A Tech Executive's Guide to

The Data Streaming Ecosystem

Table of Contents

Click the page number on any page if you'd like to return to the Table of Contents.

Introduction

BEING DATA-DRIVEN is no longer enough—today's businesses have to be ready to react in real time. This reality has driven IT leaders to make investing in real-time capabilities a top priority in 2023.

In theory, being data-driven means making decisions based on facts and metrics. In other words: reliable evidence that reflects reality. In practice, however, "data-driven" organizations are often relying on hours- or even weeks-old data to inform their decision-making.

The longer it takes to ingest, operationalize, and use new data, the more that the value businesses can extract from that data diminishes. Seconds, or even milliseconds, of latency mean that customer-facing products and back-end operations are making decisions based on a view of data that may not match the current reality.

Without data infrastructure in place that enables real-time data ingestion and analytics, organizations are making critical decisions without having a true picture of their internal systems, the customer experience, or even their threat landscape.

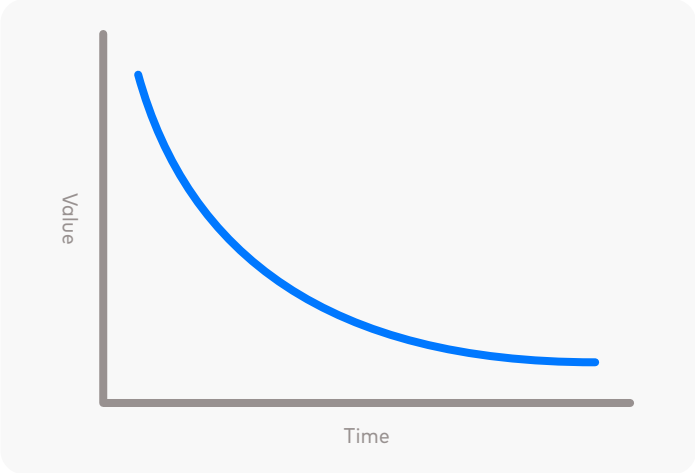


Figure 1 - What happens to data as it ages.

The companies that will rise above the rest are real-time capable—**readily learning, understanding, and reacting to what's really happening right now through data streaming**. And by putting in place event-driven systems built for scaling, these organizations will be well-equipped to predict and preempt what's next, too—using data pipelines to deliver the historical and real-time data needed to train and prompt artificial intelligence and machine learning (AI/ML) models.

Unlike batch-based systems that process stored data at periodic intervals, data streaming engines and platforms allow organizations to process continuous streams of data as soon as the data is generated or a specific event occurs. Over the last decade, this data-in-motion paradigm has been rapidly adopted across geographies, industries, and business models, upping the stakes for already competitive sectors.

But even among early adopters, few are taking advantage of the full potential of data streaming. In this ebook, we'll explore what tech executives need to know about data streaming to successfully implement and mature their real-time capabilities.

Welcome to the data streaming ecosystem.

Chapter 1

Competitive Enterprises Observe, Analyze, and Respond in Real Time

TO BE COMPETITIVE today, it's not enough to have a beloved brand or a promising product. Modern businesses need to be ready to respond and adapt to the latest information—from their customers, their partners, and their competitors. But building a data-driven business is an undertaking fraught with trade-offs and logistical dilemmas.

As organizations grow, adopt new technologies, and modernize systems over time, they inevitably accumulate an unsustainable web of infrastructure that makes real-time data integration incredibly challenging to implement.

With point-to-point connections between legacy data systems, applications, and mainframes, most technical teams are handed systems with immature governance and rigid engineering that lead to stale, siloed data—even in the most tech-savvy and innovative organizations.

Maintaining this status quo isn't an option for most tech leaders, who are charged with delivering the technical infrastructure to move the business forward.

Relying on outdated information can have disastrous implications for an organization's bottom line and adds risk to decision-making. Modernizing the data stack to streamline operations and breaking down data silos is paramount for the long-term health and growth of the business.

In the past, businesses that wanted to access new types of data and get the latest information as quickly as possible had one option: add to the problem by building even more batch pipelines or turn to messaging middleware like message queues (MQs), enterprise service buses (ESBs), and extract, transform, and load (ETL) tools.

Relying on these options has forced organizations to continue tolerating significant data latency, incompatibility between legacy and cloud-native systems, and costly maintenance, among other challenges and innovation roadblocks.

The Inevitable Rise of Apache Kafka®

Although data streaming isn't right for every scenario, there are a wealth of examples where even a second of data latency makes a significant difference in a business's performance—technologically and financially. That's why real-time data streaming has been adopted so readily across so many industries.

Over the last decade, more than 26,000 companies have decided that the solution they've been looking for is data streaming—more specifically Apache Kafka.

Kafka started as a way for LinkedIn to ingest and apply data in real time, but once open-sourced, it quickly became the foundation for use cases driving disruption in a number of industries. Now, over a decade later, there are over 1,000 Kafka use cases that practitioners and technical leaders have brought to life.

