Data Pipeline & ETL

A data pipeline is a set of tools and processes for collecting, processing, and delivering data from one or more sources to a destination where it can be analyzed and used. A well-designed data pipeline is crucial for organizations to leverage their data effectively, support decision-making, and gain insights that drive business success.

Data Pipeline deals with information that is flowing from one end to another. In simple words, we can say collecting the data from various resources than processing it as per requirement and transferring it to the destination by following some sequential activities. It is a set of manner that first extracts data from various resources and transforms it to a destination means it processes it as well as moves it from one system to another system.

A diagram of a data pipeline

AI-generated content may be incorrect.

**Why Data Pipelines are important?**

Let's think about a scenario where a data pipeline is helpful.

The improvement of the cloud has meant that modern technology for enterprises uses lots of apps with different features. The retailing team might employ a combination of Hub spot and Market for trading automation. The other retailer teams mostly depend on Salesforce to handle and some might use [MongoDB](https://www.geeksforgeeks.org/mongodb/mongodb-an-introduction/) for storing customer approaches. This leads to the waste of data across different tools and results in data silos. Data silos are nothing but they will create it difficult to fetch even business insights, like your most profitable market. It is most important for [Business Intelligence(BI)](https://www.geeksforgeeks.org/power-bi/what-is-business-intelligence/) in their day-to-day life they require everyday information to work with.

**How to build a Data Pipeline?**

An organization can decide the methods of development to be followed just to abstract data from sources and transfer it to the destination. Batch transforming and processing are two common methods of development. Then there is a decision on what transformation process- [ELT(Extract/Load/Transform) or ETL](https://www.geeksforgeeks.org/dbms/etl-process-in-data-warehouse/) -to use before the data is moved to the required destination.

**Challenges to building Data Pipeline**

Netflix, has built its own data pipeline. However, building your own data pipeline is very difficult and time is taken.

**Here are some common challenges to creating a data pipeline in-house:**

* Connection
* Flexibility
* Centralization
* Latency

**Components of Data Pipeline :**

To know deep about how a data pipeline prepares large datasets for deconstruction, we have to know it is the main component of a common data pipeline. These are –

1. Source
2. Destination
3. Data flow
4. Processing
5. Workflow
6. Monitoring

**Data Pipeline Architecture**

Here is the representation of data Pipeline Architecture:

A diagram of a pipeline architecture

AI-generated content may be incorrect.

Data Pipeline Architecture

* **Ingestion Layer**: It retrieves data from an assortment of sources ranging from databases to APIs or even event streams.
* **Processing Layer:** Another operation on data involves the analysis of the said data followed by data cleaning through tools such as spark or Hadoop.
* **Storage Layer:** Held in data lakes, warehouses or other databases, data was kept for future reference or as a back up to be analyzed later.
* **Monitoring Layer:**It is the authority that is charged with the responsibility of providing quality data in the right time and increasing the efficiency of the system.
* **Consumption Layer**: Delivers the final data to BI tools or machine learning models where the data is analyzed at the next level for decision making.

**ETL (Extract Transform Load)**

Extract, transform, and load (ETL) is the process of combining data from multiple sources into a large, central repository called a data warehouse. ETL uses a set of business rules to clean and organize raw data and prepare it for [storage](https://aws.amazon.com/what-is/cloud-storage/), [data analytics](https://aws.amazon.com/what-is/data-analytics/), and [machine learning (ML)](https://aws.amazon.com/what-is/machine-learning/). You can address specific business intelligence needs through data analytics (such as predicting the outcome of business decisions, generating reports and dashboards, reducing operational inefficiency, and more).

Why is ETL important?

Organizations today have both structured and unstructured data from various sources including:

* Customer data from online payment and customer relationship management (CRM) systems
* Inventory and operations data from vendor systems
* Sensor data from Internet of Things (IoT) devices
* Marketing data from social media and customer feedback
* Employee data from internal human resources systems

By applying the process of extract, transform, and load (ETL), individual raw datasets can be prepared in a format and structure that is more consumable for analytics purposes, resulting in more meaningful insights. For example, online retailers can analyze data from points of sale to forecast demand and manage inventory. Marketing teams can integrate CRM data with customer feedback on social media to study consumer behavior.

