CONFLUENT

Transform Your Data Pipelines, Transform Your Business

3 ways to get started

Data is the lifeblood of today's enterprises—especially when it's used in real time. It helps organizations uncover insights that inspire innovations, deliver personalized experiences, and operate more intelligently and efficiently. However, in many companies, data is siloed, fragmented, and stored in multiple formats across numerous legacy and cloud-based systems. In fact, 60% of tech leaders say that difficulty integrating multiple data sources is their biggest hurdle to accessing more real-time data.

To make data more accessible, most IT organizations centralize as much information as possible. They typically use point-to-point data pipelines to move data between operational databases and a centralized data warehouse or lake. For example, extract, transform, and load (ETL) pipelines ingest data, transform it in periodic batches, and later push it downstream to an analytical data warehouse. ETL pipelines—and reverse ETL pipelines—can also share the results of data analysis that takes place in the warehouse back to operational databases and applications.

In this guide, you'll learn about the challenges that come with legacy data pipelines and the benefits of adopting streaming pipelines. You'll also discover how four organizations are using streaming data pipelines from Confluent—and how these pipelines could transform your business, too.

60% of tech leaders say that difficulty integrating multiple data sources is their biggest hurdle to accessing more real-time data.

Why traditional data pipelines don't scale

Today's enterprises often manage numerous point-to-point data pipelines, which are challenging to maintain. A growing number of IT organizations are concluding that their pipeline approaches to sharing actionable data don't scale. Reasons include:

1 Batch processing

Traditional pipelines move data in periodic batches, and the resulting low-fidelity and stale data can't power real-time analytics or deliver data to power real-time applications.

2 Centralized data teams

Data pipelines and data warehouses are usually governed by a single centralized data engineering team that must approve individual requests for data or spend time building new pipelines and maintaining existing ones. This often becomes a bottleneck to innovation.

3 Immature governance and observability

Most companies run a patchwork of numerous point-to-point pipelines, which require a very large data engineering team to manage, maintain, and secure. Observability, data lineage, and data and schema error management are also challenges—and those lead to organizational inertia and lost opportunities.

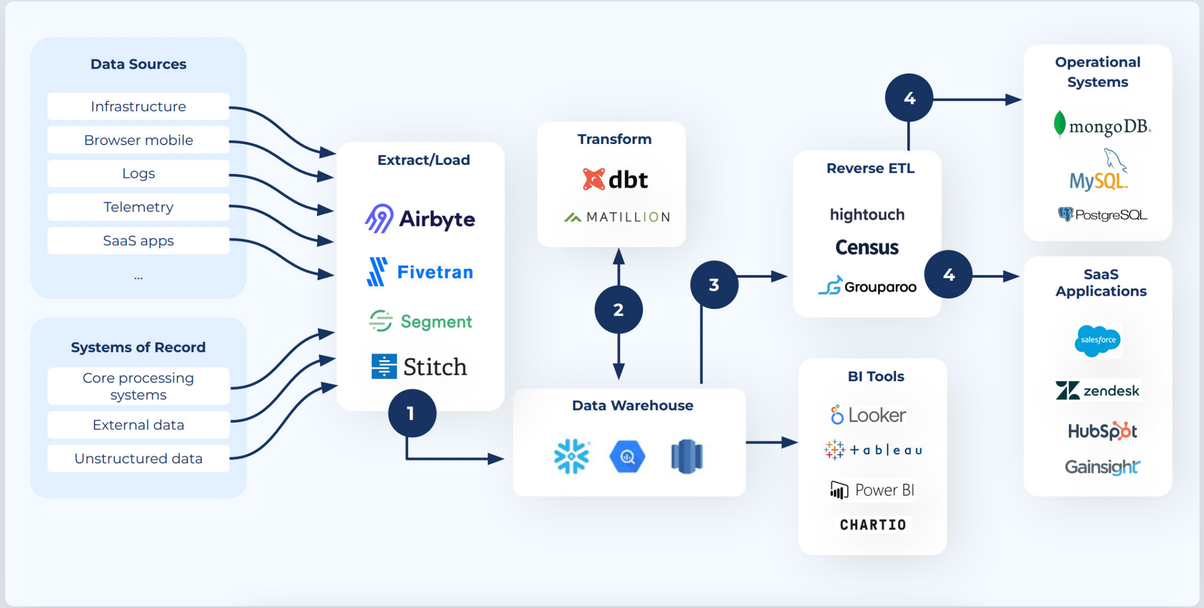
4 Resource-intensive data processing

Traditional pipelines require intensive computing power and storage, and requirements are increasing as data volumes grow. Plus, inconsistent workload variations make it difficult to predict resource requirements, which leads to scale and performance challenges and high operating expenses.

