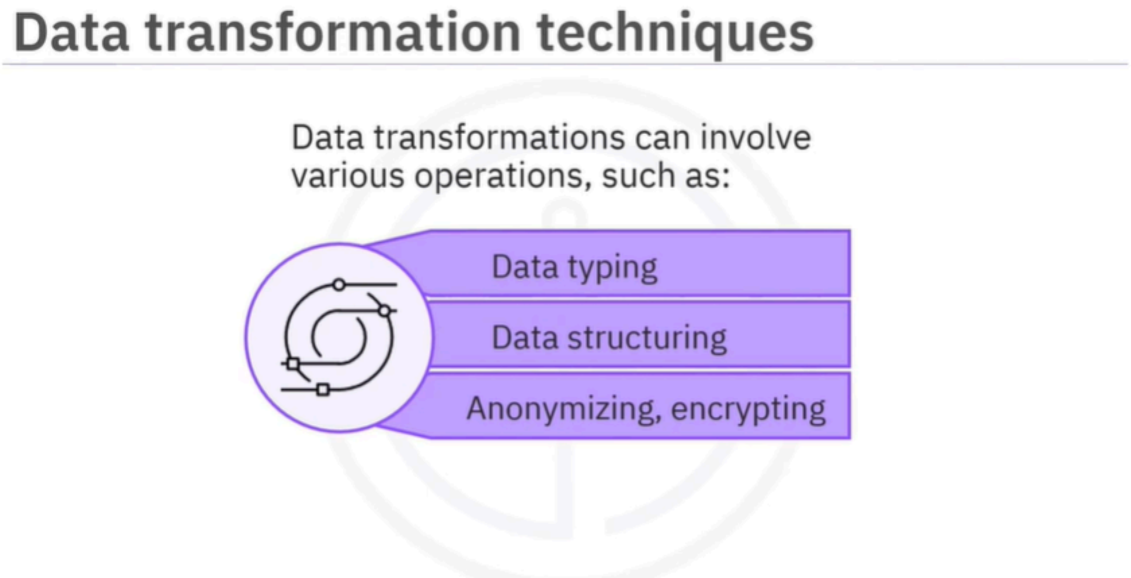
Data Transformation Techniques

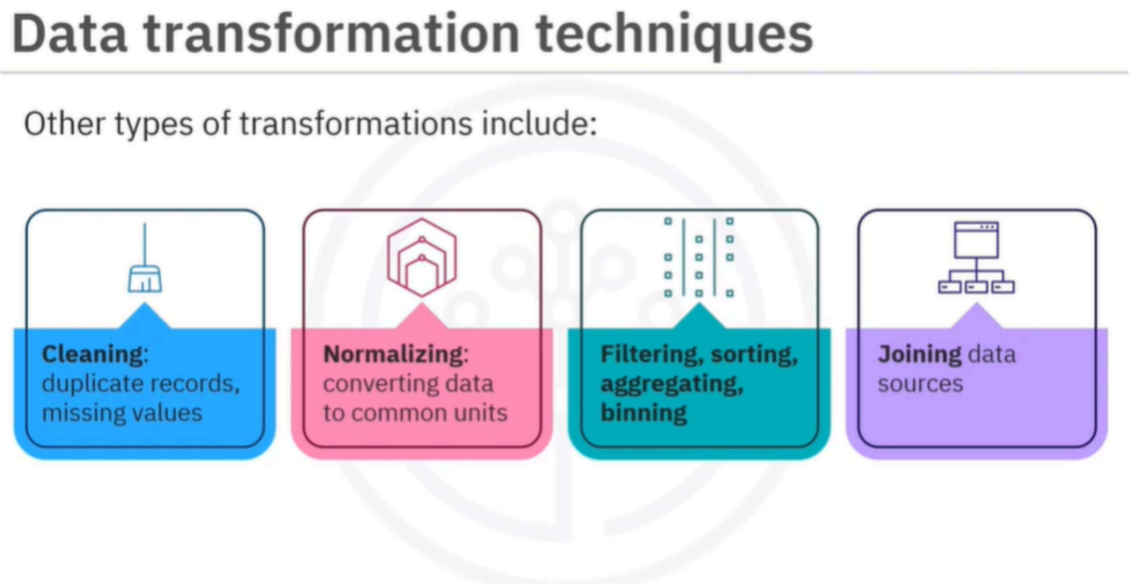


Data transformation techniques

Data transformations can involve various operations, such as:

