# DATA STREAMING EXPLAINED: PROS, CONS & HOW IT WORKS

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

## How data streaming works

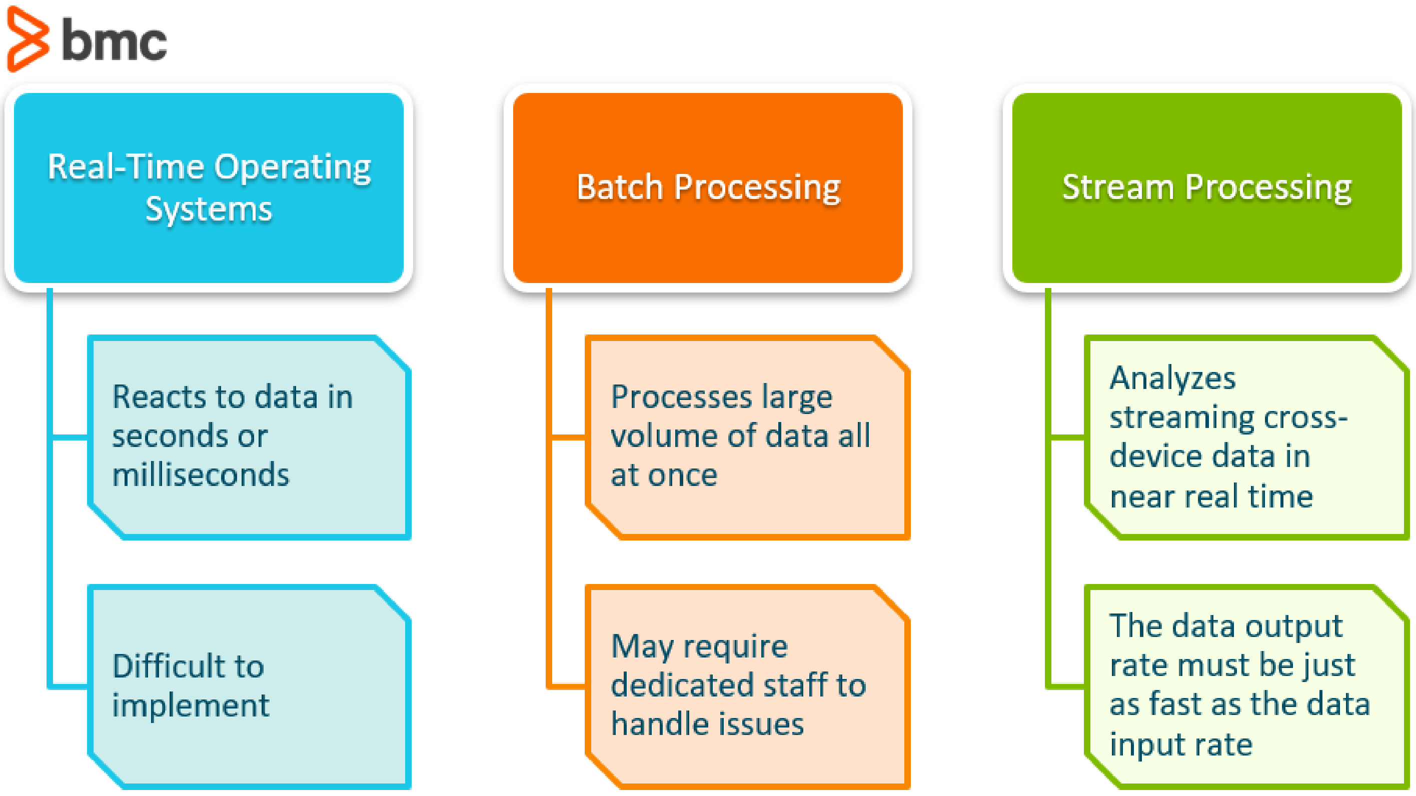
Companies can have thousands of data sources that get piped to different destinations. The data can be processed using stream processing techniques, and generally consists of small chunks of data.

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.


bmc
Real-Time Operating Systems
Reacts to data in seconds or milliseconds
Difficult to implement
Batch Processing
Processes large volume of data all at once
May require dedicated staff to handle issues
Stream Processing
Analyzes streaming cross-device data in near real time
The data output rate must be just as fast as the data input rate

*Benefits & drawbacks of common data processing types*

## 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.

## Data architecture for streaming data

Data streams require a certain kind of architecture, which is easier to adopt if you're already familiar with cloud architectures. The bulk of the learning curve has been climbed, and the rest is just adding pieces of knowledge here and there.

