

What is a Data Pipeline?

All / What is a Data Pipeline?

A data pipeline is a method in which raw data is ingested from various data sources, transformed, and then moved to a destination—such as a [data lake](#) or [data warehouse](#)—for analysis. It consists of a series of steps that are carried out in a specific order, with the output of one step acting as the input for the next step.

There are usually three key elements to any data pipeline: the source, the data processing steps and the destination, or “sink.” Data can be modified during the transfer process, and some pipelines may be used simply to transform data, with the source system and destination being the same.

In recent years, data pipelines have developed to cope with the [big data](#) demands of organizations, as large volumes and varieties of new data have become more common.

It is important that organizations take steps to ensure that pipelines experience no data loss, provide high accuracy and quality, and can scale with businesses’ varying needs. They should also be versatile enough to cope with structured, unstructured and semi-structured data.

Traditionally, data pipelines were deployed in on-premises data centers to handle the flow of data between on-prem systems, sources and tools. But, with data rapidly growing in volume and complexity, cloud data pipelines have emerged as one of the most scalable, flexible and agile types of data pipelines.