The State of Streaming Data

2023-24

Independent analysis of the size, shape and scope of modern streaming data deployments.

Commissioned by

Redpanda

Contents

Introduction

Streaming data refers to an endless current of real-time information generated continuously from countless sources. As cutting-edge technologies expand, torrents of data pour in simultaneously in small chunks from websites, apps, IoT devices, and more. This firehose includes everything from online customer behavior, social media trends and financial trades to location data and metrics from instruments and sensors. Streaming data can have both transactional and analytical applications.

For example, astute retailers use streaming data to optimize transactional customer experiences, supply chains and marketing campaigns. By monitoring inventory levels, browsing patterns and promotion response times in real-time, businesses tailor offerings, adapt pricing and fine-tune recommendations. Meanwhile, streaming analytics empowers airline crews to provide predictive aircraft maintenance, reducing delays. It allows utility providers to dynamically balance loads, integrate renewable energy and strengthen grid resilience.

We can all agree that real-time streaming data is important, to the point of constituting the new normal. But what is the true current state of streaming data in terms of the drivers and challenges for adoption, its size, shape, and environment?

This report summarizes findings of a survey conducted among both current users (59%) and prospective users (41%) of streaming data technology. A total of 300 survey respondents with an understanding for streaming data participated in the study.

The survey was carried out by independent research firm Material and commissioned by Redpanda Data. The complete survey methodology and an overview of the survey audience are at the end of this report.

Executive Summary

Some of the main findings from current and prospective users of steaming data systems...

1. **AI/ML is a primary trend driving adoption and expansion of streaming data usage.**
Delivering real-time Artificial Intelligence (AI) and Machine Learning (ML) to business users is likely to be the biggest driver of growth in the streaming data segment in the next 12-24 months. Roughly three-in-four respondents identified AI/ML as far and away the top trend fueling future adoption of streaming data.

2. **Real-time analytics is the clearest current or near-term use case for streaming data.**
Two-thirds to three-quarters of survey participants are using or expecting to use streaming data to support real-time analytics. Real-time analytics is also a leading driver of new adoption among prospective users.

3. **Data security and the lack of necessary in-house skills are key concerns.**
Two-thirds of respondents identified lack of in-house skills as a core issue, especially among prospective users. Over half of the technical challenges identified pertain to data security, privacy or governance.

4. **Users and prospective users alike see value and business applicability from the use of streaming data.**
On average, survey respondents mentioned five to six motivators for adopting streaming data. They expect to use streaming data to support a range of business goals and use cases, now and in the near future. Between two-thirds and three-quarters expect increases to their date volumes and streaming infrastructure – consistent across analytical and transactional workload types.

5. **The streaming data category is a multi-platform space with wide adoption of both Kafka-compatible and Kafka non-compatible platforms.**
Current users and prospective users work with (or expect to work with) on average between two to three streaming data platforms. The majority of current users (54%) and prospective users (75%) use or expect to use both Kafka-compatible and Kafka non-compatible platforms.

SECTION 01

Meet Current Users and Prospective Users

Let's start with a look at the profiles that emerged from surveying our two primary audiences.

Profile of a current user

Having already implemented a streaming data solution or presently doing so, current users see a large number of benefits accruing from its use — now and in the future.

Today they're using it to support a variety of business goals and use cases, including those more strategic (e.g., phasing out legacy technology, enabling growth into new markets). The majority are running both analytical and transactional workloads, and they're using multiple solutions including Kafka and non-Kafka platforms.

In the near-term future, they expect meaningful increases in the size of their streaming data infrastructure and the amount of real-time data they stream, especially as the development of their AI/ML systems boosts the need for streaming data services. However, their expansion is inhibited by a lack of in-house mastery and security concerns.

Profile of a prospective user

Implementing a streaming data solution is on their near-term radar, and they already have a list of platforms in their consideration set. They understand the tactical use cases for streaming data, the fundamental benefits it can bring to their businesses, and how it can support their AI/ML initiatives. However, they're not aware of its full potential.

Their lack of in-house expertise in this category has primarily hindered their adoption of streaming data technology. Concerns about solution complexity, integration with existing systems, and the time/effort required to set up the system and train users have further compounded these barriers.

Prospective users have greater diversity than users in their role type. They were more likely to select another role in conjunction with one of the four that were required to qualify for this survey (IT, Software or Data Engineering, Infrastructure Operations, or Application Development).

This signals that prospective users have a greater range of areas to oversee and are less likely to be exclusively focused on these responsibilities. Our learnings about their concerns about in-house skill gaps as a barrier to adoption further support this linkage.

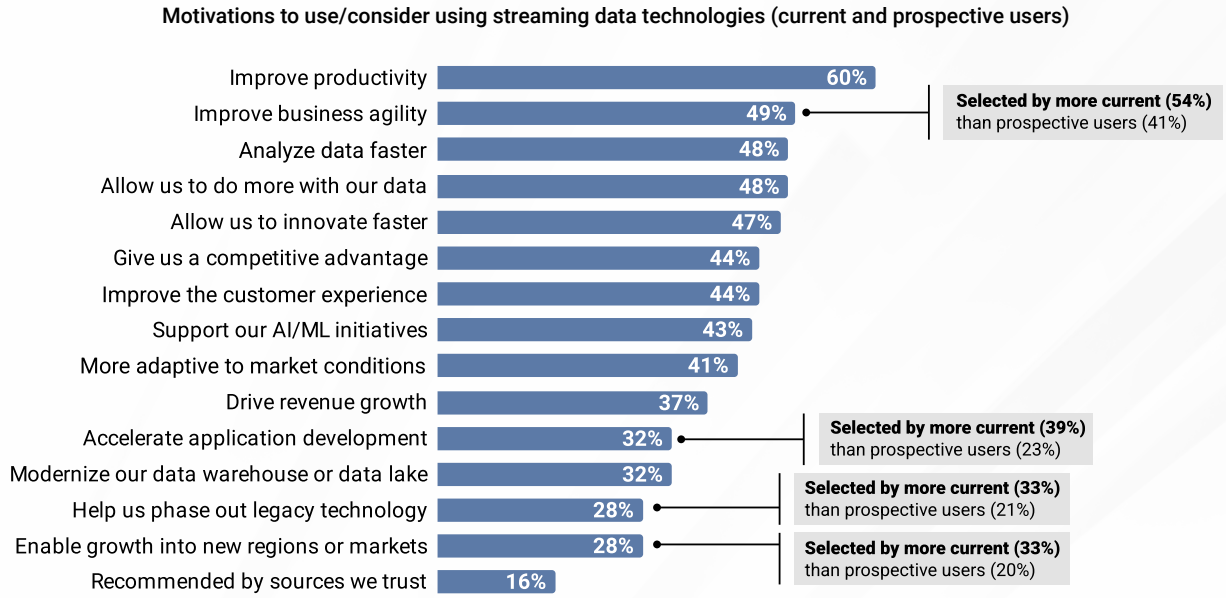
SECTION 02

Drivers for Streaming Data Adoption

This section aims to catalog and quantify the use cases driving adoption of streaming data systems, now and in the near future. For this, we surveyed both current users and prospective users of streaming data. We analyzed perceived benefits, projected growth and popular use cases.

