STREAMING DATA EXPLAINED: BENEFITS, ARCHITECTURE AND CHALLENGES



Data streaming is data that continuously flows from a source to a destination to be processed and analyzed in near real-time.

What is data streaming?

Companies can have thousands of data sources that get piped to different destinations. The data can be processed using stream processing techniques, and generally consists of small chunks of data.

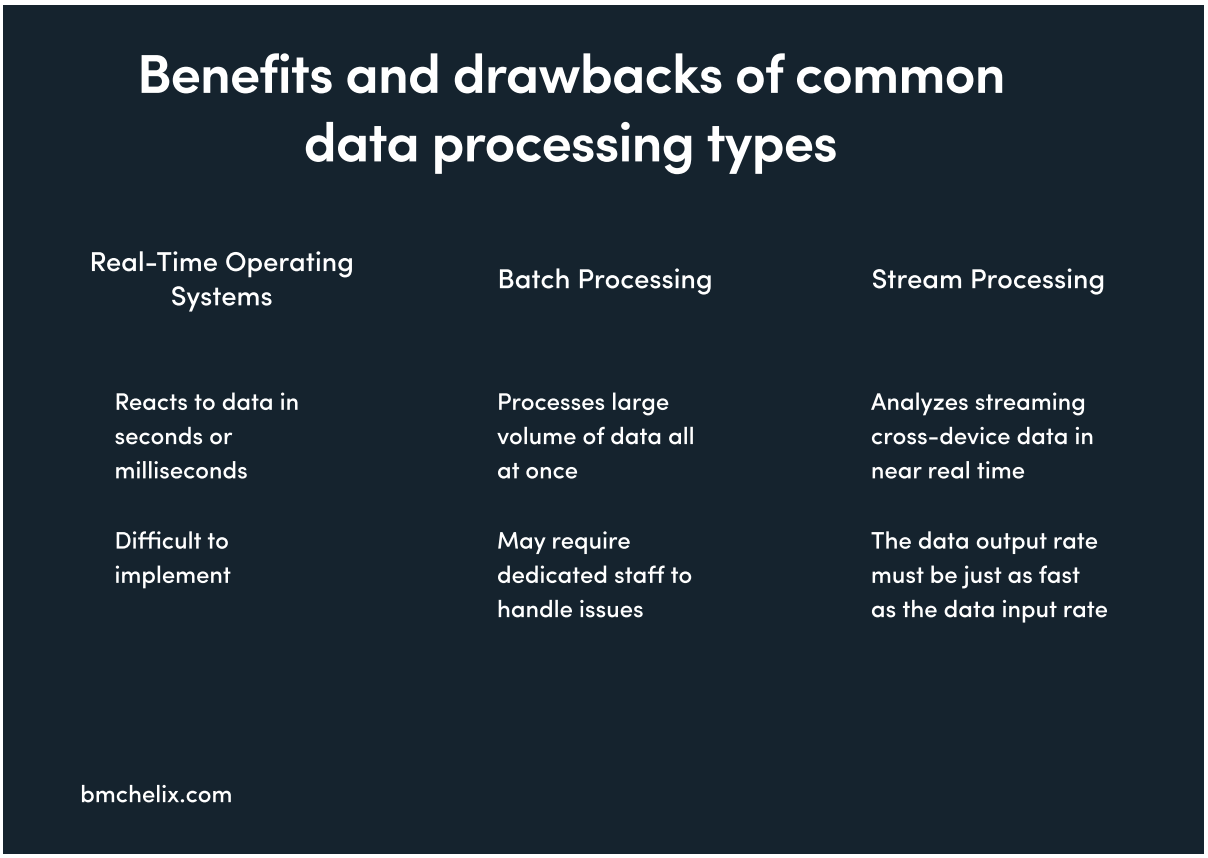
Streaming data allows pieces of data to be processed in real or near real-time. The two most common use cases for data streaming:

- Streaming media, especially video
- Real-time analytics

Data streaming used to be reserved for very select businesses, like media streaming and stock exchange financial values. Today, it's being adopted in every company. Data streams allow an organization to process data in real-time, giving companies the ability to monitor all aspects of its business.

The real-time nature of the monitoring allows management to react and respond to crisis events

much quicker than any other data processing methods. Data streams offer a continuous communication channel between all the moving parts of a company and the people who can make decisions.



Streaming media

Media streaming is one example. It allows a person to begin watching a video without having to download the whole video first.

This allows users to begin viewing the data (video) sooner, and, in the case of media streaming, prevents the user's device from having to store large files all at once. Data can come and go from the device as it is processed and watched.

Real-time analytics

Data streams enable companies to use real-time analytics to monitor their activities. The generated data can be processed through time-series data analytics techniques to report what is happening.

The Internet of Things (IoT) has fueled the boom in the variety and volume of data that can be streamed. Increasing network speeds contribute to the velocity of the data.