- Data typing
- Data structuring
- Anonymizing, encrypting

- Data transformation is mainly about formatting the data to suit the application. This can involve many kinds of operations, such as data typing, which involves casting data to appropriate types, such as integer, float, string, object, and category, data structuring, which includes converting one data format to another, such as JSON, XML, or CSV to database tables, as well as anonymizing and encrypting transformations to help ensure privacy and security.



Data transformation techniques

Other types of transformations include:

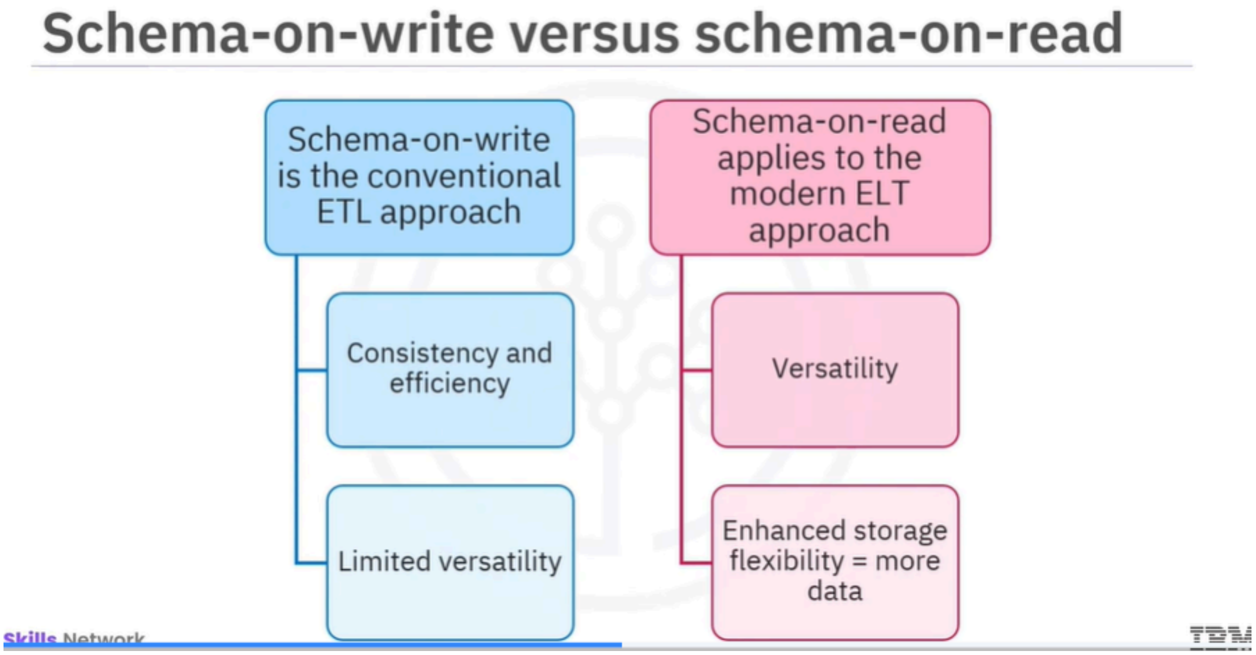
Cleaning: duplicate records, missing values

Normalizing: converting data to common units

Filtering, sorting, aggregating, binning

Joining data sources

- Other types of transformations include cleaning operations for removing duplicate records, and filling missing values, normalizing data to ensure units are comparable, for example, using a common currency, filtering, sorting, aggregating, and binning operations for accessing the right data at a suitable level of detail and in a sensible order, and joining or merging disparate data sources.



