WHAT IS A DATA PIPELINE?

Data is the oil of our time—the new electricity. It gets collected, moved, refined.

The **data pipeline** encompasses how data travels from point A to point B; from collection to refining; from storage to analysis. It covers the entire data moving process, from where the data is collected, such as on an edge device, where and how it is moved, such as through data streams or batch-processing, and where the data is moved to, such as a data lake or application.

The data pipeline should seamlessly get data to where it is going and allow the flow of business to run smoothly. If the pipeline gets held up, quarterly reports can be missed, KPIs uninformed, user behaviour not processed, ad revenue lost, etc. Good pipelines can be the lifeblood of an organization.

It used to be that trustworthy members on teams were the endpoints to send this information from one point to another. In today's world, there are reliable software systems that move the data around. A good pipeline will get your data from its source to its destination timely and securely.

Data processing: streamed or batched?

Data can be processed in a few ways, but streaming and batching are the most common.

Streamed data gets moved from A to B in near real-time. It is a form of reactive programming, and the data stream gets triggered upon a specific user event. When a user posts on Twitter, the tweet is

a part of a data stream that gets submitted to the user profile and gets moved immediately to a sort of "global access" data viewing area so all users can see the Tweet. When Twitter runs a fact-check on President Trump, it is processing the tweet as a piece of data through a stream, combined with a micro-service, to offer the analysis.

Batch processing is good for processing high volumes of data. Its endpoints can wait a day, a week, a month for the information, so the data can get moved at scheduled times. Examples of data that gets batched might be end-of-quarter reports or marketing data—data that isn't needed immediately. Data that is used for analysis, where the analysts wish to do a one-time, or infrequent, detailed report on something can be batched.

Data transformations

Data in the pipeline doesn't have to be transformed. But, if transformations do occur, they'll be part of the data pipeline. Data transformations can be of many, many kinds. Data transformations might convert Word documents and PDF file formats submitted by a user to raw text documents for uniform storage in a data lake.

Transformed data could be something as simple as changing the data type from an integer value to a string value. It could be something more complex, where picture data gets classified as Emotional Indicator for marketing, Striking Visuals for the video content team, and Contains a Plant for the image classification team. The picture can get mixed and mangled, chopped and distorted, partitioned so it arrives at each party the way they wish to receive it.

Destinations

Whether the data is streamed or batched, and how it is transformed, depends on where that data is heading. The medium is the message, and the destination is the medium in which that data is presented.

Whether the data is used on a personal device to view the stats of a ball game, processed for facial recognition, compiled for a quarterly report, or prepared to train a machine learning model, data is passed around and arrives at a destination, presented in a format appealing to a reader. And that appealing format is often helped by data visualization techniques.

Data pipeline use cases

Not every business needs to do it and not every application requires a pipeline. Data pipelines are feature specific. The kinds of features that may require a data pipeline are ones that:

1. Store large amounts of data.
2. Acquire data from multiple sources.
3. Store data in the cloud.
4. Require quick access for analysis or reporting.

Each major cloud provider has its own tools to help build a data pipeline.

Data versioning

Data versioning is an important part of the Data Pipeline. The CI/CD DevOps workflow needs to rollback its versions occasionally, like when a new one fails and the old one proves to work. The same concept appears with organization's data.

Other KPIs of the data pipeline are:

1. **Versioning.** Keep a version history of the data.
2. **Latency.** Length of time it takes to pass data from point A to point B.
3. **Scalability.** The pipeline's ability to handle small or large amounts of data flow.
4. **Querying.** The ability to query the data from sources for analysis.
5. **Monitoring.** Check details throughout the data pipeline from event trigger, to transformation, to the final output.
6. **Testing.** Test the pipeline works.

Additional Resources

For more on this topic, browse the BMC Machine Learning & Big Data Blog or these articles:

- Simplifying and Scaling Data Pipelines in the Cloud
- Modern Batch: Managing the Madness?
- Basics of Graphing Streaming Big Data
- What is Stream Processing? Event Stream Processing Explained
- What is the "Intelligent Edge"?
- Cold vs Hot Data Storage
- Structured vs Unstructured Data in 2020: A Shift in Privacy