5 Monolithic design

Traditional pipelines are known for their rigidity, and developers often perceive them as "black boxes" because they lack customizations and are difficult to port across different environments. When business logic or data needs to change, engineers are thus fearful of changing existing pipelines. Instead, they prefer to add "just one more" pipeline to address the new requirement, increasing pipeline sprawl and technical debt.

The "modern" data stack is built on a legacy paradigm



How streaming data pipelines are different

Streaming is a modern approach to building data pipelines that allows teams to share real-time data in many different contexts through a decoupled architecture rather than integrating it in a centralized data warehouse. It uses change data capture (CDC) capabilities to continuously intercept changes to databases as streams—combining, enriching, and analyzing them in motion even before they reach at-rest systems such as the database or data warehouse.

Unlike traditional pipelines, streaming data pipelines can be designed using declarative languages such as SQL to specify the logic of what needs to happen while abstracting away the low-level operational details. This approach helps to maintain the delicate balance between centralized continuous data observability, security, policy management, and compliance standards and the need to make data easily searchable and discoverable so that developers and engineers can innovate faster.

In addition, streaming data pipelines invite IT organizations to apply Agile engineering practices in order to build modular, reusable data flows that can be tested and debugged using version control and CI/CD systems. This characteristic makes streaming data pipelines easier to evolve and maintain—and reduces their total cost of ownership compared to traditional approaches.

Getting started with streaming data pipelines

Streaming data pipelines are a modern approach to delivering data as a self-service product. Instead of sending data to a centralized warehouse or an analytics tool before making it available to applications, streaming data pipelines can capture changes to data in real time, enrich them on the fly, and send them to downstream systems. And because teams get self-service access, they can find, browse, create, share, and reuse data—wherever and whenever it's needed.

Confluent makes it easy and cost-effective to build streaming data pipelines and to evolve them as business and data needs change. Using Confluent, you can build and deploy modern data flows in five simple steps:

1 Connect

Create and manage data flows with an easy-to-use visual user interface and pre-built connectors.

2 Govern

Centrally manage, tag, audit, and apply policies for trusted high-quality data streams.

3 Enrich

Use SQL to combine, aggregate, clean, process, and shape data in real time, increasing the safety, efficiency, and usability of your data streams to power operational, analytical, and business intelligence use cases.

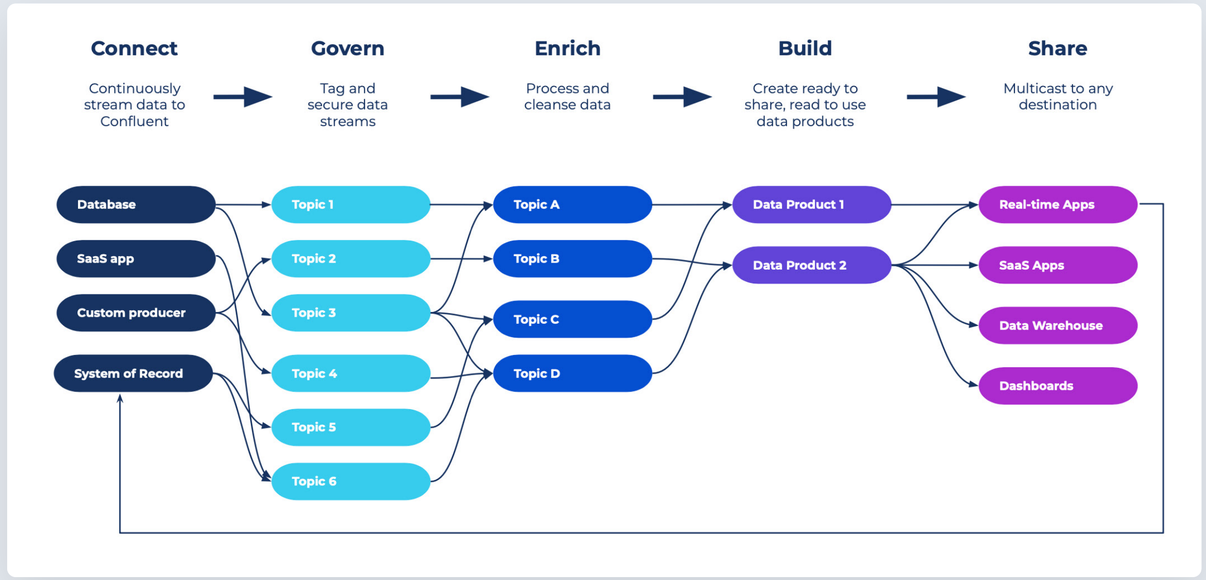
4 Build

Prepare well-formatted, trustworthy data products for downstream systems and apps.