The key finding is that users and prospective users alike see a large number of benefits, value, and business applicability from the use of streaming data. Real-time analytics is the clearest current or near-term use case for streaming data, and AI/ML is a primary trend driving adoption and continued expansion of streaming data usage.

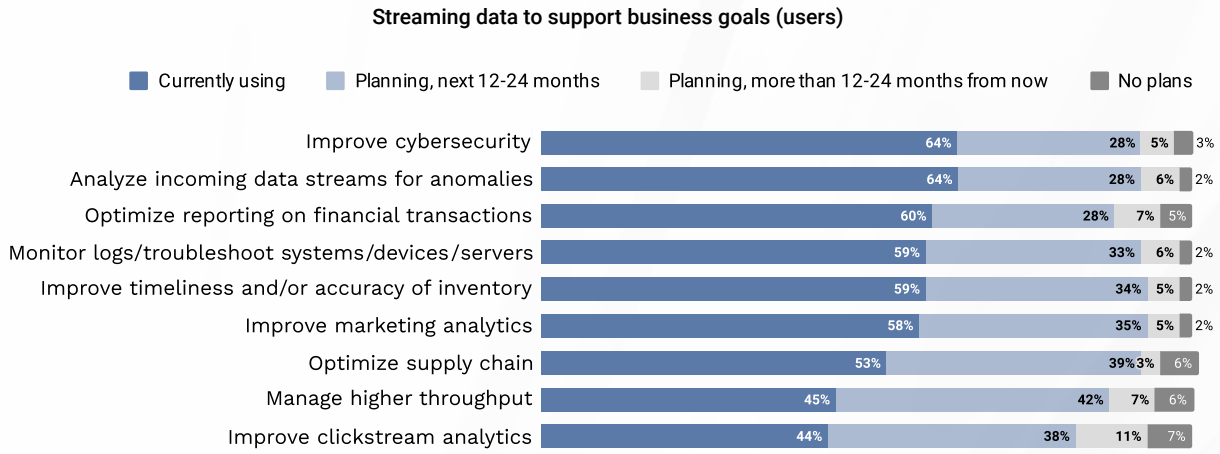
Benefits of streaming data



Both users and prospective users recognized a broad range of business benefits and applicability in this technology. At 60%, improving productivity was the most frequently selected benefit of streaming data. Use cases such as predictive maintenance and real-time analytics typically drive greater efficiencies and productivity. Event-driven systems can drive operational agility, automate manual processes, and help get products to market faster. Gains in productivity can also be had by replacing batch-oriented data processing workflows with more automated streaming-based ones.

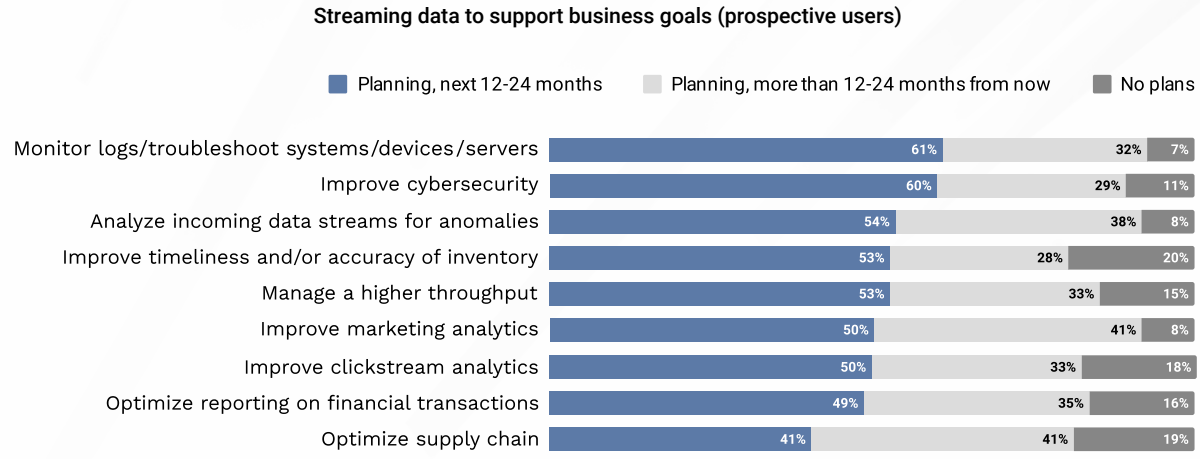
The survey found that current users in particular are motivated by agility and innovation. This reflects the intentions of businesses that have moved to real-time decision making in order to respond more quickly to customers and market conditions.

Business goals: Current users



Close to two thirds of current users are leveraging streaming data to improve cybersecurity and analyze incoming data streams for anomalies. Other business goals cited by the majority of current users include optimizing reporting on financial transactions, monitoring logs and troubleshooting systems/devices/servers, improving the timeliness/accuracy of inventory, improving marketing analytics and optimizing supply chain. All of these use cases are projected to grow further still in the near future.

Business goals: Prospective users

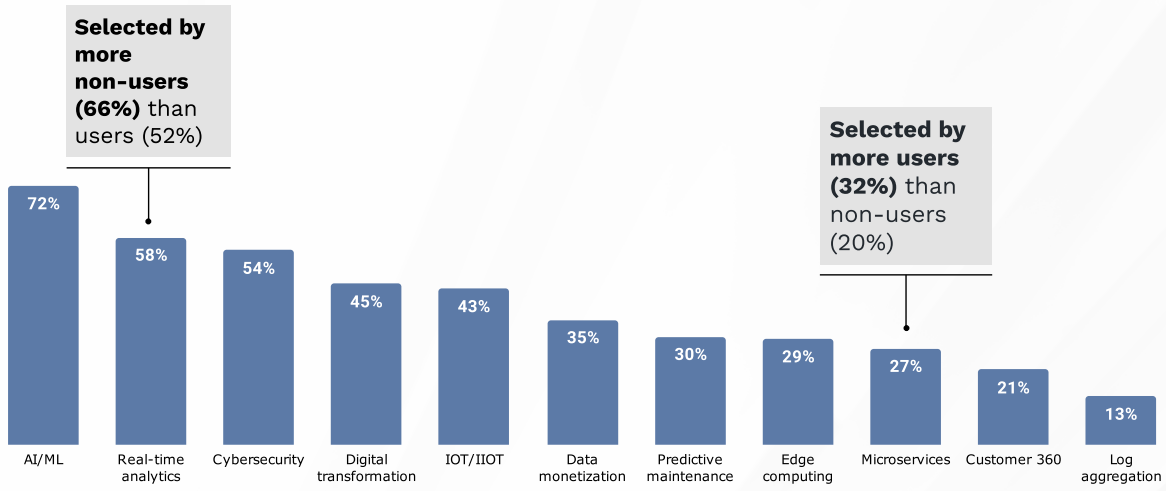


Similar to the results for current users, prospective users also expect streaming data will support a range of business goals. Troubleshooting systems and improving cybersecurity top the list for prospective users.