**How does ETL work?**

Extract, transform, and load (ETL) works by moving data from the source system to the destination system at periodic intervals. The ETL process works in three steps:

1. Extract the relevant data from the source database
2. Transform the data so that it is better suited for analytics
3. Load the data into the target database

A diagram of a stage

AI-generated content may be incorrect.

**1. Extraction**

The Extract phase is the first step in the ETL process, where raw data is collected from various [data sources](https://www.geeksforgeeks.org/data-analysis/different-sources-of-data-for-data-analysis/). These sources can be diverse, ranging from structured sources like [databases](https://www.geeksforgeeks.org/dbms/what-is-database/) (SQL, NoSQL), to [semi-structured data](https://www.geeksforgeeks.org/dbms/what-is-semi-structured-data/) like JSON, XML, or [unstructured data](https://www.geeksforgeeks.org/dbms/what-is-unstructured-data/) such as emails or flat files. The main goal of extraction is to gather data without altering its format, enabling it to be further processed in the next stage.

Types of data sources can include:

* **Structured:** SQL databases, ERPs, CRMs
* **Semi-structured:** JSON, XML
* **Unstructured:** Emails, web pages, flat files

**2. Transformation**

The [Transform](https://www.geeksforgeeks.org/data-analysis/what-is-data-transformation/) phase is where the magic happens. [Data extracted](https://www.geeksforgeeks.org/data-analysis/what-is-data-extraction/) in the previous phase is often raw and inconsistent. During transformation, the [data is cleaned](https://www.geeksforgeeks.org/data-analysis/what-is-data-cleaning/), aggregated, and formatted according to business rules. This is a crucial step because it ensures that the data meets the quality standards required for accurate analysis.

Common transformations include:

* **Data Filtering:** Removing irrelevant or incorrect data.
* **Data Sorting:** Organizing data into a required order for easier analysis.
* **Data Aggregating:** Summarizing data to provide meaningful insights (e.g., averaging sales data).

The transformation stage can also involve more complex operations such as currency conversions, text [normalization](https://www.geeksforgeeks.org/dbms/introduction-of-database-normalization/), or applying domain-specific rules to ensure the data aligns with organizational needs.

**3. Loading**

Once data has been cleaned and transformed, it is ready for the final step: Loading. This phase involves transferring the transformed data into a [data warehouse](https://www.geeksforgeeks.org/dbms/data-warehousing/), [data lake](https://www.geeksforgeeks.org/data-engineering/what-is-data-lake/), or another target system for storage. Depending on the use case, there are two types of loading methods:

* **Full Load:** All data is loaded into the target system, often used during the initial population of the warehouse.
* **Incremental Load:** Only new or updated data is loaded, making this method more efficient for ongoing data updates.

**Importance of ETL**

* **Data Integration:** ETL combines data from various sources, including [structured](https://www.geeksforgeeks.org/dbms/what-is-structured-data/) and [unstructured formats](https://www.geeksforgeeks.org/dbms/what-is-unstructured-data/), ensuring seamless integration for a unified view.
* **Data Quality:** By transforming raw data, ETL cleanses and standardizes it, improving data accuracy and consistency for more reliable insights.
* **Essential for Data Warehousing:** ETL prepares data for storage in data warehouses, making it accessible for analysis and reporting by aligning it with the target system's requirements.
* **Enhanced Decision-Making:** ETL helps businesses derive actionable insights, enabling better forecasting, resource allocation, and strategic planning.
* **Operational Efficiency:** Automating the data pipeline through ETL speeds up data processing, allowing organizations to make real-time decisions based on the most current data.

**Challenges in ETL Process**

The ETL process, while essential for data integration, comes with its own set of challenges that can hinder efficiency and accuracy. These challenges, if not addressed properly, can impact the overall performance and reliability of data systems.

* **Data Quality Issues:** Inconsistent, incomplete, or duplicate data from multiple sources can impact transformation and loading, leading to inaccurate insights.
* **Performance Bottlenecks:** Large datasets can slow down or cause ETL processes to fail, particularly during complex transformations like cleansing and aggregation.
* **Scalability Issues:** Legacy ETL systems may struggle to scale with growing data volumes, diverse sources, and more complex transformations.