Among some of the most innovative engineering teams and companies, we've already seen data streaming become the default for mission-critical use cases. But even among early adopters, few are taking advantage of the full potential of data streaming.

To realize that potential, technical leaders need to:

- Define strategic business objectives for their team or department that map to a broader company strategy
- Determine when to use data streaming over batch processing to realize those business objectives
- Understand the wide range of technologies available within the data streaming ecosystem and the enterprise capabilities needed
- Give technical teams the tools to effectively evaluate data streaming solutions based on business and IT requirements
- Develop a strategic road map for prioritizing data streaming use cases
- Guide developer enablement and holistic implementation

The last point is perhaps the most important responsibility that tech executives have when it comes to data steaming adoption. The true value of this technology is realized when different teams can leverage the same data streams for unlimited use cases, creating a network effect that can't be achieved with open-source Kafka alone.

Although this technology is supplanting the traditional batch pipeline in many mission-critical use cases, data streaming implementations can and should offer much more.

When designed for enterprise use cases, a fully managed data streaming platform can help companies break down data silos that block innovation and efficiency, introduce new products, and accelerate time to value for strategic work.

Early Kafka Adopters

Exploring What's Possible, Then and Now

From FinServ and digital media to transportation and logistics, early adopters of Kafka have invested thousands of hours learning to develop, operate, and scale mission-critical use cases. Take a look at what these innovators were already achieving in the early days of data streaming:

- Goldman Sachs visualizes real-time transaction data
- Uber builds real-time transit infrastructure
- Audi reimagines transportation with real-time IoT data
- The New York Times stores all its content in Kafka
- Comcast monitors over 2 million miles of cable
- Pinterest builds a real-time ads platform with Kafka Streams
- Target simplifies omni-channel retail at scale

Many of these use cases still drive valuable customer experiences and business outcomes today, only now the barrier to entry can be much lower. Keep reading to learn how fully managed Kafka services like Confluent Cloud have made the benefits of data streaming more attainable and valuable than ever.