What's driving adoption?

AI/ML and real-time analytics are key drivers for adoption of streaming data. Both current and prospective users ranked AI/ML at the top of their list when citing use cases for the next 12-24 months.

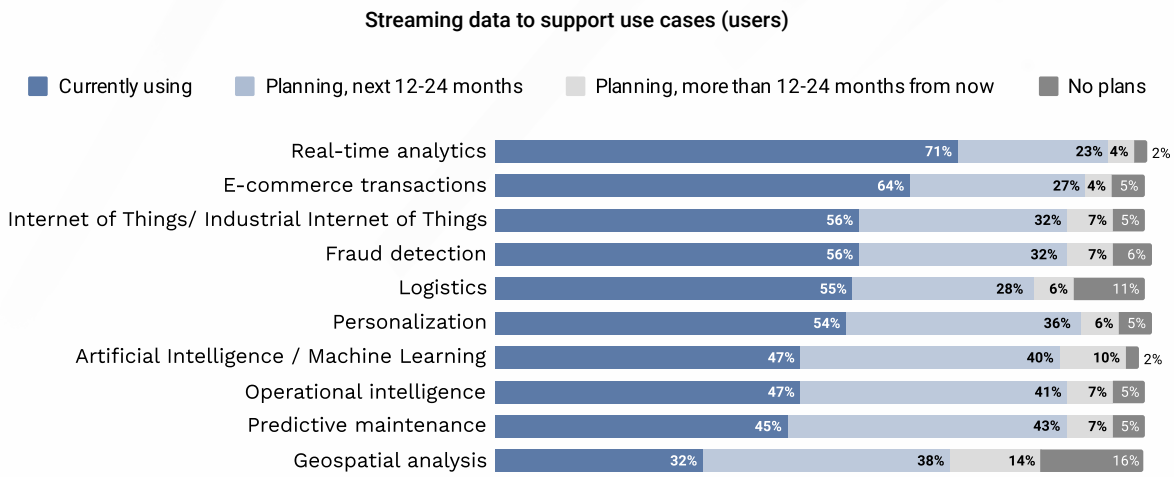
Expected trends driving streaming data adoption in next 12-24 months (current and prospective users)



Perhaps it's not surprising that the move to real-time decision making is more popular with prospective users, who would benefit from initially migrating to more responsive business operations. Current users were more likely to select microservices as a driver for streaming data, suggesting that they have moved into greenfield event-driven application use cases and not just traditional streaming data pipelines for analytics.

Use cases: Current users

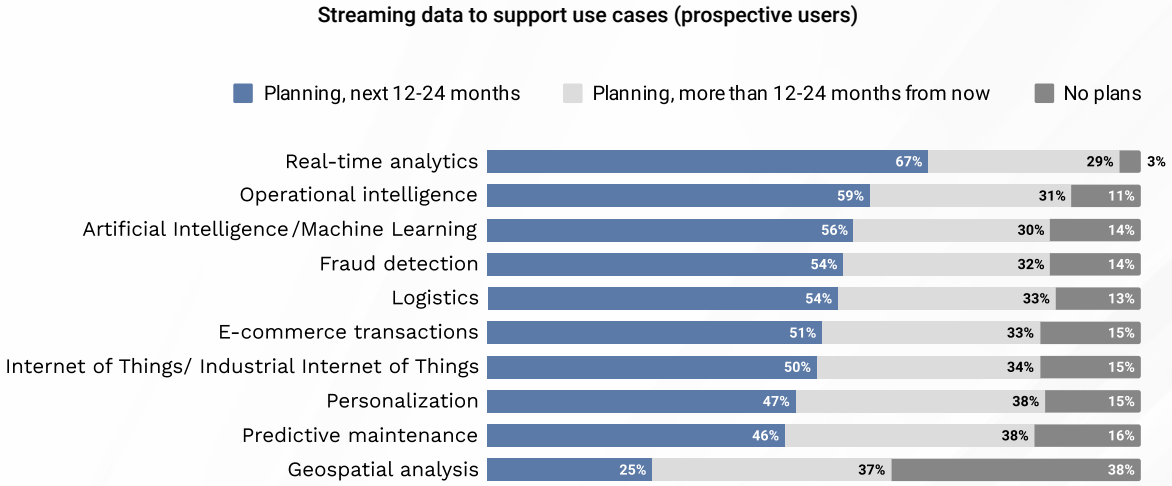
Streaming data to support use cases (users)



For users, real-time analytics and e-commerce transactions lead among the many use cases for which they're relying on streaming data. Looking ahead, use cases pertaining to AI/ML, operational intelligence, and predictive maintenance show meaningful expected growth over the next 12-24 months.

About 20% of existing users mentioned other current use cases. Some examples include data breach security, customer service feedback analysis, intelligent meter leak detection, fraud detection and monitoring online behavioral patterns. About 30% of existing users also mentioned other future use cases, such as fraud detection, IoT sync data, and customer experience reviews.

Use cases: Prospective users



Among prospective users, real-time analytics is also far and away the clearest use case. E-commerce transactions (another clear leader among current users) do not resonate as strongly with prospective users, but are still pertinent.

34% of prospective users also mentioned other future use cases, such as synchronized warehousing, fraud detection and custom content delivery. These are a mix of established use cases and domain-specific ones.