Schema-on-write versus schema-on-read

Schema-on-write is the conventional ETL approach

Consistency and efficiency

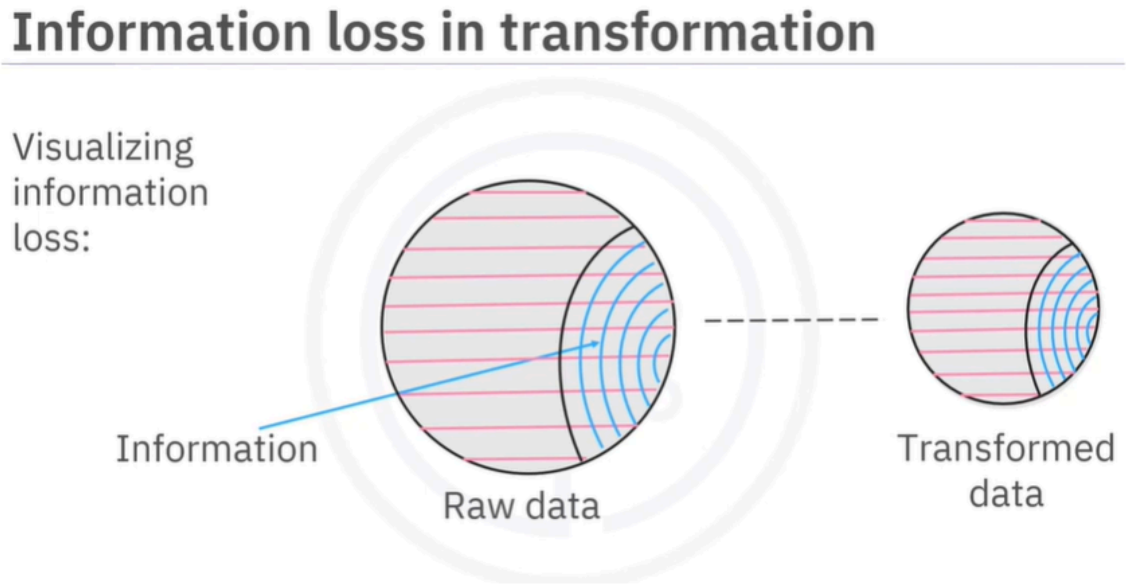
Limited versatility

Schema-on-read applies to the modern ELT approach

Versatility

Enhanced storage flexibility = more data

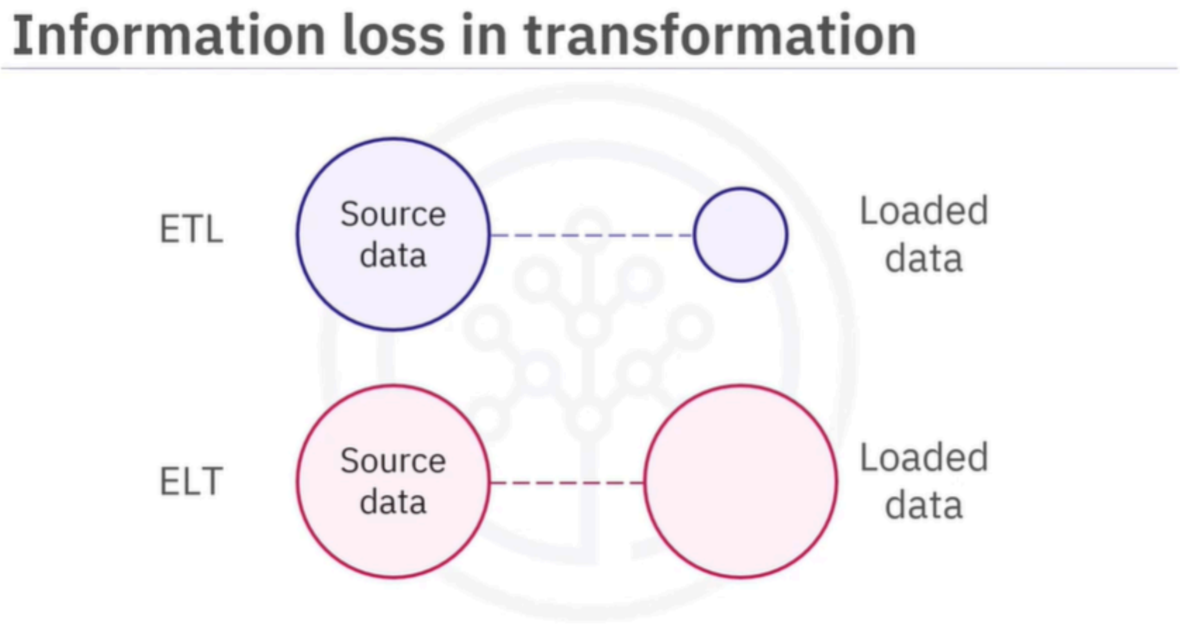
- Schema-on-write is the conventional approach used in ETL pipelines, where the data must be conformed to a defined schema prior to loading to a destination, such as a relational database. The idea is to have the data consistently structured for stability and for making subsequent queries much faster. But this comes at the cost of limiting the versatility of the data.
- Schema-on-read relates to the modern ELT approach, where the schema is applied to the raw data after reading it from the raw data storage. This approach is versatile since it can obtain multiple views of the same source data using ad hoc schemas. Users potentially have access to more data, since it does not need to go through a rigorous preprocessing step.



Information loss in transformation

Visualizing information loss:

- Whether intentional or accidental, there are many ways in which information can be lost in transformation. We can visualize this loss as follows: Raw data is normally much bigger than transformed data. Since data usually contains noise and redundancy, we can illustrate the information content of data as a proper subset of the data. Correspondingly, we can see that shrinking the data volume can also mean shrinking the information content.



