their competitors.

CHALLENGE 5

Fragmented legacy and cloud technologies



Kafka isn't limited only to ingesting streaming data. In fact, it's moved beyond this niche, becoming a central nervous system that bridges the divide between legacy systems and modern microservices architectures.

These otherwise obsolete technologies, such as Managed Queues (MQ), Managed File Transfer (MFT), and SOAP/XML, remain within digital environments for a number of reasons. They may power mission-critical applications or complex workflows, or simply be hidden away from view. Whatever the case may be, these technologies often create fragile, undocumented dependencies, and removing them could lead to outages or worse.

Kafka is the solution, serving as a strategic integration layer with its rich ecosystem of source/sink connectors. Because it can stream data asynchronously, delivering events from one producer to multiple consumers, it frees up data environments from the limits of point-to-point transmission. This flexibility is ideal for mixed architectures, allowing interoperability without the need to create, configure, and maintain custom connectors or pipelines.

Even so, using Kafka as a strategic backbone does present challenges. Without the right systems and standards in place, teams may duplicate efforts, misuse data, or struggle to find and trust the streams they need. Additionally, the absence of clear ownership and access policies can lead to bottlenecks or accidental exposure of sensitive data.

Assuming that organizations can standardize, manage, and govern their Kafka infrastructure, they can fill in the chasm between obsolete, holdover technologies and newer ones. This brings high-level benefits, including:

Future proofing your environment

Oftentimes, these outdated systems hold back organizations from truly moving forward and becoming competitive. By using Kafka as the connective tissue, your teams can ensure that you can progress forward without abandoning vital, if older, infrastructure.

Migrate and upgrade without disruptions

Equally important, streaming data between systems via Kafka buys time for your teams, so that they can migrate on their terms, with minimal impact on business-critical or customer-facing applications and products.

Unify data across platforms and clouds — and remove vendor lock-in

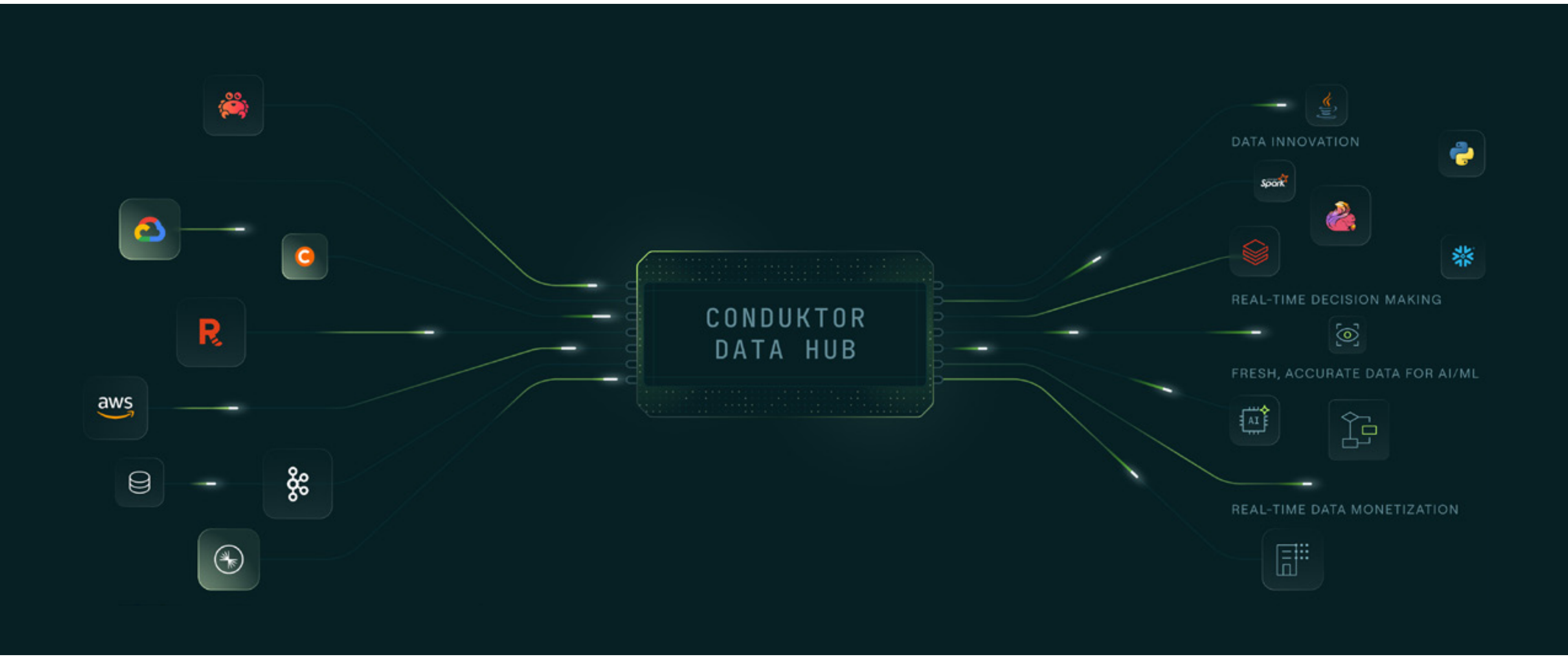
Sharing data via Kafka enables data to move seamlessly across clouds, platforms, technologies, and ecosystems. This removes data silos by cloud provider or technology, freeing your teams from vendor lock-in and allowing them to work with the most optimal solutions.

Increase efficiency and ROI via reuse and standardization

By using Kafka as the connective tissue between your microservices components, teams no longer have to build redundant pipelines by business unit or use case, instead consuming data as needed. This efficient use (and reuse) of resources improves ROI, centralizes governance, and speeds up both developer productivity and time-to-value for data.

Operationalizing Kafka at scale with Conduktor

Apache Kafka has become essential infrastructure for modern digital environments — and the organizations that run them. But while Kafka moves data exceptionally well, it doesn't include features for managing that data securely, efficiently, or at scale. Kafka also lacks native tools for governance, cost control, access management, and collaboration — making it risky and expensive for enterprises to utilize without additional control.


aws
CONDUKTOR
DATA HUB
DATA INNOVATION
Spark
REAL-TIME DECISION MAKING
FRESH, ACCURATE DATA FOR AI/ML
AI
REAL-TIME DATA MONETIZATION

Conduktor fills that gap. As a data hub built specifically for Kafka, Conduktor helps organizations transform Kafka from a technical enabler into a true business asset — making it secure, scalable, and aligned to the goals of the enterprise.

Conduktor provides in-stream observability and validation to improve data quality and prevent downstream issues. Teams can replace slow, ticket-driven workflows with standardized self-service abilities, so that developers and analysts can access data quickly (and without compromising governance). Conduktor also bridges the divide between operational and analytical systems, offering centralized catalogs and metadata tagging that make Kafka data easier to find, understand, and use across teams.

Conduktor also helps teams reduce waste and control costs. It identifies unused topics, excessive partitions, and long-retention policies that quietly accumulate in legacy environments — giving platform teams visibility and levers to clean up infrastructure without disrupting active pipelines. Finally, it enables Kafka to serve as a strategic integration layer across cloud and on-prem environments, replacing brittle point-to-point connections and reducing vendor lock-in through standardized, Kafka-native data sharing.

The business impact is significant. With Conduktor, organizations accelerate time-to-value, reduce platform and cloud costs, strengthen compliance, and unlock more value from their real-time data. Teams can move faster, stay secure, and scale Kafka without losing control.

If you're ready to bring clarity, control, and confidence to your Kafka environment, we'd love to show you how Conduktor can help.

Book a demo today