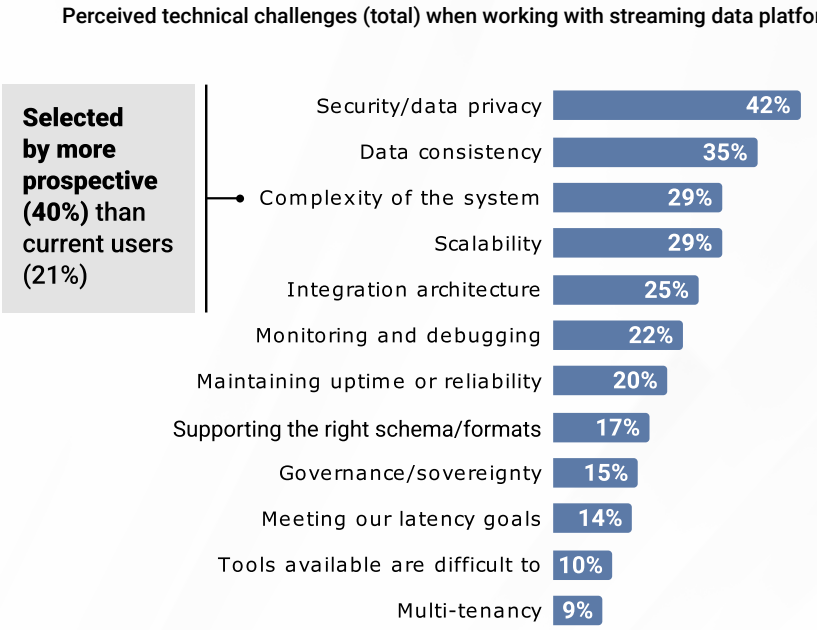
It's worth noting that AI/ML use cases rank very high in the lists of both users and non-users.

SECTION 03

The Challenges of Streaming Data

In this section, we catalog and quantify the key barriers to working with streaming data platforms, as identified by both current users and prospective users of streaming data systems. For better clarity, the survey separated possible barriers into technical and business challenges.

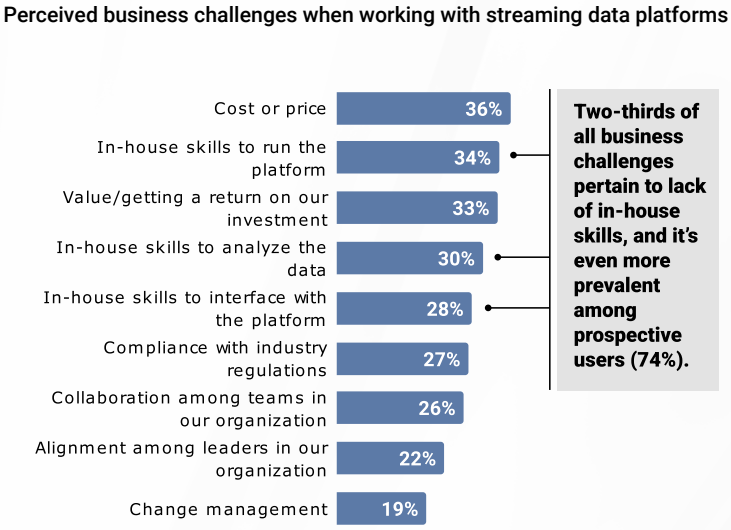
Technical challenges: Current and prospective users



Security and governance are top of mind for current users and prospective users alike. The prevalence of security concerns could suggest the audience does not yet have solutions in place to apply proper governance and security to their new or coming streaming systems.

Considerations about complexity of the solution and its ability to integrate with existing systems are also key concerns, especially for prospective users, who are twice as likely as current users to mention system complexity as a potential technical challenge.

Business challenges: Current and prospective users



Rather predictably, cost and ROI are the most frequently cited business challenges. In today's economic climate, these would probably be listed as concerns for any new technology.

Lack of in-house expertise is universally a key challenge when working with streaming data platforms. Two-thirds of our sample cited concerns about not having the required skills in-house to run/interface with the platform or analyze data from it. This is especially a concern among prospective users, three quarters of whom selected this issue.

Related to lack of in-house expertise, prospective users worry about the time and effort required to set up and train their users on the system. More than half of prospective users indicated these as primary reasons why they haven't yet adopted or fully implemented a solution.

Barriers to adoption: Prospective users

Reasons for not yet adopting streaming data (prospective users)

Concerned about integration with existing or legacy systems	**33%**
Data security or privacy concerns	**31%**
Concerned about how much training or set-up it would require	**26%**
Compliance, legal, or regulatory concerns	**24%**
Concerned about the value/getting a return on our investment	**24%**
Don’t have skills in-house needed for this technology	19%
Too expensive	17%
Isn't commonly used in our industry	16%
Hesitant or slow to adopt new technology	14%
Don’t understand capabilities or benefits of the technology well enough	10%
Don’t have clear picture of our data or are unable to accurately map it	9%
Internal leaders or external partners don't recommend this technology	9%
Don’t have a use case for the technology	5%
Don’t believe the technology can support the throughput we require	3%
Don’t believe the technology will solve our needs	3%

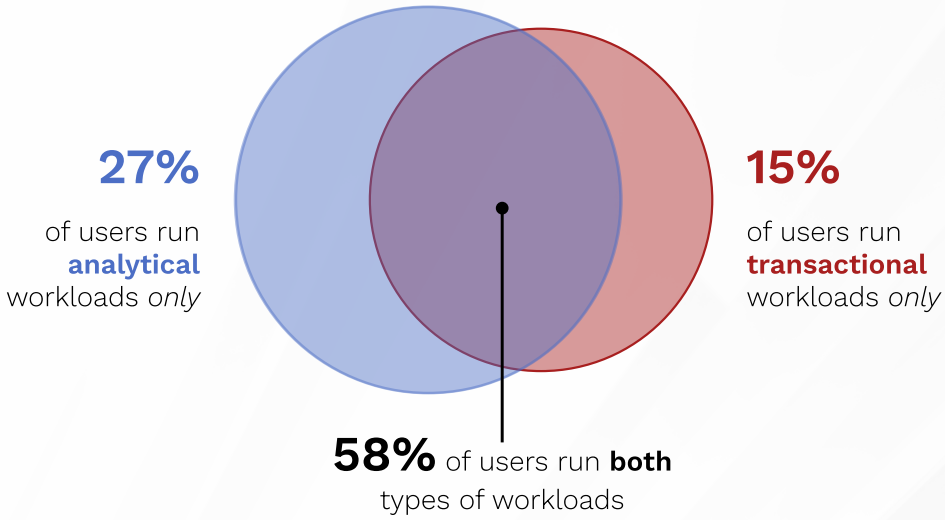
For non-users, four out of the five most cited reasons for not yet adopting a solution pertain to integration/setup and security/regulation. This further underscores the potential complexity of these systems and importance of addressing these concerns.