The Rise of Data Streaming

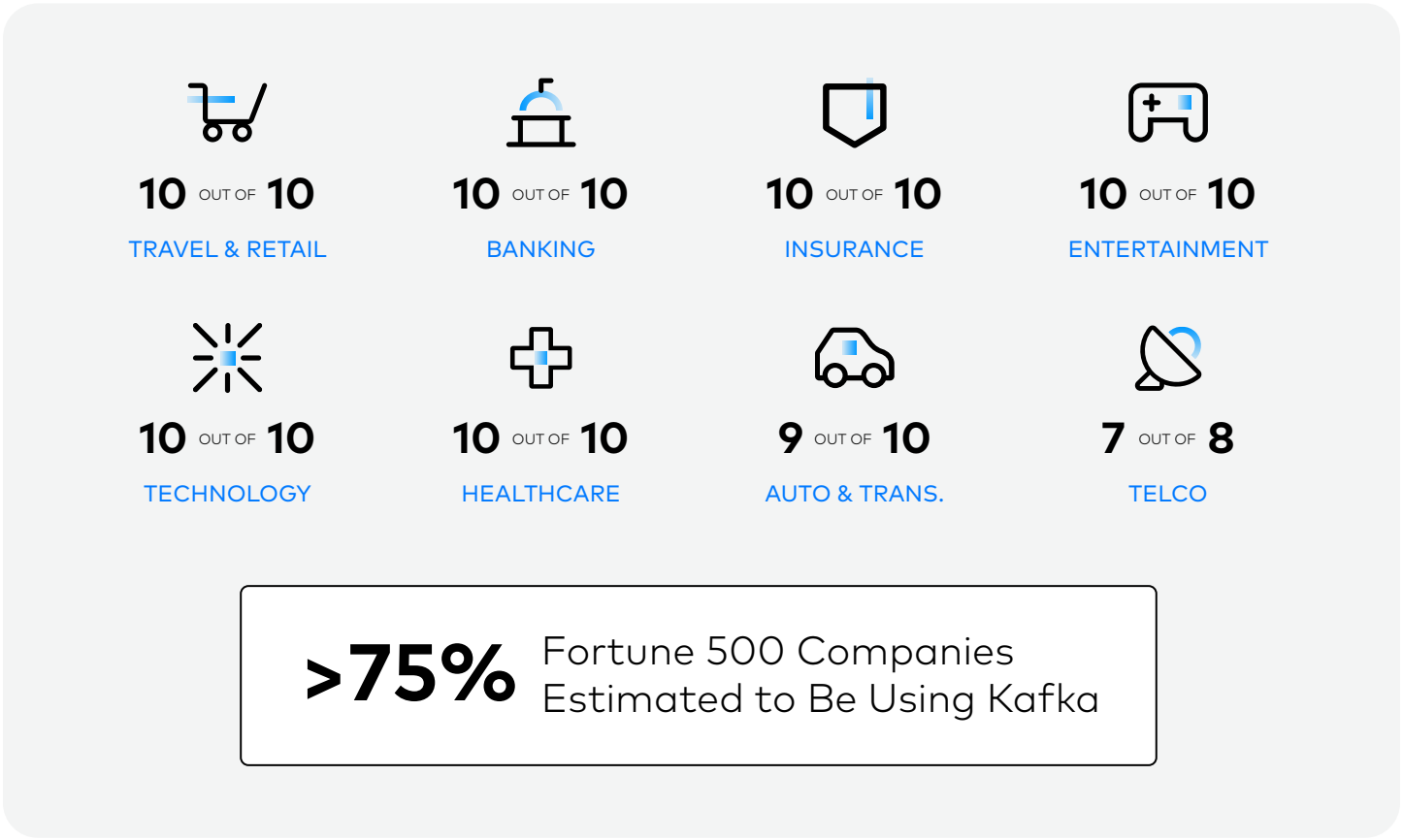


Figure 2 - Top company count by industry and >75% are calculated using the 2022 Fortune 500 list, released in August 2022, and are as of March 31, 2023.

Hear from Confluent customers on why they started using Apache Kafka

"Everything we do is in real time because batch processing is an old way of thinking. The longer your data waits, the less value it has. So, as data comes through, you need to be able to act on it, or enrich it quickly."

— Rajay Rai, Chief Information Officer, Trust Bank

"We chose event-driven architecture as the core of our platform, for which we needed a messaging service that gave us all the guarantees... not to mention that it had to be extremely scalable, highly available, and simple to use. Kafka hit all of these markers."

– Ravi Vankamamidi, Senior Director of Technology, Expedia Group

"Kafka is our central nervous system. It is a part of everything that we do. Most services across 110 different engineering teams with hundreds of services touch Kafka in some way, shape, or form in our company, so it really is mission-critical."

– Andrew Hartnett, VP of Software Engineering, New Relic

Chapter 2

The Business Impact of Data Streaming and Real-Time Processing

IN INDUSTRIES LIKE banking, e-commerce, media, digital advertising, and logistics, technical teams have invested significant resources into learning, experimenting with, and implementing Kafka use cases drive business value.

Even among organizations that are reliant on periodic, batch-based processing, technology leaders are choosing to use both—maintaining traditional data infrastructure while gradually migrating to Kafka to power the real-time analytics and operational use cases their businesses need to compete.

This kind of incremental modernization is especially valuable in highly regulated industries like financial services, healthcare, and the public sector where maintaining security and reliability is of the utmost importance. With this stepwise approach, organizations in these fields can integrate streaming technologies when and where it makes sense within their existing data stack.

That flexibility is critical because many data streaming use cases don't rely on the real-time storage and streaming capabilities of Kafka alone—they also take advantage of real-time data integration and stream processing tools like Kafka connectors, Kafka Streams, ksqlDB, or Apache Flink®.

Together, these tools extend the value of data streaming platforms, allowing organizations to ingest and transform data in flight to make it ready for use.

For example, when dealing with nonstop streams of transaction data or monitoring metrics of IT systems, real-time applications and services often need to combine multiple data points to turn raw financial or operational data into meaningful information that gives a true picture of what's happening with a specific customer account or IT system. Then, that enriched data can trigger the right response in downstream systems and apps.

The same is true for any customer-facing products that update existing information with real-time data (e.g., banking apps, real-time location tracking for transportation).

Organizations and business functions that don't leverage these tools to ingest, process, and operationalize data in real time will continue to act on outdated information as they wait for improvements to their existing data infrastructure.

Benefits from using real-time data streaming

Business efficiency

IT efficiency

Business responsiveness

Faster decision making

Enhanced customer experience

Want to learn more about the benefits and challenges organizations are seeing today as they advance in their data streaming journey?

Download the 2023 Data Streaming Report

Challenges When Scaling Kafka Usage

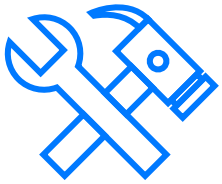
Some companies have had the internal expertise, timing, and capital to succeed with Kafka adoption, while others have struggled to realize its business benefits when it came time to scale.

Early Kafka adopters have had to invest just as much time and energy in infrastructure management as they have in hiring and retaining Kafka developers. Not all businesses that want to leverage data streaming have that option available to them.

Although the open-source streaming engine is purpose-built for storing real-time data at scale, the reality is that Kafka alone is not the complete, enterprise-ready platform businesses need. To maximize the value of real-time data and event-driven architectures, technical leaders need to put in place a data streaming platform that's ready to meet the demands of enterprise use cases and simplify operations across heterogeneous systems and environments.

Organizations adopting Kafka have two roads to choose from:

- Committing to a huge investment in building home-grown or partial streaming solutions while managing infrastructure operations and support as their data streaming initiatives scale.
- Relying on fully managed Kafka services, like Confluent, to successfully and cost-effectively mature their data streaming practice.



Is Self-Managed Kafka the Best Use of Your Team's Time?

"Although we had some experience running open-source Kafka ourselves, we recognized that we needed a trusted partner to help us resolve issues that we'd encounter with a larger-scale deployment. Confluent not only had a local office to provide that support, they also had the best known, verified solution, which was what we wanted for our mission-critical e-commerce platforms."

