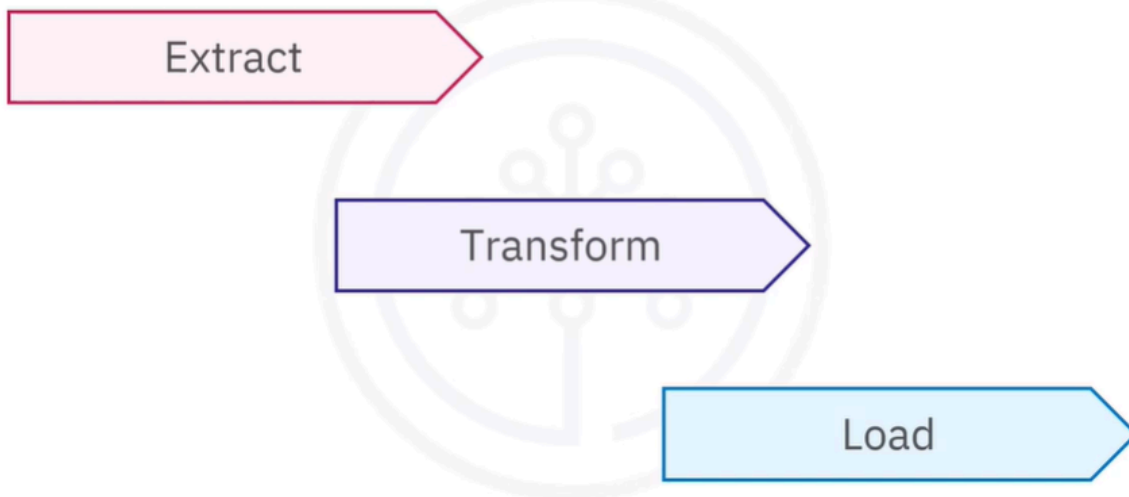


# ETL

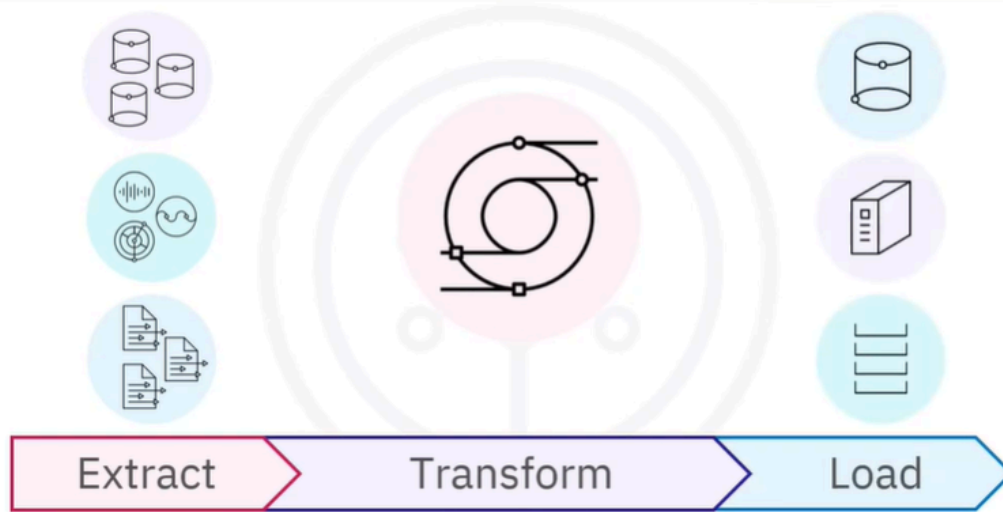
## What is an ETL process?

---



- What is an ETL process? ETL stands for Extract, Transform, and Load. ETL is an automated data pipeline engineering methodology, whereby data is acquired and prepared for subsequent use in an analytics environment, such as a data warehouse or data mart.

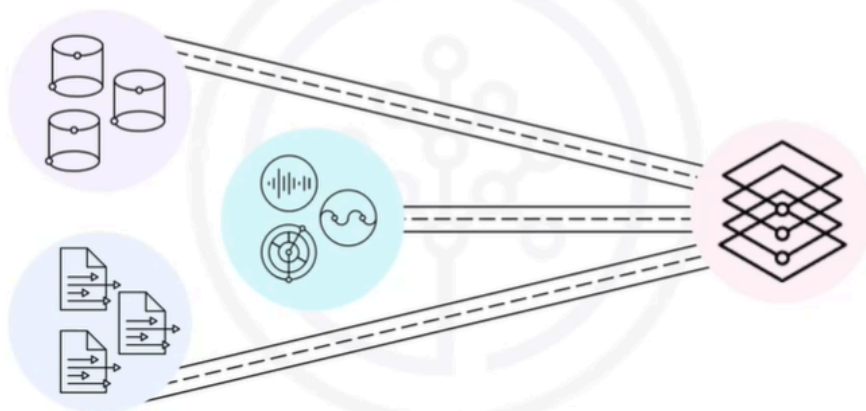
## What is an ETL process?