SECTION 04

The Size and Shape of Streaming Data

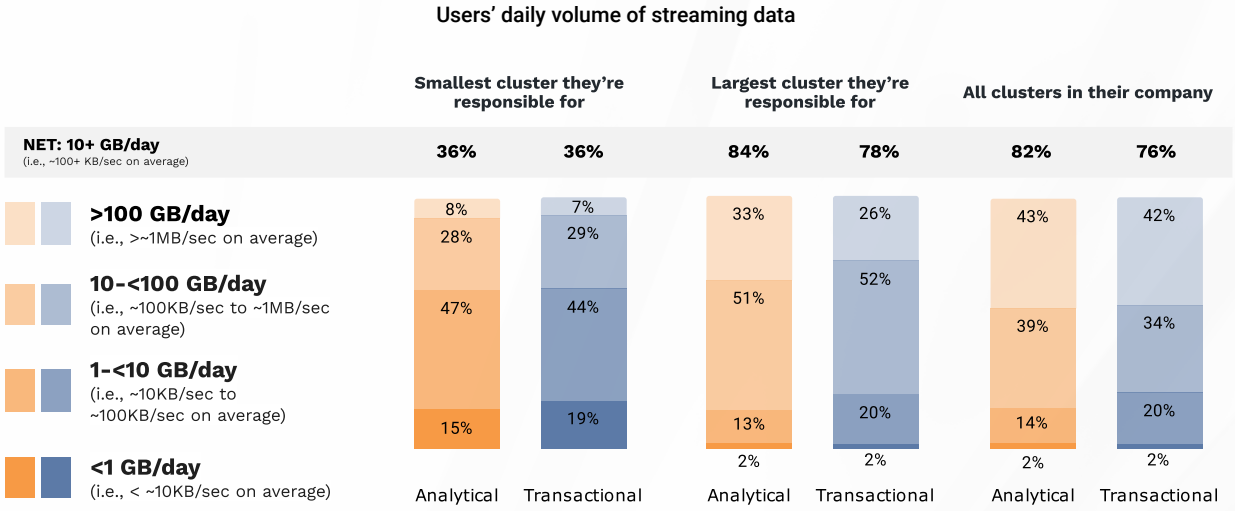
In this section we provide a baseline snapshot of streaming data and the characteristics of streaming data usage today. We include only current users in this analysis, split by the types of workloads in production – analytical versus transactional. We also look at user environments in terms of volume, throughput, latency and data retention.

Analytical vs transactional workloads

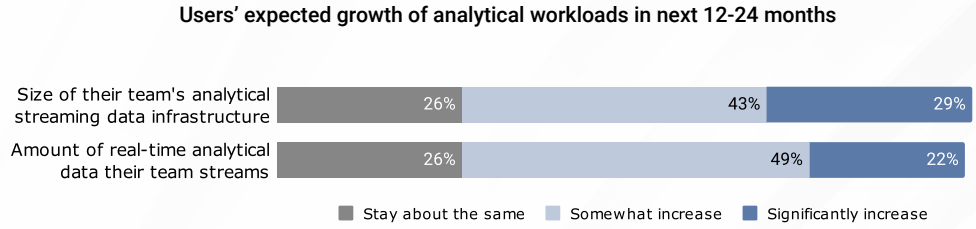


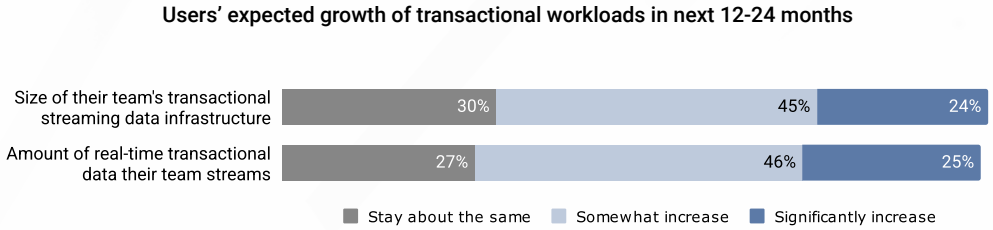
The majority of current users (58%) are currently running both analytical and transactional workloads on streaming data platforms. However, of those who run exclusively one or the other, the percentage of users running only analytical workloads (27%) is nearly double that of users running only transactional workloads (15%).

Data volume



The daily volume of data varies by cluster size. Roughly half of the workloads fall in the 1 to 10 GB/day range, while workloads of 10 - 100 GB/day come in as a distant second. For the respondents' largest clusters, a third to a quarter of workloads are greater than 100 GB/day. Interestingly, there is no significant difference between analytical and transactional workload volumes. However, the majority of current users expect their data volumes to grow, as we see below.





Nearly all users expect to see an increase in the size of their streaming data infrastructure and the amount of real-time data their teams stream. These increases hold consistent across both analytical and transactional workloads, with between two-thirds and three-quarters of current users expecting an increase of some magnitude. Respondents report two reasons for the growth in data volumes: Either their business is growing/expanding or their tools, systems and needs are evolving.

Here's what some of the users had to share on this topic:

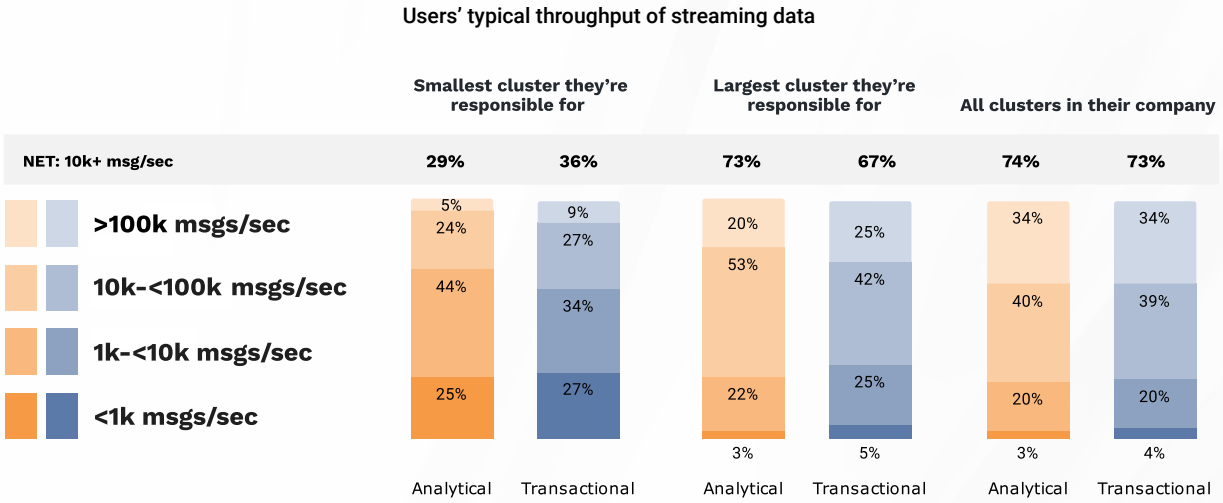
> *This is because we're expanding into new markets and currently deploying new infrastructure that will increase our data bandwidth and significantly improve the current transactional streaming to an all-time high.*