Thus we get the widely accepted three V's of data analytics and data streams:

1. Variety
2. Volume
3. Velocity

Paired with IoT, a company can have data streams from many different sensors and monitors, increasing its ability to micro-manage many dynamic variables in real-time.

From a chaos engineering perspective, real-time analytics is great because it increases the company's ability to monitor the company's activities. So, if equipment were to fail, or readings were to send back information that needed quick action, the company has the information to act.

Data streams directly increase a company's resilience.

Benefits of streaming data

Organizations enjoy advantages when they can stream data, instead of waiting for traditional batch processing. Getting an uninterrupted flow of information in real time can make for faster decisions. Instead of reacting, executives can get ahead of trends, customer demands, and operational irregularities. By streaming data into your processes, your organization can be more agile, resilient, and innovative. Benefits of streaming data include:

1. **Real-time insight:** You don't have to wait for a batch process to get data to analyze. You can use it right away to make time-sensitive decisions. Detect fraud immediately, jump on a consumer trend, and spot a quality issue before it evolves into a bigger problem.
2. **Enhanced operational efficiency:** Slowdowns, process inefficiencies, and weak links are easier to spot in workflows when you can analyze continuous data in real time. Not only can you boost productivity and trim costs, you can also prevent problems and downtime.
3. **Predictive analytics**: Streaming data into predictive models strengthens your ability to look ahead and make decisions. For example, you can forecast demand so that you can optimize inventory or monitor equipment to prevent failures.
4. **Personalized customer experiences:** The ability to analyze customer behavior in real time provides insights to fuel product recommendations, offer relevant promotions, and engage in ways that drive loyalty.
5. **Better risk management:** Live data makes it possible to detect irregular trends and patterns that could indicate fraud as they emerge. You may also be able to detect anomalies that could signal a coming supply chain disruption, giving you time to make adjustments.
6. **Improved agility and responsiveness:** Streaming data empowers teams to make fast and precise changes in response to evolving realities. Having real-time data helps you go in the right new direction.
7. **Shared company-wide perspectives:** When all departments can see what is happening when it is happening, it is easier to coordinate data-driven strategies and responses. No one has to wait for IT to generate reports or to worry that one department is operating in a data silo.

Data architecture for streaming data

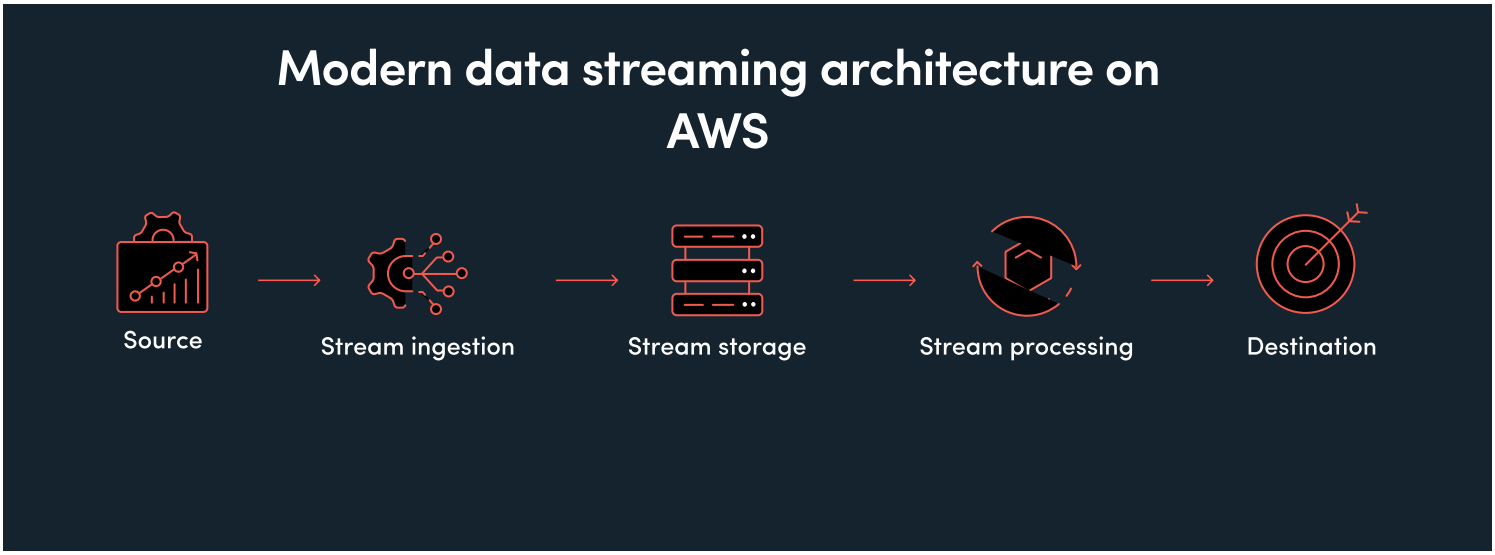
Streaming data architecture involves creating systems to ingest and use real-time data from multiple sources, including IoT devices, transaction systems, social media, application logs, and more. Streaming data architecture supports real-time processing and analysis so users can make decisions without waiting for batch processing to give them the information they need.