— Hudson Lee, Leader of Platform Engineering, eBay

"We decided Confluent was necessary when we realized the limitations of self-managing open source Kafka on our own— inability to scale cost-effectively, monitoring and security limitations, and no well-defined path to the cloud. Confluent offered a complete platform with all the enterprise-grade capabilities we needed to run mission-critical use cases, end to end."

— Olivier Jauze, CTO of Experiences Business Line, Michelin

"We only had a handful of people managing our entire infrastructure. We didn't want to overburden them with also managing the complexities of Kafka...Without Confluent managing our Kafka deployment, we'd easily need three or four additional employees to keep things ticking along. But now, we can put those resources into app development."

— Jaydeep Punjani, Principal Engineer, Mobile Premier League

Chapter 3

Enterprises Need More Than a Data Streaming Engine

KAFKA ON ITS own doesn't have all the capabilities that organizations need to make the most of real-time data. Although the data integration and stream processing capabilities that Kafka connectors and ksqlDB provide are valuable, these aren't the only capabilities that enterprises need to successfully scale data streaming use cases.

On top of allowing an organization to stream data in real time—to be truly complete and enterprise-ready—data streaming platforms must also deliver these key capabilities:

- **Connect** to data sources and sinks across environments
- **Process** data and store transformed data in real time
- **Stream** data with elastic scaling and resiliency
- **Govern** data at scale while promoting data discovery and data access
- **Share** metadata and replicate data across Kafka clusters and topics

Although open-source connectors and stream processing tools allow organizations to connect and process data streams, they lack the governance features needed to standardize and secure streaming data.

Open-source Kafka also lacks built-in tooling for data sharing, which means self-managed platforms often fall short when it comes to breaking down the data silos and bottlenecks that are plaguing most modern businesses.

Some organizations choose to write their own connectors and build out their own enterprise capabilities internally rather than taking advantage of pre-built connectors and managed Kafka services. This path comes with significant upfront and long-term costs, especially when operating in the cloud.