> *I expect the size of my team's transactional streaming infrastructure will significantly increase within the next 12-24 months considering we're constantly managing and improving technology within the workplace.*

> *"We are selling more augmented data services which are getting a lot of buy-in from clients, so we definitely see much more streaming and data augmentation."*

> *We plan to increase our analytics capabilities by introducing new tools and insights to our business. As our analytical streaming platform grows, we expect our infrastructure to scale accordingly.*

Throughput



Findings for throughput closely reflect findings for volume. Current users' typical message throughput (messages/sec) varies by cluster size. Their analytical clusters handle a slightly higher throughput than their transactional clusters. In addition, roughly 80% of clusters handle more than 10K msgs/sec.

Latency requirements

Latency Measurements for analytical data

% Ranked #1 Most Important	
40%	End-to-end latency
30%	Consume/fetch latency
30%	Producer (publish) latency

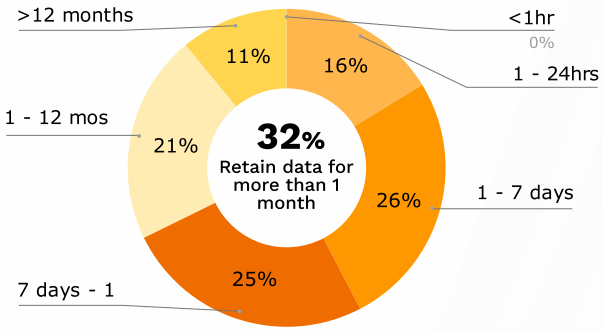
Latency Measurements for transactional data

% Ranked #1 Most Important	
38%	Consume/fetch latency
36%	End-to-end latency
25%	Producer (publish) latency

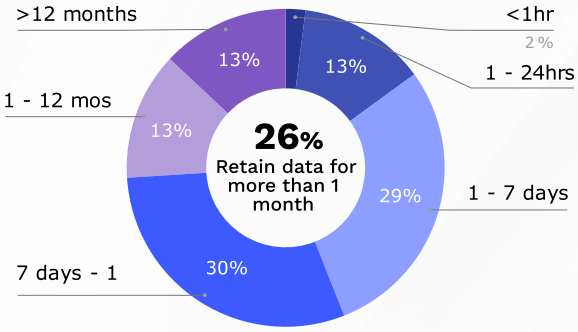
When asked which latency measurement is most important to them, current users cited that end-to-end latency is highly important to both analytical and transactional projects. Consume/fetch latency is relatively more important to transactional than analytical projects. This is what one would expect, as analytical use cases are usually less latency sensitive because the data is being used for things like reporting and long-term analysis, while for transactional use cases consume/fetch latency is critical for some real-time, event-driven functionality.

Data retention

Users' analytical data retention



Users' transactional data retention



Data retention policies seem to vary significantly from company to company. They also vary slightly when comparing analytical and transactional workloads. For example, 32% of respondents retain data for analytical workloads for more than 1 month, compared to 26% for transactional workloads.

Most respondents probably retain their data for as long as possible. Relatively short-lived data retention can mostly be attributed to two factors. First and foremost is cost. A second factor is regulatory concerns. In many industries, data retention is subject to specific requirements. So even though longer retention periods would probably be beneficial, in many cases that may not be possible.

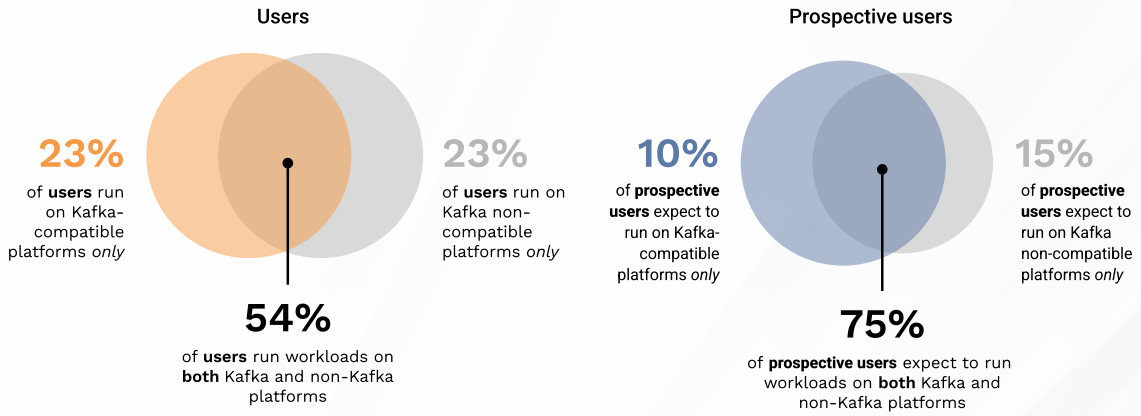
SECTION 05

Streaming Data Environments

This section is intended to help understand the platforms, pipelines, libraries, frameworks, tools, and hosting that are part of the streaming data ecosystem. Here we surveyed both current users to see what they now have in place, as well as prospective users to gauge their intentions. We analyze Apache Kafka-compatible solutions vs. those that are not compatible with Kafka.

The key take-away is that the streaming data category is a multi-platform space.

Kafka vs. non-Kafka



The majority of respondents run (or expect to run) streaming data workloads on a combination of Kafka and non-Kafka platforms. This is particularly pronounced in prospective users, where three quarters of respondents expect to have a mixed environment.

While Apache Kafka is the incumbent in the streaming data category, several solutions (Kafka-compatible or not) are vying for market share, including solutions from hyperscale cloud providers. Pragmatic users seem to understand their environments are (going to be) mixed.