- ETL refers to the process of curating data from multiple sources, conforming it to a unified data format or structure, and then loading the transformed data into its new environment.

## What is an ETL process?

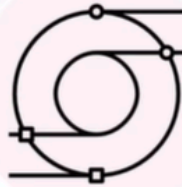
**E=Extract:** Extracting data from a source



## What is an ETL process?

---

**T=Transformation:**



Transforming data into the format for the output

## What is an ETL process?

---

**L=Load:** Loading data into a database, data warehouse or other storage

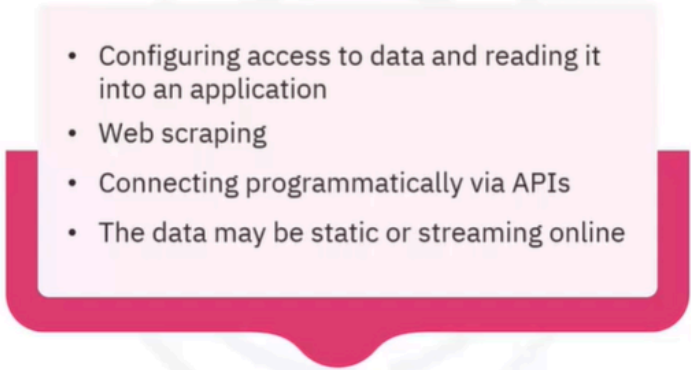


- The extraction process obtains or reads the data from one or more sources.
- The transformation process wrangles the data into a format that is suitable for its destination and its intended use.
- The final loading process takes the transformed data and loads it into its new environment, ready for visualization, exploration, further transformation, and

modeling. The curated data may also be utilized to support automation and decision-making.

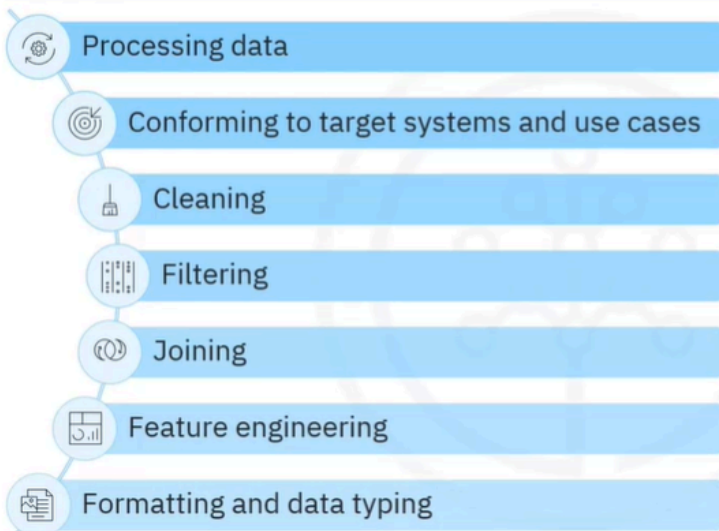
## What is extraction?

---

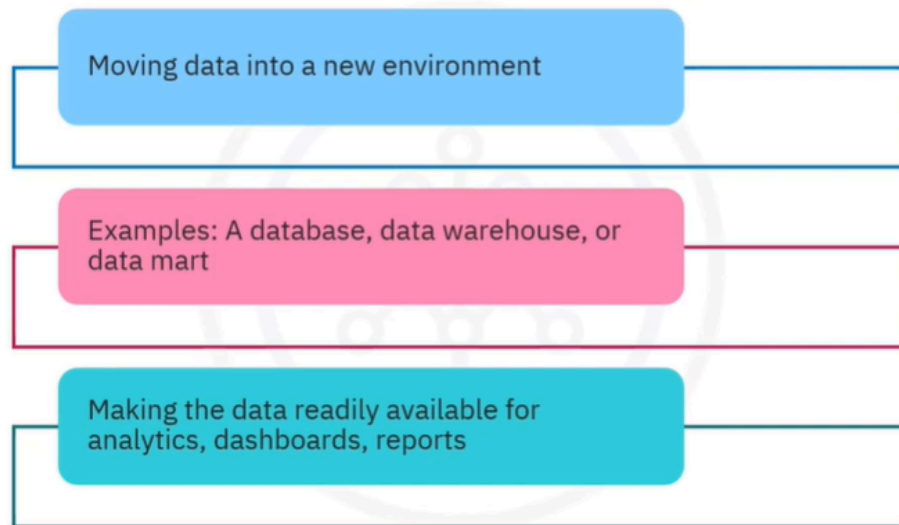
- 
- A diagram illustrating the extraction process. It features a light pink rectangular box with a dark pink border and a decorative drop shadow at the bottom. Inside the box, there is a bulleted list of four items. The background of the slide shows a faint circular graphic with arrows pointing inwards.
- Configuring access to data and reading it into an application
  - Web scraping
  - Connecting programmatically via APIs
  - The data may be static or streaming online