5 Share

Collaborate securely on live streams with self-service data discovery and sharing.

Next-gen data lifecycle with Confluent



Streaming data pipelines can capture changes to data in real time, enrich them on the fly, and send them to downstream systems. Teams can find, browse, create, share, and reuse data—wherever and whenever it's needed.

Three key use cases for streaming data pipelines

In this section, we'll outline the real-world stories of how four organizations are using streaming pipelines from Confluent to deliver rich front-end customer experiences and real-time back-end operations.

Use case #1: Database pipelines

Streaming pipelines can move data between on-prem databases—such as MySQL, SQL Server, and PostgreSQL and cloud databases such as MongoDB, DynamoDB and CosmosDB—where it can be quickly combined, enriched, and used to make critical operational decisions in real time. Organizations need to be able to harness data for real-time insights into how their business is performing, and to enable data teams to make timely operational decisions in response to market conditions.

Database pipelines can accelerate and bring intelligence to key operational activities, such as customer 360, credit applications, ecommerce checkout, and product fulfillment. By enabling faster, smarter operations, operational database pipelines can have a major impact on both revenue and costs.

DATABASE PIPELINES IN ACTION:

SecurityScorecard builds streaming data pipelines to scale faster, govern data better, and ultimately lower TCO.

The global leader in cybersecurity ratings and the first cybersecurity ratings company to offer digital forensics and incident response services, SecurityScorecard tracks potential weaknesses—exposed servers, suspicious IP addresses, breached employee accounts, malware-infected devices, out-of-date endpoints, and much more—with certainty.

Accurate, up-to-date data is the lifeblood of this business, and it needs to come from a myriad of sources across the internet. "Our platform relies on data collection and processing being done extremely accurately, in real time and at scale," says Jared Smith, Senior Director, Threat Intelligence, SecurityScorecard. This is what led the company's engineering teams to adopt data streaming, and it's how they ended up using Confluent to build streaming data pipelines. Now the company can create a source of truth in their PostgreSQL database, share this data to downstream systems using a combination of polling and change data capture (CDC) connectivity, and ultimately lower its total cost of ownership (TCO) by offloading the management of this key piece of infrastructure to Confluent.

Use case #2: Data warehouse pipelines

A data warehouse, whether it's on-premises or in the cloud (examples include Databricks, BigQuery, and Redshift), is only as good as the information it contains.

Unlocking real-time insights requires a streaming architecture that's continuously ingesting, processing, and provisioning data in real time. With Confluent, you can build streaming data pipelines from hybrid and multicloud data sources to the cloud data warehouse of your choice, unlocking real-time analytics and decision-making while reducing total cost of ownership (TCO) and time to value (TTV).

AN IN-DEPTH LOOK INTO DATA WAREHOUSE PIPELINES:

Toolstation takes a modern approach to data streaming in the cloud.

Toolstation is an omnichannel retailer of building tools and materials with more than 500 branches in the UK and more than 100 in the Netherlands, Belgium, and France. During the pandemic, the company struggled to scale click-and-collect fulfillment with its legacy MySQL database, which relied on multiple polling processes and batch jobs to stay up to date. With Confluent, the retailer used stream processing and connectors to gradually switch to data streaming—and removed the limitation on click-and-collect orders. The update was tested, scaled, and moved into production in just six weeks.

Picnic processes more than 300 million unique events per week to power predictive analytics in their data warehouse.

Picnic is Europe's fastest-growing online-only supermarket and relies heavily on data-driven decisions to provide the lowest-price guarantee to its customers. Their blazing growth underpinned the need to seek out a solution for real-time processing and easy scalability of event data to power predictive analytics in their data warehouse. With Confluent, Picnic was able to use fully managed sink connectors to load data seamlessly from upstream systems into Snowflake and Amazon S3 for analysis by data science teams while reducing infrastructure costs by 40%.

Use case #3: Mainframe integration

Mainframe systems have demonstrated tremendous staying power. As important as these systems are, though, it can be difficult for today's data-driven businesses to integrate them with other systems. Obtaining data from these systems in real time can require the development and ongoing maintenance of costly custom connectors.

Streaming data pipelines allow mainframe data to be accessed in real time without the complexity and expense of sending ongoing queries to mainframe databases. They give enterprises the capabilities they need to power cloud applications and systems with real-time data streams from mainframe systems while lowering consumption costs.