Producers and consumers

Users' Most Common **Producers** of Streaming Data

63%	Transaction systems
50%	Sensors (IOT/IIOT)
40%	Clickstream
40%	Log collection
38%	Financial data/Operational data
31%	Event sourcing
24%	Telemetry
7%	Other streaming data clusters
3%	This is not applicable to our platform

Selected most often by users in large companies

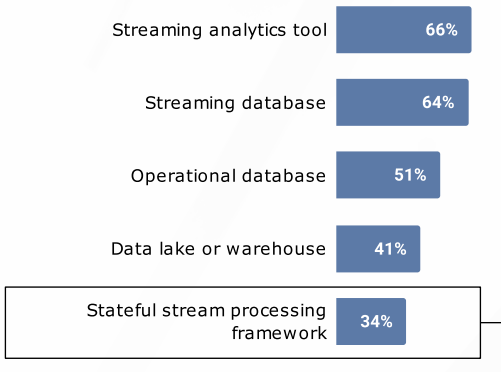
Users' Most Common **Consumers** of Streaming Data

60%	AI/ML pipeline
58%	Applications
55%	Analytic environments
49%	Data lake or warehouse
34%	Microservices
8%	Other streaming data clusters
3%	This is not applicable to our platform

Transaction systems and sensors lead as common producers of streaming data. Both are obvious options, as they are known to generate high volumes of data at high velocities.

AI/ML pipelines, applications, and analytic environments are the top-listed consumers for the majority of respondents. These results are in line with the drivers for adoption.

Pipeline components



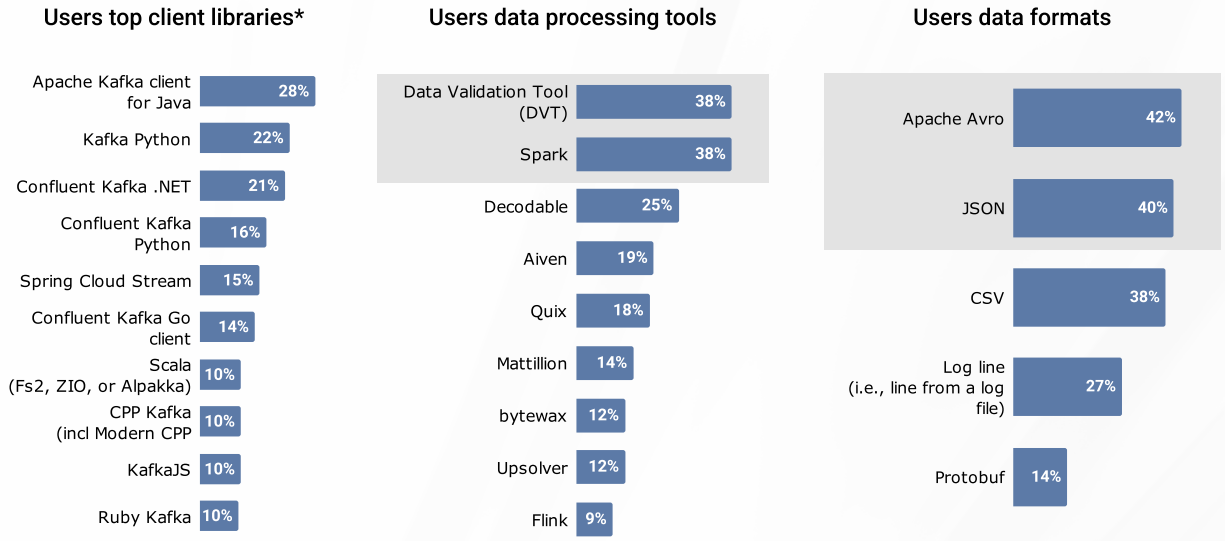
TOP STREAM PROCESSING FRAMEWORKS USED

Google Dataflow	39%
Apache Flink	28%
Apache NiFi	25%
Kafka Streams	21%
Spark Streaming/Spark Structured Streaming	21%
Apache Beam	20%
DeltaStream	16%
Apache Flume	16%
AKKA Streams/Alpakka	13%
Kafka Connect	13%
Materialize	11%
Quix Streams	11%

Frameworks with <10% not shown

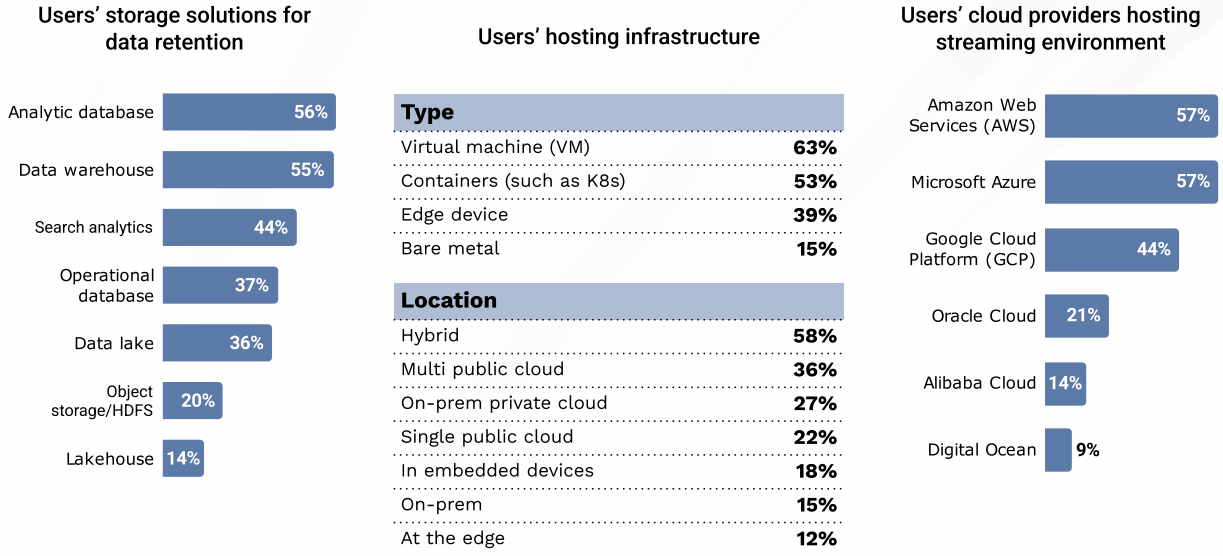
Streaming analytics tools and streaming databases are core components of the data pipeline for nearly two-thirds of users. More than half report using an operational database and slightly more than a third use a stateful stream processing framework.

Tools and libraries



The top client libraries used are fairly fragmented, but if we aggregate similar responses, leaders do emerge. In client libraries, top responses are all variants of Kafka-specific libraries. In validation tools, DVT and Spark are the leaders. In data formats, it's Apache Avro and JSON.

