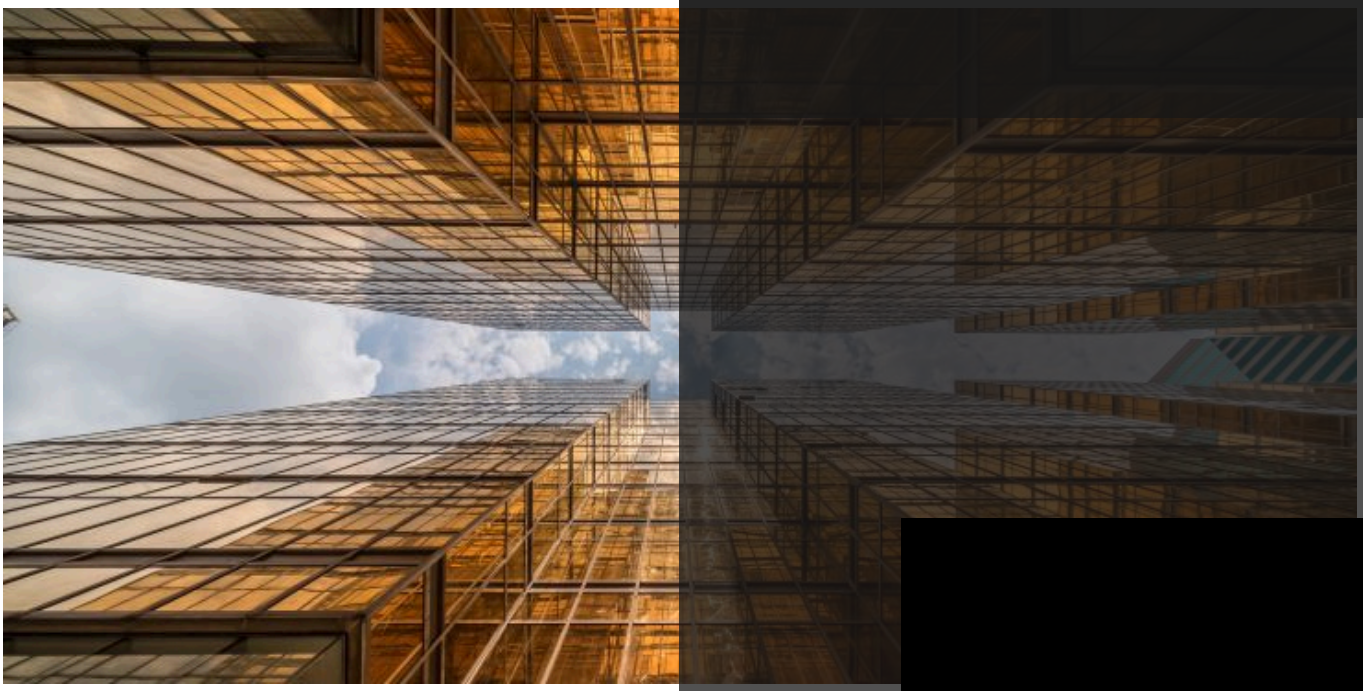
What is a data pipeline?



Analytics

14 June 2024

Authors

Cole Stryker
Editorial Lead, AI Models

What is a data pipeline?

A data pipeline is a method in which raw data is ingested from various data sources, transformed and then ported to a data store, such as a data lake or data warehouse, for analysis.

Before data flows into a data repository, it usually undergoes some data processing. This is inclusive of data transformations, such as filtering, masking, and aggregations, which ensure appropriate data integration and standardization. This is particularly important when the destination for the dataset is a relational database. This type of data repository has a defined schema which requires alignment—that is, mat types—to update existing data with new data.

As the name suggests, data pipelines act as the "piping" for business intelligence dashboards. Data can be sourced throu APIs, SQL and NoSQL databases, files, et cetera—but unfortu ready for immediate use. During sourcing, data lineage is trac relationship between enterprise data in various business and where data is currently and how it's stored in an environment data lake or in a data warehouse.

Data preparation tasks usually fall on the shoulders of data s who structure the data to meet the needs of the business use cases and handle huge amounts of data. The type of data processing that a data pipeline requires is usually determined through a mix of exploratory data analysis and defined business requirements. Once the data has been appropriately filtered, merged, and summarized, it can then be stored and surfaced for use. Well-organized data pipelines provide the foundation for a range of data projects; this can include exploratory data analyses, da visualizations, and machine learning tasks.

