Overview of data pipeline stages and benefits

So far, you've learned that the data pipeline is a series of processes that transports data from different sources to their final destination for storage and analysis. The data pipeline is a critical element in the data journey, ensuring that data goes through the same process every time. You can think of a data pipeline like an assembly line; it ensures consistent data processing from extraction to storage. In this reading, you'll learn more about how the data pipeline is used to address business needs.

The data pipeline process

The purpose of a data pipeline is to automate data collection, transformation, and storage for businesses. The data pipeline begins when data is collected in the collection phase of the data journey. The data pipeline moves collected data to where it's stored for future analysis in the storage phase. There are three processes of a data pipeline, including extract, transform, and load. These data pipeline processes are also known as the E-T-L process.

Not all data pipelines are identical. Data pipelines can vary based on specific tasks, and the nature of the data. The purpose of the process is the same: to automate the process of extracting, transforming, and loading data. Data pipelines can be customized to specific business needs the same way you would customize an assembly line based on the product being manufactured.

Extract

In the extract stage of the ETL process, data is collected from different sources. ETL tools like Data Fusion, Data Prep, Talend, and Apache NiFi perform data collection. Analysts can also connect to databases, data warehouses, or other storage systems to pull out the required data. Sometimes, analysts write a custom script in Python or Java to extract data from specific sources. For example, a business might send a survey to be completed by email or through their app. Then, ETL tools can extract this survey data from a CSV file of responses about how often customers use certain products.

Transform

Once the data is collected, the raw data is cleaned and converted into a standard format determined by the business. For example, the business may choose to convert date to number format (MM/YYYY) even though it was collected in text format (Month Name and Year). In the transformation stage, errors and duplicates are also addressed. In a CSV file of customer survey responses, a duplicate is flagged when the same user information is connected to two surveys - one from the app, and one from an email. Duplicate responses must be handled so analysts know they are working with the most accurate information.

Load

The last step of the data pipeline involves loading the data into a designated storage solution. In the loading phase, analysts save the prepared data in an operational database, data lake, or data warehouse. Any time a decision is made about where data should be stored, the data's size, security requirements, and frequency of access must be considered.

Benefits of using a data pipeline

The primary function of a data pipeline is to move data from collection to storage. Data pipelines serve other useful functions including automation, data-driven decision making, and improved data quality.

Automation

A data pipeline allows large volumes of data to be handled in an automated and repeatable way, ensuring consistency. When working for a large organization, there won't be enough time to handle data manually. As a reminder, there are plenty of tools available like Data Fusion, DataPrep, Talend, and Apache NiFi to help handle data quickly, efficiently, and with built-in error detection.

Make data driven decisions

Businesses that implement a data pipeline have a steady supply of data to make more informed decisions. With readily available transformed data, analysts can get to work to address business needs as they arise. In the event that more data is needed, the data pipeline ensures there's already a process in place to collect it during the extraction process.

Improved data quality

Since every piece of data in the pipeline goes through the same process, there's a higher assurance of quality. While not all data is the same, businesses can rest assured that the processes in the data pipeline will standardize their data according to their needs.

Key takeaways

A data pipeline ensures consistent data processing. Data pipelines consist of three main stages: extract, transform, and load. While all data pipelines are unique, their purpose is to automate the extraction, transform, and load the data. Businesses benefit from data pipelines because they enhance data quality, facilitate automation, and enhance data-driven decision making.

Resources for more information

Dive deeper into how data pipelines are created and used for a variety of businesses with these resources:

- Click this link to learn more about different types of data pipelines and how businesses use them: Data Pipelines: An Overview
- Click this link to learn how to choose a data pipeline architecture that works for your needs: What Data Pipeline Architecture should I use?