Information loss in transformation

ETL

Source data

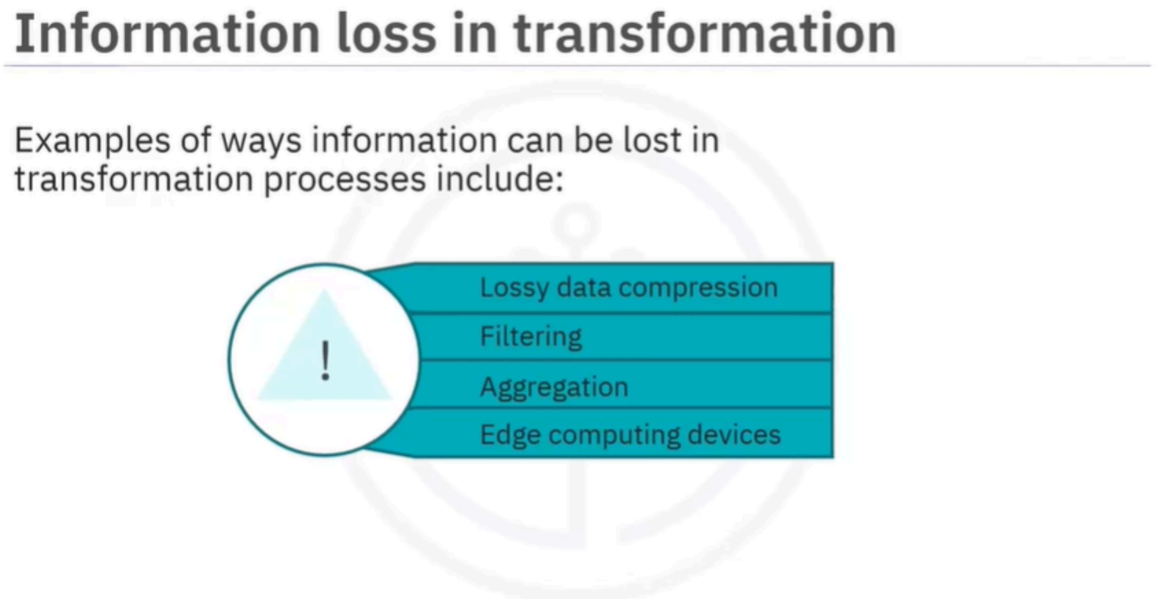
Loaded data

ELT

Source data

Loaded data

- Either way, for ETL processes, any lost information may or may not be recoverable, whereas with ELT, all the original information content is left intact because the data is simply copied over as is.

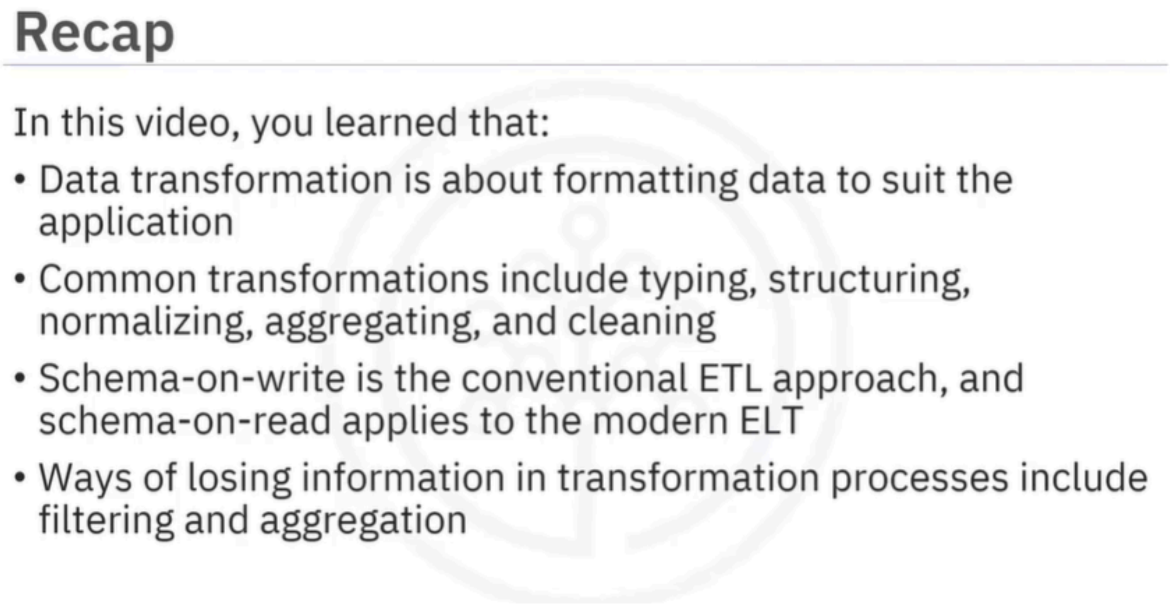


Information loss in transformation

Examples of ways information can be lost in transformation processes include:

- Lossy data compression
- Filtering
- Aggregation
- Edge computing devices

- Examples of ways information can be lost in transformation processes include lossy data compression, for example, converting floating point values to integers, reducing bit rates on audio or video. Filtering, for example, filtering is usually a temporary selection of a subset of data, but when it is permanent, information can easily be discarded. Aggregation, for example, average yearly sales versus daily or monthly average sales, and edge computing devices, for example, false negatives in surveillance devices designed to only stream alarm signals, not the raw data



Recap

In this video, you learned that:

- Data transformation is about formatting data to suit the application
- Common transformations include typing, structuring, normalizing, aggregating, and cleaning
- Schema-on-write is the conventional ETL approach, and schema-on-read applies to the modern ELT
- Ways of losing information in transformation processes include filtering and aggregation