Types of data pipelines

There are several main types of data pipelines, each appropriate for specific tasks on specific platforms.

Batch processing

The development of batch processing was a critical step in b that were reliable and scalable. In 2004, MapReduce, a batc patented and then subsequently integrated into open-source CouchDB and MongoDB.

As the name implies, batch processing loads "batches" of da time intervals, which are typically scheduled during off-peak other workloads aren't impacted as batch processing jobs te volumes of data, which can tax the overall system. Batch pro data pipeline when there isn't an immediate need to analyze example, monthly accounting), and it is more associated with the ETL data integration process, which stands for "extract, transform, and load."

Batch processing jobs form a workflow of sequenced commands, where the output of one command becomes the input of the next command. For example, one command migh kick off data ingestion, the next command may trigger filtering of specific columns, a

the subsequent command may handle aggregation. This series of commands will continue until the data quality is completely transformed and rewritten into a data repository.

Streaming data

Unlike batching processing, streaming data pipelines—also known as event-driven architectures—continuously process events generated by various sources, such as sensors or user interactions within an application. Events are processed and analyzed, and then either stored in databases or sent downstream for further analysis.

Streaming data is leveraged when it is required for data to be continuously updated. For example, apps or point-of-sale systems need real-time data to update inventory and sales history of their products; that way, sellers can inform consumers if a product is in stock or not. A single action, such as a product sale, is considered an "event," and related events, such as adding an item to checkout, are typically grouped together as a "topic" or "stream." These events are then transported via messaging systems or message brokers, such as the open-source offering, Apache Kafka.

Since data events are processed shortly after occurring, streaming processing systems have lower latency than batch systems, but aren't considered as reliable as batch processing systems as messages can be unintentionally dropped or spend a long time in queue. Message brokers help to address this concern through acknowledgements, where a consumer confirms processing of the message to the broke queue.

Data integration pipelines

Data integration pipelines concentrate on merging data from unified view. These pipelines often involve extract, transform that clean, enrich, or otherwise modify raw data before storin repository such as a data warehouse or data lake. Data integ for handling disparate systems that generate incompatible fo example, a connection can be added to Amazon S3 (Amazon service that is offered by Amazon Web Services (AWS) that provides object storage through a web service interface.

Cloud-native data pipelines

A modern data platform includes a suite of cloud-first, cloud-native software products that enable the collection, cleansing, transformation and analysis of an organization's data to help improve decision making. Today's data pipelines have become increasingly complex and important for data analytics and making data-driven decisions. A modern data platform builds trust in this data by ingesting, storing, processing and transforming it in a way that ensures accurate and timely information, reduces data silos, enables self-service and improves data quality.

Data pipeline architecture

Three core steps make up the architecture of a data pipeline.

1. Data ingestion: Data is collected from various sources—including software-as-a-service (SaaS) platforms, internet-of-things (IoT) devices and mobile devices—and va
data structures, both structured and unstructured data. Within streaming data, these raw

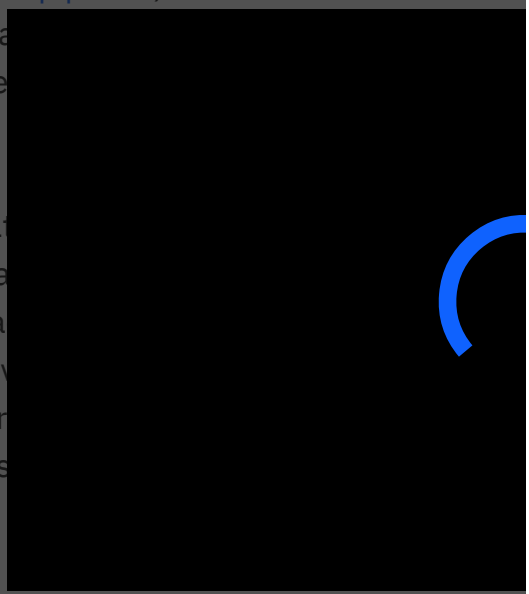
data sources are typically known as producers, publishers, or senders. While businesses can choose to extract data only when ready to process it, it's a better practice to land the raw data within a cloud data warehouse provider first. This way, the business can update any historical data if they need to make adjustments to data processing jobs. During this data ingestion process, various validations and checks can be performed to ensure the consistency and accuracy of data.