Storage and hosting

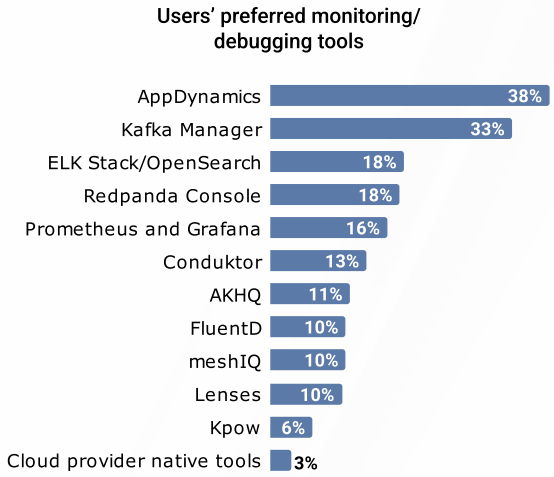


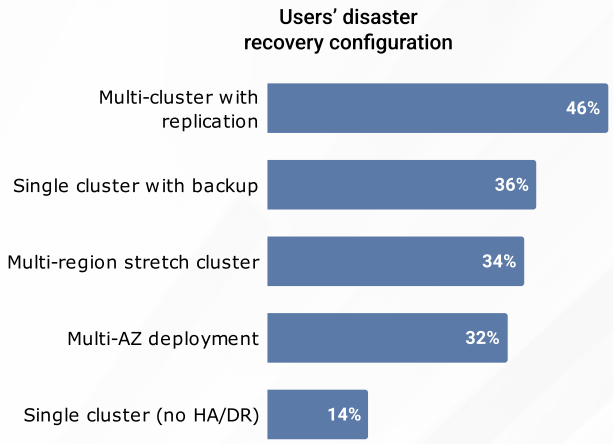
It is surprising that the utilization rate of data lakes (36%) and object storage (20%) are not higher, as these are typically the most economically viable solutions (remembering that costs were cited as a leading business concern around streaming data adoption).

VMs and containers are the most common types of hosting infrastructure at 63% and 53% respectively. Edge devices are also gaining ground to the point of becoming a mainstay, as they are used by 39% of respondents. Hybrid environments are used by 58% of respondents, and multi-public clouds by 36%. Those are the prevalent options for hosting infrastructure location, reflecting overall trends.

AWS and Azure are the top cloud providers for streaming data storage, tied at 57%. Google Cloud Platform comes in third at 44%.

Monitoring / debugging and disaster recovery





AppDynamics and Kafka Manager lead as most preferred tools for debugging streaming data pipelines at 38% and 33% respectively. The ELK stack and Redpanda Console share third place, tied at 18%. AppDynamics and the ELK stack are both market leaders in monitoring, so that seems to have carried over to streaming data. Kafka Manager and (Kafka-compatible) Redpanda Console, on the other hand, are streaming data-specific solutions.

In terms of disaster recovery configuration, we see a number of variations of high availability. Multi-cluster with replication is the most common high availability disaster recovery environment. 14% of respondents utilize only a single cluster without HA/DR features – we speculate (and hope!) that these are likely non-production use cases.

Conclusion

The picture that emerges from the findings of this survey is an encouraging one. The importance of streaming data is widely understood. The majority of respondents are working on projects that help bring the benefits of streaming data to production, albeit at different stages of maturity.

Users and prospective users alike see a large number of benefits, value, and business applicability from the use of streaming data, with real-time analytics and AI/ML being the drivers for adoption. However, there is still much work to be done. Both current users and prospective users have concerns around critical technical issues such as data security, privacy and governance as well as business issues including costs and the lack of in-house expertise to implement and manage these complex systems.

Overall, the streaming data category reflects usage patterns prevalent in technology today. It's a multi-cloud, multi-platform space, with lots of diversity. Going forward in this rapidly growing industry, we expect to see platforms evolving to accommodate the needs of this diverse audience with many different applications, ranging from the critical to the innovative.

SECTION 06

Methodology and Audience

The purpose of this research was to generate data to understand the extent to which organizations are currently deploying streaming data technology, their specific solutions, footprints and performance, and their future plans for the use of this technology.

The research was conducted in July and August 2023 using an online survey. Respondents were screened according to rigorous criteria. All 300 respondents are based in the United States, employed full-time in a company of at least 20 employees, and knowledgeable about streaming data.

Respondents are current or prospective users of streaming data technology who work in IT, Software or Data Engineering, Infrastructure Operations, and Application Development. They are influential in decisions about implementing or managing data streaming at their company, with a variety of levels of seniority represented—from individual contributors (with a minimum experience requirement) through C-suite.

Adoption status

Current users of streaming data are the majority of respondents at 59%. Of those, 32% already have a solution in place and 27% are in the process of implementing a solution.

The remaining 41% of respondents are not currently using a streaming data system. Of those, 17% are researching or gathering information about the technology, 14% are piloting or participating in a proof-of-concept for a solution, and 10% are planning to evaluate the technology within 12 months.

Company size and seniority level

Company size

- 20-100 employees | **12%**
- 101-500 employees | **26%**
- 501+ employees | **62%** — **Users skew toward larger companies**

Level of seniority

- Individual Contributor (min. 5 years experience) | **1%**
- Manager | **17%** or Senior Manager | **18%**
- Director | **20%** or Senior Director | **8%**
- Vice President | **3%** or Senior Vice President | **3%**
- C-Level Executive (not CEO) | **19%** — **Users skew toward more senior roles**
- President or CEO | **4%**
- Owner | **5%**

The vast majority of respondents work in large companies (500+ employees). At the same time, a considerable part of respondents are C-level executives. That signifies that there is a cohort of C-level executives in mid-market companies who work closely with streaming data.

More senior roles skew toward smaller companies. Respondents in the C-suite and above are disproportionately more likely to be at smaller companies. Of participants from small companies, 47% are in the C-suite or above, compared to just 27% in larger companies. Users lean more heavily into more strategic motivations for using data streaming.

Prospective users are more likely to hold less-senior roles. 81% of prospective users are in roles below the C-suite compared to 65% of users who are below the C-suite. This may help explain why prospective users seem to be more focused on tactical goals and use cases.

Industries and functional areas

Industries

- Technology/Consumer Electronics | **39%**
- Software/Internet | **30%**
- Manufacturing | **11%**
- Retail/Wholesale Trade | **7%**
- Financial Services/Insurance | **5%**
- Healthcare | **5%**
- Telecommunications | **2%**
- Media & Entertainment | **1%**

Functional area

To qualify, respondents must work in one of the following roles. Users were more heavily concentrated in these roles, while non-users had more role diversity.

- Information Technology | **90%**
- Software or Data Engineering | **43%**
- Infrastructure Operations | **19%**
- Application Development | **25%**

About the Sponsor

This report was commissioned by Redpanda Data. Redpanda provides a simple but powerful Apache Kafka®-compatible streaming data platform that works both as a fully managed cloud service and a self-hosted platform. The easiest way to try Redpanda is signing up for a free 14-day trial of Redpanda Serverless, which spins up a cluster in seconds. You can learn more about Redpanda and its technology at redpanda.com.