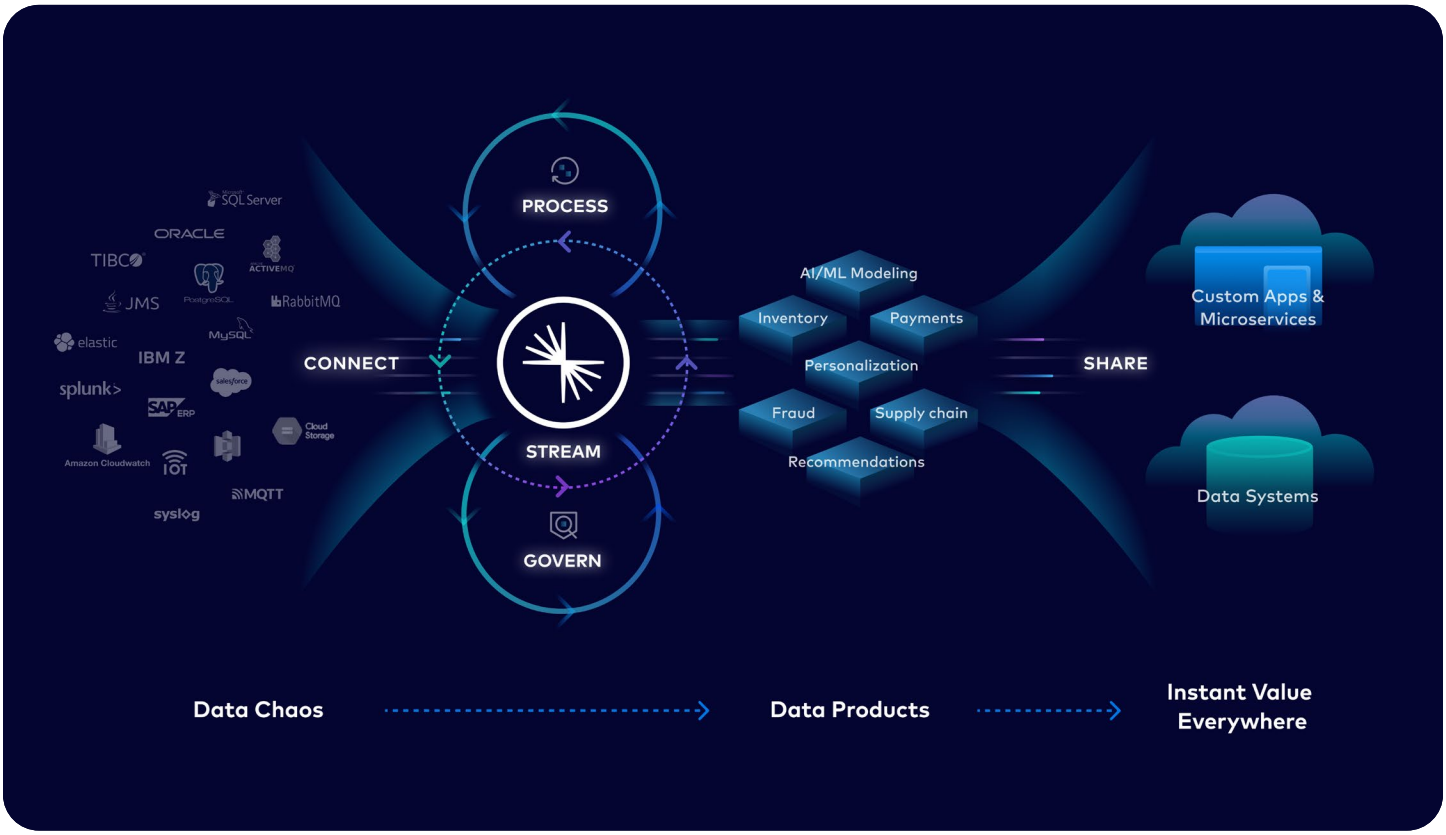


Figure 3 - The Five Key Capabilities of an Enterprise-Ready Data Streaming Platform

Anticipating Cloud Complications—Why a Fully Managed Kafka Service Is Key

While many organizations have had success managing open-source Kafka on-premises, the full value of its scalability and performance capabilities can only be realized in the cloud.

Migrating Kafka workloads to the cloud provides significant advantages for a number of key use cases, but migrations also exponentially increase the complexity and costs of deploying and managing Kafka at scale.

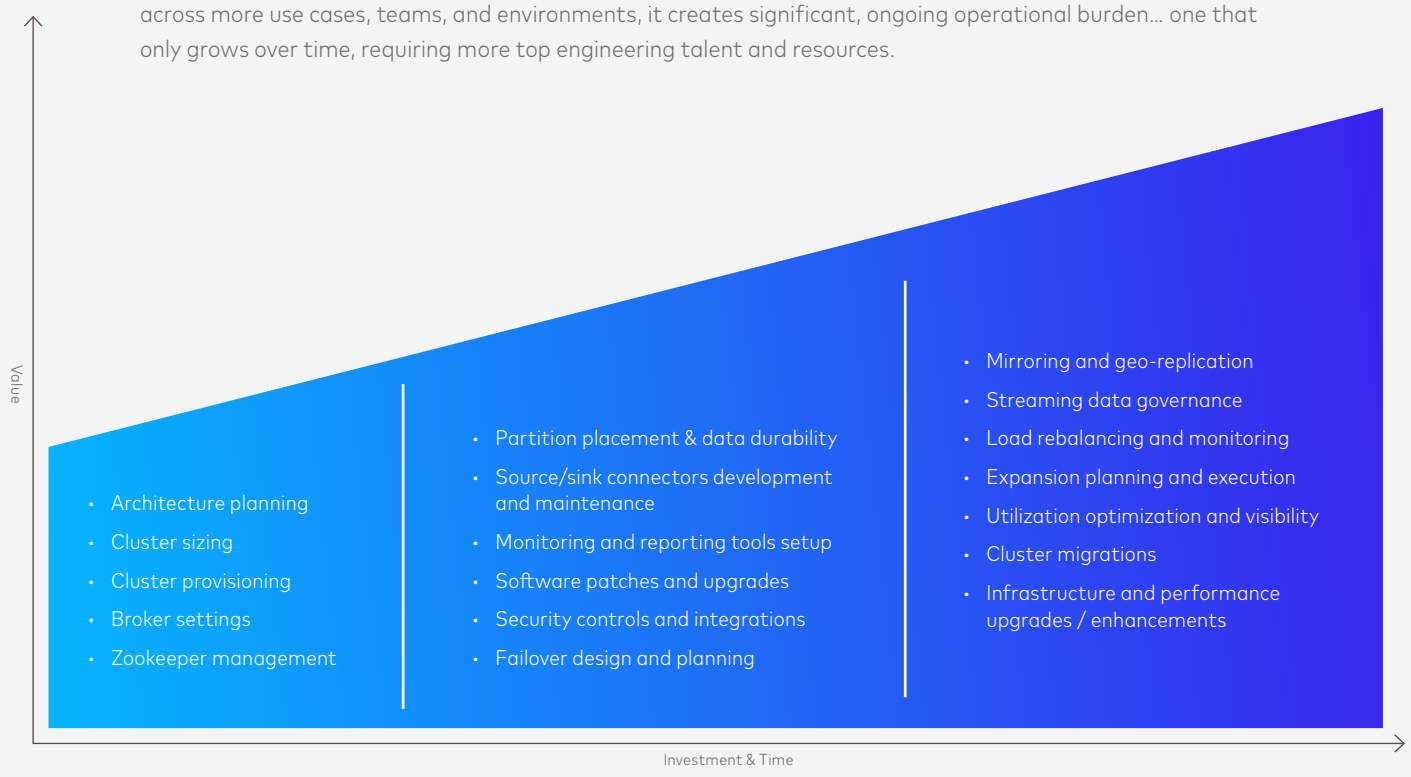
Being able to automate Kafka deployment, operation, and management is key to minimizing downtime and optimizing cloud costs. And building these capabilities can take, literally, millions of engineering hours. Additionally, few technical teams have the time and resources to tackle strategic, high-value work while providing the ongoing support and maintenance that a cloud-native platform requires.