## What is data transformation?

---



# What is data loading?



Skills Network

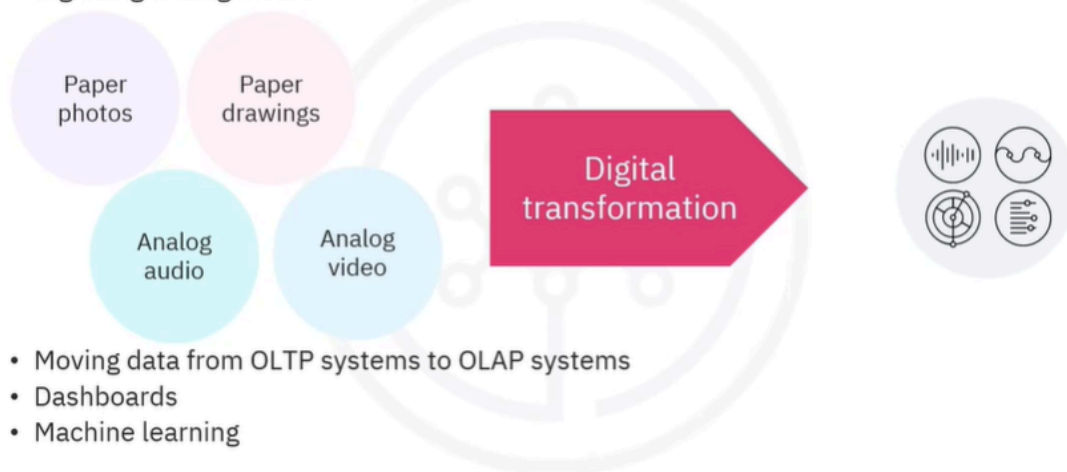
IBM

- What is extraction? To extract data is to configure access to it and read it into an application. Normally this is an automated process. Some common methods include: Web scraping, where data is extracted from web pages using applications such as Python or R to parse the underlying HTML code, and Using APIs to programmatically connect to data and query it. The source data may be relatively static, such as a data archive, in which case the extraction step would be a stage within a batch process. On the other hand, the data could be streaming live, and from many locations. Examples include weather station data, social networking feeds, and IoT devices.
- What is data transformation? Data transformation, also known as data wrangling, means processing data to make it conform to the requirements of both the target system and the intended use case for the curated data. Transformation can include any of the following kinds of processes: Cleaning: fixing errors or missing values. Filtering: selecting only what is needed. Joining disparate data sources: merging related data. Feature engineering: such as creating KPIs for dashboards or machine learning. Formatting and data typing: making the data compatible with its destination.
- What is data loading? Generally this just means writing data to some new destination environment. Typical destinations include databases, data

warehouses, and data marts. The key goal of data loading is to make the data readily available for ingestion by analytics applications so that end users can gain value from it. Applications include dashboards, reports, and advanced analytics such as forecasting and classification.

## Use cases for ETL pipelines

- Digitizing analog media



- Moving data from OLTP systems to OLAP systems
- Dashboards
- Machine learning

- There are many use cases for ETL pipelines. A very large amount of information is either already recorded or being generated, but is not yet captured, or accessible, as a digital file. Examples include paper documents, photos and illustrations, and analog audio and video tapes. Digitizing analog data includes extraction by some form of scanning, analog-to-digital transformation, and, finally, storage into a repository. Online transaction processing (OLTP) systems don't save historical data. Accordingly, ETL processes capture the transaction history and prepare it for subsequent analysis in an online analytical processing (OLAP) system. Other use cases include engineering 'features', or KPIs, from data sources, as preparation for ingestion by dashboards used by operations, sales and marketing, customers, and executives. Training and deploying machine learning models for prediction and augmented decision-making.

## Recap

---

In this video, you learned that:

- ETL stands for Extract, Transform, and Load
- Extraction means reading data from one or more sources
- Transformation means wrangling data to meet destination requirements
- Loading means writing the data to its destination environment
- ETL is used for curating data and making it accessible to end users