Many cloud service companies offer the tech to build a data stream—the usual Amazon, Azure, Google.

You can also build your own data stream. The first step is building a stream processor, just something to capture the streaming data from an application or device.

To build a stream processor, you can use tools like:

- Amazon MSK
- Amazon Kinesis
- Apache Kafka
- Google Pub/Sub

The second step is to add some way to analyze or query the streaming data, using tools like:

- Amazon Kinesis Data Analytics
- Google BigQuery, Dataflow

Third, the analysis needs to output somewhere, either to an app or a dashboard, so users can be alerted and respond to the information.

Finally, the streamed data needs to get stored somewhere. The cost of storage is cheap, so general practice is to keep everything. When it comes to data storage, the common belief is that "If it's not useful now, maybe it will be later."

For storing streaming data, these are good options:

- Amazon Redshift
- Kafka
- Amazon S3
- Google Storage

*(Learn more about working with data in Redshift and S3.)*

## Challenges with data streaming

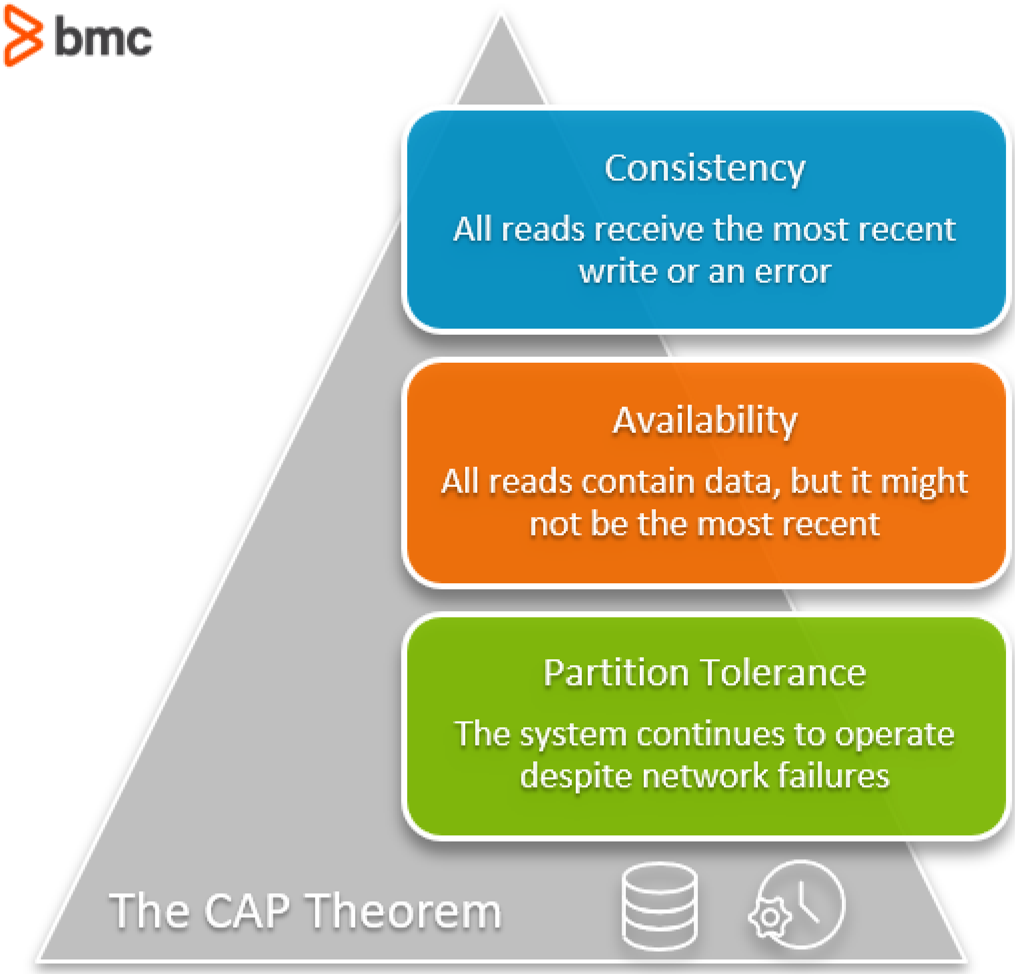
Data streams offer continuous streams of data that can be queried for information.

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.



Consistency

All reads receive the most recent write or an error

Availability

All reads contain data, but it might not be the most recent

Partition Tolerance

The system continues to operate despite network failures

## Related reading

- BMC Machine Learning & Big Data Blog
- 3 Simple Data Strategies for Companies
- Data Ethics for Companies
- What Is Data Gravity?