A successful streaming data architecture ensures that its components work together efficiently. It has to scale to handle high-velocity and high-volume data without latency or failure. Current best practices for performance and reliability include modular design, distributed computing, and event-driven processing. The key components are:

- **Data ingestion layer:** This component captures data in real time using tools like Apache Kafka,

AWS Kinesis, or RabbitMQ. They handle high-velocity and high-volume data from sensors, applications, and services, with high throughput and low latency.

- **Processing layer:** Processing has to match the volume and velocity of the data ingestion component to process events, detect patterns, apply business rules, and support immediate action. Stream processing frameworks that fit the bill today include Apache Flink and Apache Spark Streaming.
- **Storage layer:** How streaming data is stored ensures that it not only fuels real-time analysis, but that it can also be used for future analysis. Time-series storage solutions like InfluxDB and NoSQL databases like Cassandra or MongoDB can be optimized for fast read and write, and make data quickly accessible for future retrieval and queries.
- **Data output layer:** This component ensures that processed data can feed into real-time reporting tools, data visualization dashboards, and other downstream applications. It supports the instant and accurate decision-making that defines agility and responsiveness.



Challenges with data streaming

Data streams offer continuous streams of data that can be queried for information.

Generally, the data will need to be in order, which is sometimes the point of having a stream. (After all, any messaging app needs to have all the messages in order.)

Because data may come from different sources, or even the same source, but it moves through a distributed system, it means the stream faces the challenge of ordering its data and delivering to its consumer.

So data streams directly encounter the CAP theorem problem in its build. When choosing a database or a particular streaming option, the data architect needs to determine the value between:

- **Having consistent data**, where all the reads received are the most recent write, and, if not return an error.
- **Having highly available data**, where all the reads contain the data, but they might not be the most recent.



Data streaming technologies

Streaming data architecture built to handle high-volume, high-velocity data makes it possible to process data, gain insights, and take action on information as it happens. Cloud-based and modular data streaming technologies support the foundational principles of real-time data processing, including scalability, fault tolerance, and minimal latency.

Teams already familiar with cloud environments can access a robust suite of tools covering every layer of streaming architecture from Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP). With a flattened learning curve, teams with cloud expertise can concentrate on customizing and optimizing the architecture to best support organizational needs.

The first step in building out your streaming data architecture is setting up the data ingestion layer or stream processor. Common tools to handle diverse and distributed data sources with high throughput and low latency include:

- Amazon MSK
- Amazon Kinesis
- Apache Kafka

- Google Pub/Sub

The second step is to choose and configure the stream processing and analysis layer that performs on-the-fly data transformations, aggregations, and filtering. You will also need to apply business logic, make queries, detect anomalies, track trends, trigger alerts, and make data actionable in real time. You can use tools like:

- Amazon Kinesis Data Analytics
- Google Dataflow
- Google BigQuery

Third, you need to get the analyzed data to decision-makers in a useful form. Typically, you will use a dashboard, alert system, or an operational application that gives teams the ability to visualize data and capture insights. You can customize a data destination or use tools like Amazon QuickSight or Looker.

Finally, you need to store the streamed data somewhere. The cost of storage is cheap, so the general practice is to keep everything. When it comes to data storage, the common belief is that if it's not useful now, it may be later.

For storing streaming data, these are good options:

- Amazon Redshift
- Kafka
- Amazon S3
- Google Storage

Popular streaming data technologies

A growing list of data streaming technologies that power robust and scalable real-time analytics includes:

- **Apache Kafka:** This distributed event streaming platform offers fault-tolerant data ingestion with high throughput. Kafka's strength is handling large data volumes for event-driven applications.
- **AWS Kinesis:** This fully managed cloud service offers components like Kinesis Data Streams, Firehose, and Data Analytics to give you end-to-end streaming capabilities for collecting, processing, and analyzing data in real time.
- **Apache Flink:** This high-powered stream processing framework handles both unbounded and bounded data streams with low latency.
- **Apache Spark Streaming:** As part of the Apache Spark ecosystem, this component supports scalable, high-throughput, and fault-tolerant data stream processing. It is a versatile choice that supports hybrid architectures that include batch processing.
- **RabbitMQ:** This message broker uses message queuing to support real-time communication between services, with reliable message delivery and flexibility.

Leveraging data streaming in your organization

Data streaming technologies and architectures are critical to your organization's ability to make informed decisions in real time. Speed, agility, and responsiveness are all strategic advantages.

Instant, data-driven decision making is the competitive edge for uncovering new opportunities, optimizing operations, reducing risks, and spurring innovation. Data streaming enables your organization to work smarter and faster, for agile responsiveness and proactive decision-making.

Related reading

- BMC Machine Learning & Big Data Blog
- 3 Simple Data Strategies for Companies
- What is data normalization? Benefits and normal forms explained.
- Data Ethics for Companies
- What Is Data Gravity?