2. Data transformation: During this step, a series of jobs are executed to process data into the format required by the destination data repository. These jobs embed automation and governance for repetitive workstreams, such as business reporting, ensuring that data is cleansed and transformed consistently. For example, a data stream may come in a nested JSON format, and the data transformation stage will aim to unroll that JSON to extract the key fields for analysis.

3. Data storage: The transformed data is then stored within a data repository, where it can be exposed to various stakeholders. Within streaming data, this transformed data are typically known as consumers, subscribers, or recipients.

Data pipeline vs. ETL pipeline

You might find that some terms, such as data pipeline and ETL pipeline, are used interchangeably in conversation. However, you should think a subcategory of data pipelines. The two types of pipelines are features:



- ETL pipelines follow a specific sequence. As the abbreviat data, transform data, and then load and store data in a da pipelines need to follow this sequence. In fact, ELT (extra have become more popular with the advent of cloud-nati generated and stored across multiple sources and platfor occurs first with this type of pipeline, any transformations has been loaded into the cloud-based data warehouse.

- ETL pipelines also tend to imply the use of batch processing, but as noted above, the scope of data pipelines is broader. They can also be inclusive of stream processing.

- Finally, while unlikely, data pipelines as a whole do not necessarily need to undergo data transformations, as with ETL pipelines. It's rare to see a data pipeline that do utilize transformations to facilitate data analysis.

Use cases of data pipelines

As big data continues to grow, data management becomes an ever-increasing priority. While data pipelines serve various functions, the following are for business applications:

- **Exploratory data analysis**: Data scientists use exploratory data analysis (EDA) to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the needed answers, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis or check assumptions.

- **Data visualizations**: To represent data via common graphics, data visualizations such as charts, plots, infographics, and even animations can be created. These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand.

- **Machine learning**: A branch of artificial intelligence (AI) and computer science, machine learning focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. Through the use of statistical methods, algorithms are trained to make classifications or predictions, uncovering key insights within data mining projects.

- **Data observability**: To verify the accuracy and safety of the data being used, data observability applies a variety of tools for monitoring, tracking and alerting for both expected events and anomalies.

applications, analytics and AI workflows.

Read the report →

Resources

Report

Managing data for AI and analytics at scale

Learn how an open data lakehouse approach can provide trustworthy data and faster analytics and AI projects execution.

Read the report →

Report

2024 Gartner® Magic Quadrant™ for Data Integration Tools

IBM named a Leader for the 19th year in a row in the 2024 Gartner® Magic Quadrant™ for Data Integration Tools.

Read the report →

Guide

The data differentiator

Explore the data leader's guide to building a data-driven organization and driving business advantage.

Read the guide →

Report

Increase AI adoption with AI-ready data

Discover why AI-powered data intelligence and data integration are critical to drive structured and unstructured data preparedness and accelerate AI outcomes.

Read the report →

Ebook

The hybrid, open data lakehouse for AI

Simplify data access and automate data governance. Discover the power of integrating a data lakehouse strategy into your data architecture, including cost-optimizing your workloads and scaling AI and analytics, with all your data, anywhere.

Read the ebook →

Insights

IBM Research® data management publications

Explore how IBM Research is regularly integrated into new features for IBM Cloud Pak® for Data.

Explore articles →

Report

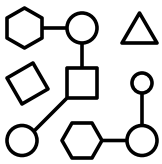
Gartner® predicts 2024: How AI will impact analytics users

Gain unique insights into the evolving landscape of ABI solutions, highlighting key findings, assumptions and recommendations for data and analytics leaders.

Read the report →

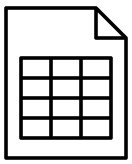
1 / 2

Related solutions



Data management software and solutions
Design a data strategy that eliminates data silos, reduces complexity and improves data quality for exceptional customer and employee experiences.

Explore data management solutions →



IBM watsonx.data
Watsonx.data enables you to scale analytics and AI with all your data, wherever it resides, through an open, hybrid and governed data store.

Discover watsonx.data →

Data and analytics consulting services
Unlock the value of enterprise data with IBM Consulting, building an insight-

driven organization that delivers business advantage.

Discover analytics services →

Take the next step

Design a data strategy that eliminates data silos, reduces complexity and improves data quality for exceptional customer and employee experiences.

Explore data management solutions →

Discover watsonx.data →