The challenges involved have made it difficult for many organizations to rationalize the risks and costs associated with migrating their self-managed data streaming infrastructure to the cloud. Organizations who chose to self-manage Kafka have to weigh the opportunity costs of remaining on-premises against the risks and significant investment required for operating a hybrid or multicloud Kafka architecture.

As a result, many organizations that are invested in building future-ready capabilities with data streaming are evaluating the variety of managed Kafka services available. But not all Kafka services are created equal.

Self-Managed Kafka–Challenges Along the Adoption Journey

Operating Kafka on your own is hard enough during early experimentation. As your Kafka footprint expands across more use cases, teams, and environments, it creates significant, ongoing operational burden… one that only grows over time, requiring more top engineering talent and resources.



Chapter 4

Unlocking Value With Key Data Streaming Use Cases

WITH A COMPLETE, fully managed data streaming platform—one that allows organizations to store, connect, process, govern, and share streaming data without the operational burden of open-source Kafka—organizations can unlock a limitless number of use cases.

At its core, data streaming is an evolution of integration technology that allows organizations to respond and adapt to events as they happen. And two primary categories of IT workloads drive all the opportunities for innovation that data streaming has to offer:

- Streaming data pipelines
- Streaming applications

From real-time fraud detection and prevention to omnichannel experiences in retail, businesses can use a fully managed data streaming platform to build pipelines and applications, maximizing the value of real-time data.

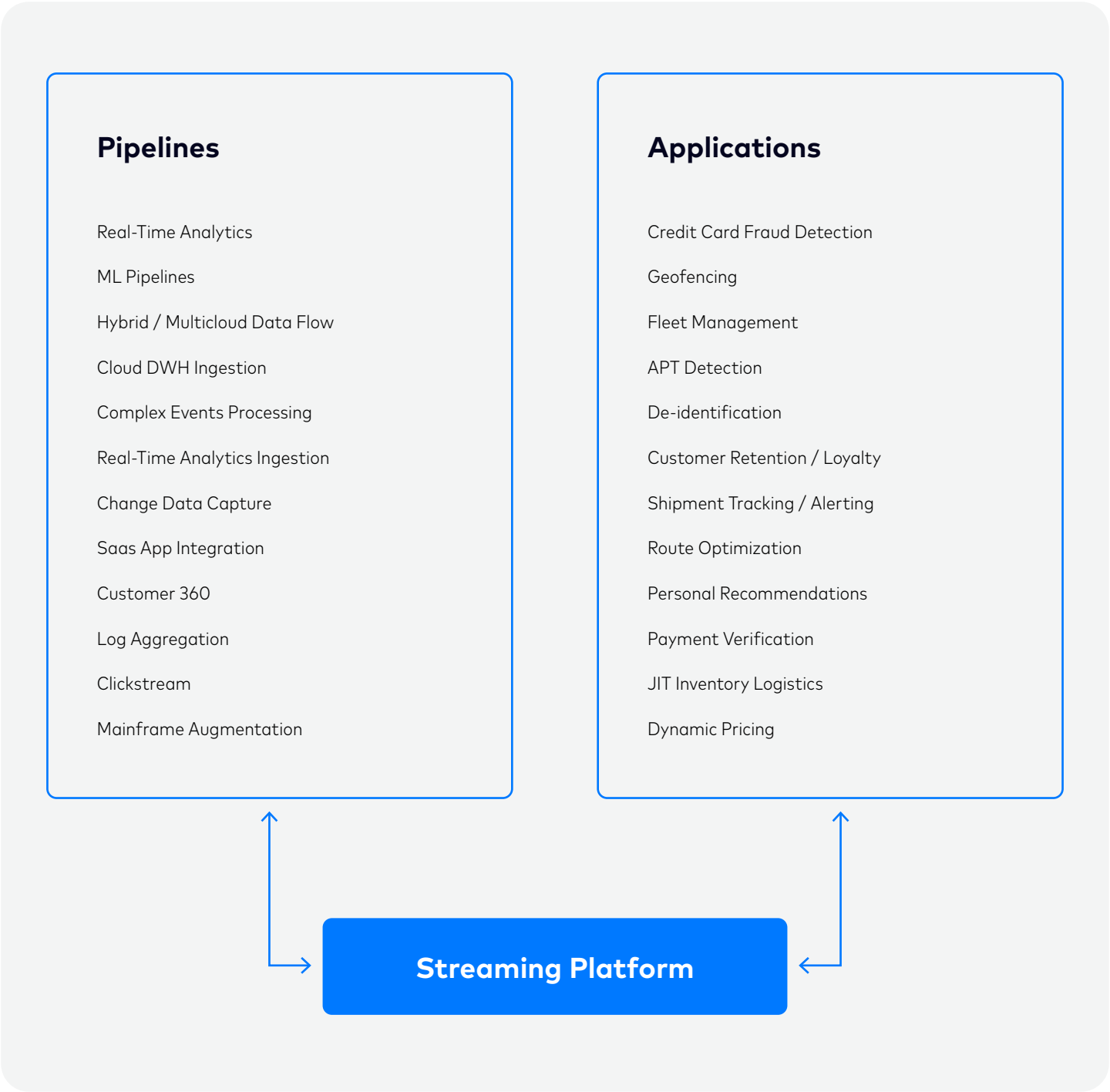
The typical "modern" data stack is built on a legacy paradigm—one that treats a chaotic and unscalable data foundation as an inevitability. Instead, organizations that choose to gradually implement streaming data pipelines across teams and functions can unlock an endless number of real-time solutions and streaming applications that solve previously insurmountable technical or business challenges.