A PRACTICAL APPLICATION OF MAINFRAME INTEGRATION:

Amway unlocks mainframe data for new use cases.

With annual sales of more than $8 billion, Amway is the world's largest direct-selling company. Its nutrition, beauty, personal care, and home products are sold in more than 100 countries. IT and business solutions teams within Amway placed a high priority on modernizing their digital footprint and moving fast to deliver needed solutions on tight schedules. The company's legacy IT architecture, however, was not well aligned with this priority.

To solve this problem, Amway decided to use capabilities within Confluent, including ksqlDB, to accelerate streaming app development and build connectors to easily integrate data from legacy mainframes to modern systems. Combining cloud scalability with an event-driven architecture has allowed Amway to rapidly stand up new experiences—such as mobile selling—and plug them into the data mesh being built with Confluent.

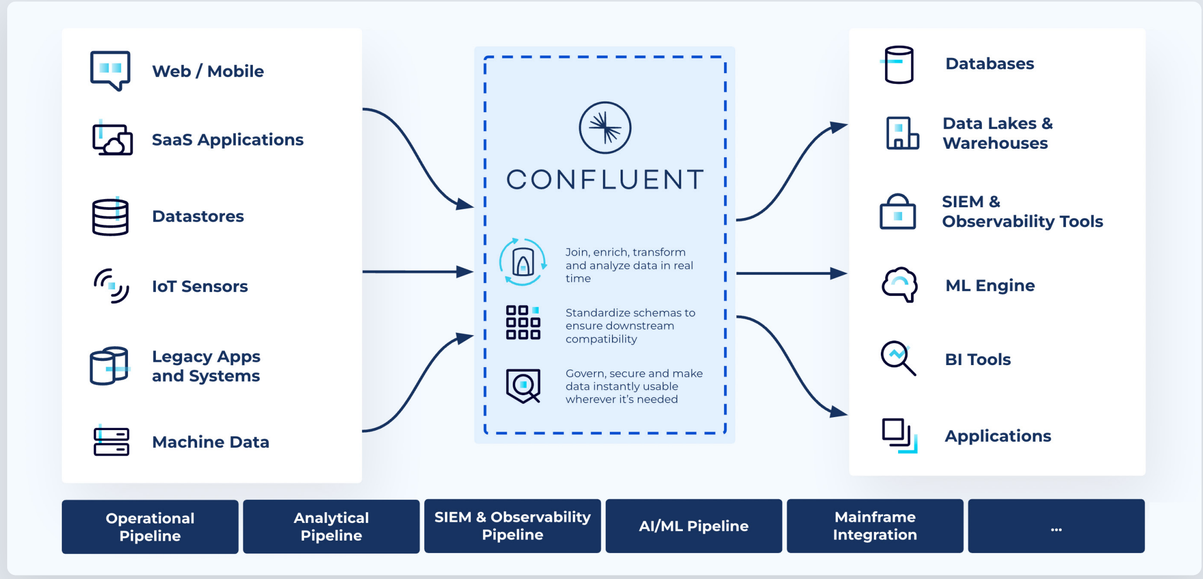
It's time for streaming data pipelines

Modern leading businesses run on real-time data. Point-to-point and batch-based data pipelines are no longer suitable for any organization that hopes to lead its sector. Today's status quo—batch-oriented, centralized, ungoverned, and inflexible—isn't enough.

Streaming data pipelines will help you solve your data needs today, and they will serve your business needs in the future. With streaming data pipelines from Confluent, you can transform your business and meet ambitious goals such as these:

- Power all your use cases with high-quality, real-time data streams
- Bring new products to market faster with self-service data access
- Boost developer productivity by simplifying data flow development and iteration
- Enable agile pipeline development to meet changing data needs
- Accelerate innovation while maintaining trust and governance

Integrate streaming data across apps and data systems to unlock unlimited real-time use cases



Analyze data in real time, ensure downstream compatibility, and make data instantly usable wherever it's needed.

Take the next step

Sign up for our free trial.

Watch a demo of how easy it is to build streaming data pipelines with Confluent.

Visit our streaming data pipelines solutions page.

CONFLUENT

ABOUT CONFLUENT

Confluent is pioneering a fundamentally new category of data infrastructure focused on data in motion. Confluent's cloud-native offering is the foundational platform for data in motion—designed to be the intelligent connective tissue enabling real-time data from multiple sources to constantly stream across the organization. With Confluent, organizations can meet the new business imperative of delivering rich digital front-end customer experiences and transitioning to sophisticated, real-time, software-driven back-end operations.

To learn more, please visit

www.confluent.io