From Data Infrastructure to Real-World Results

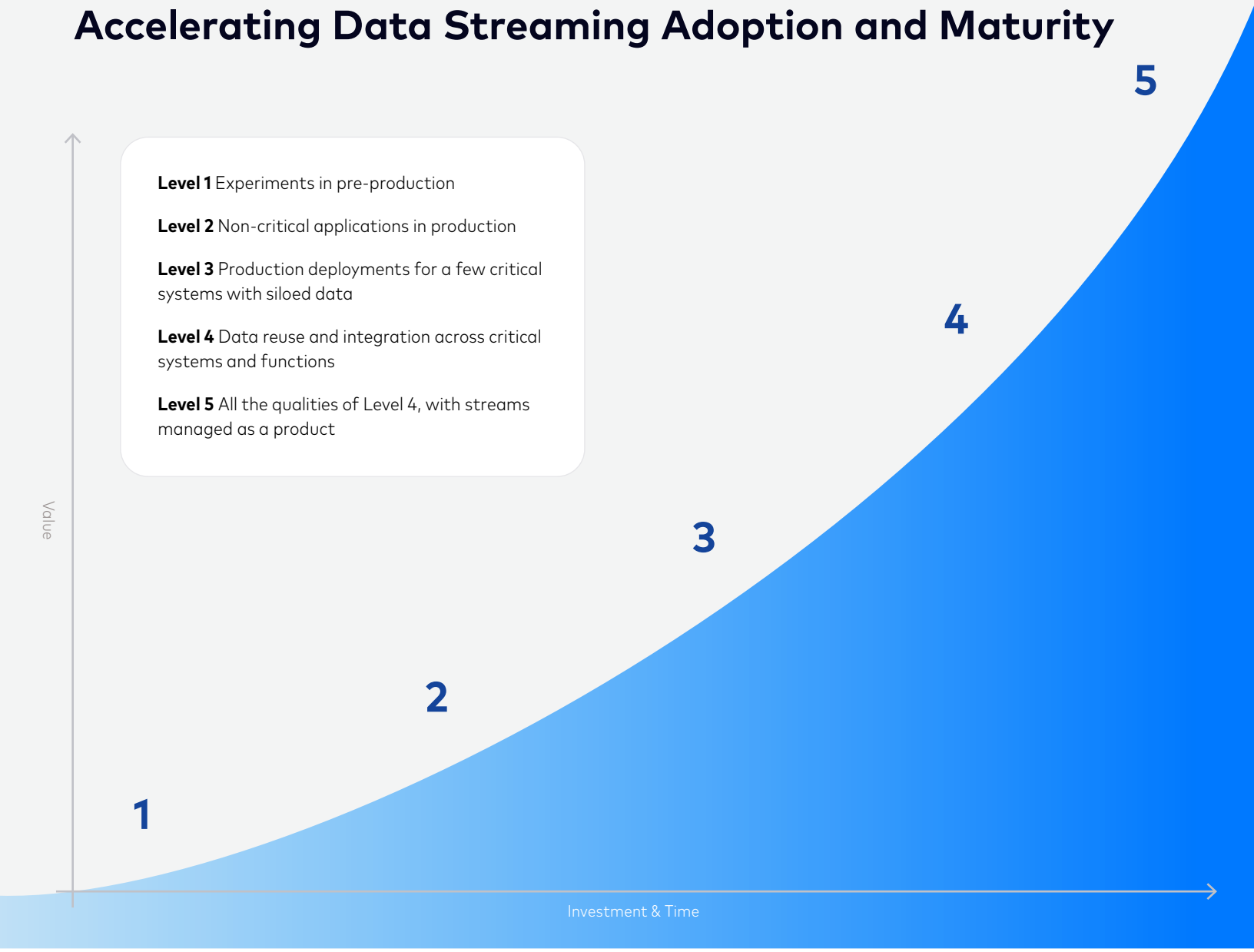
With streaming data pipelines and streaming applications as the IT foundation, organizations can use data streaming to develop countless, industry-specific solutions to problems in three key areas:

- Develop responsive and personalized customer experiences
  Instacart scales real-time grocery delivery
- Implement real-time back-end operations
  Michelin manages inventory in real time
- Build competitive products and services
  10x Banking builds digital core banking platform

Organizations can implement streaming data pipelines across teams and functions to unlock an endless number of real-time solutions and streaming applications.



Accelerating Data Streaming Adoption and Maturity



Level 1 Experiments in pre-production

Level 2 Non-critical applications in production

Level 3 Production deployments for a few critical systems with siloed data

Level 4 Data reuse and integration across critical systems and functions

Level 5 All the qualities of Level 4, with streams managed as a product

Figure 4 - The 5 Stages of the Data Streaming Maturity Curve

Across their adoption journey, organizations move through five key stages—from experimentation and siloed projects with data streaming to deployments in production across multiple critical systems.

As a company's use of a fully managed data streaming platform becomes more widespread across departments and business functions, data silos are broken down through democratized data ownership and self-service access. As a result, teams can more easily govern and maintain data quality while other business functions access, share, and process trusted sources of real-time data.

In their progression to Level 5 of data streaming adoption, organizations will have real-time streams in place, representing the vast majority of their business operations, customer activity, and transactions taking place. At this point, the data streaming platform can serve as the key integration layer or "central nervous system" across all of the company's systems and applications.

But taking a company from one stage to the next often isn't as simple as just onboarding and training new users. According to the 2023 Data Streaming Report, the majority of companies using data streaming today (63%) remain at Level 3 because of the operational and logistical complexities involved with moving from traditional data integration to an event-driven architecture.

Adopting a fully managing data streaming platform can help overcome many of these challenges. But for engineering orgs to turn their data streaming projects into a true enterprise-wide platform, they also need a vision to guide their progress and prioritize their efforts.

Tech leaders that want to help more teams adopt and leverage data streaming across the organization need a strategy centered around decentralized data ownership, self-service, and standardization. Putting in place a common data streaming platform can be the first step toward significantly lowering the barrier to entry for newly onboarded teams.

From there, leadership should provide guidance for using a fully managed platform to make incremental improvements toward data streaming maturity, offering resources like:

Assessments of the cost and returns of data streaming initiatives comparing self-managed platforms vs. managed Kafka services

An opinionated road map of which technology evaluations, proofs of concept (POCs), and production use cases to prioritize based on business needs

Education and reference material for developer enablement, including training, demos, and documentation that help increase productivity

Support from Kafka experts, whether hired in internal teams or accessed through partnerships, consultants, or managed service providers

Chapter 5

Accelerating Your Data Streaming Journey With Cloud-Native Kafka

WHEN EVALUATING MANAGED Kafka services, technical leaders need to look for platforms that are:

Cloud-native A fully managed service that's ready to instantly and elastically scale to meet demand to power streaming apps while alleviating operational burdens

Complete A platform that completes Kafka, providing the data integration, stream processing, security, data governance, and developer-friendly tooling enterprises need

Everywhere A cloud-native service that can span hybrid and multicloud architectures, allowing organizations to connect and unify all their data using only one platform.

Confluent Cloud is the only fully managed Kafka service that fulfills all three criteria, allowing businesses to realize the value of real-time data without taking on the costs of self-managing Kafka. At the same time, this cloud-native service allows organizations to operate 60%+ more efficiently and reduce their total cost of ownership (TCO) by over $2.57M compared to open-source Kafka.

We've applied millions of hours of Kafka expertise to build a data streaming platform that gives modern businesses the tools they need to liberate their data and unlock its full potential. Powered by the Kora engine, Confluent Cloud delivers the infinite scalability, elasticity, and resiliency that today's organizations need in their mission-critical use cases.

With Confluent Cloud, you can enable your teams to create better customer experiences and more efficient back-end operations. Confluent provides the data governance and integration capabilities needed to break down data silos and accelerate innovation—making it easier, faster, and more cost-effective to build innovative apps and products.

By replacing complex point-to-point batch pipelines and legacy messaging solutions with Confluent, businesses can empower all their teams to explore, understand, process, and trust data in motion.

Why Industry Leaders Choose to Stream Data With Confluent

"With Confluent Cloud, we can now provide operational data in real time to any team that needs it. This is really powerful and significantly reduces our operational burden."

— Siegfried Polysius, VP Cloud & Architecture, BestSecret

"Confluent plays an integral role in accelerating our journey to becoming a data-first and digital business... We use Confluent Cloud as an essential piece of our data infrastructure to unlock data and stream it in real time."

— Yves Caseau, Chief Information Officer, Michelin

"The scale and robustness of the system we built with Confluent Cloud have played a key role in accelerating our success in our mission of discovering new treatments and has helped us bring new treatments to human clinical trials."

— Ben Mabey, VP of Engineering, Recursion Pharmaceuticals

See How Confluent Delivers the Business Value of Data Streaming

FOUNDED BY THE original co-creators of Kafka, Confluent brings the industry's deepest expertise in data streaming to ensure your mission-critical use cases are implemented successfully. Today's organizations need a trusted partner that they can rely on to realize the business value of data streaming without having to take on the operational burden of self-managed Kafka.

Is a fully managed data streaming platform the right choice?

See how Confluent beats the cost of managing open-source Kafka.

Keep learning more about the business benefits of data streaming.