Data transformation

- Every business generates **a good amount of data** daily, but the same is **not useful** until it is transformed into a useful format. To get benefitted from raw data, its transformation is necessary. With data transformation, you can make different pieces of **data compatible** with one another, **move them** to another system, and **join with other data** to drive useful business insights.

- Data transformation is a process of converting **raw data into a single and easy-to-read format** to facilitate **easy analysis**. To turn your data into something meaningful, you must have the right **data transformation tool** by your side.

- Data transformation is also known as **ETL (Extract, Transform, Load)**

- As per ETL, the data is first extracted from multiple sources, transformed into a required format, and then loaded into a data warehouse for powering analysis and reporting processes.

- Data transformation may be constructive (adding, copying, and replicating data), destructive (deleting fields and records), aesthetic (standardizing salutations or street names), or structural (renaming, moving, and combining columns in a database).

- An enterprise can choose among a variety of ETL tools that automate the process of data transformation. Data analysts, data engineers, and data scientists also transform data using scripting languages such as Python or domain-specific languages like SQL.

How to transform data?

1. Extraction and parsing: [Data aggregation](#) starts with **extracting the data from multiple source systems** and copying the same to its destination. The transformation process starts with structuring the data into a single format, so it becomes compatible with the system in which it is copied and the other data available in it. Parsing is a process of analyzing data structures and confirming the same with the rules of grammar.

2. Translation and mapping: Translation and mapping are part of the basic steps of data transformation. Data translation is a process of converting big amounts of data from one format to a preferred one when it is transferred from one system to another. At the same time, data mapping is all about finding matching fields between two distinct data models.

3. Filtering, aggregation, and summarization: Data combined from different sources may bring unnecessary columns, fields, and records with them. Irrelevant data can be omitted from the extraction process by using data filtering.

4. Enrichment and imputation: Data from diverse sources can be merged to create enriched information. For example, merging the customers' transactions with their information table can make the process of customer analysis more efficient. The long fields can be split into multiple columns to fill the missing values, or corrupted values can be removed for enriching the available data. This will boost the process of data analysis and provide you relevant and accurate business insights.

5. Indexing and ordering: Data must be transformed to become logical and comply with the data storage scheme. You can create indexes to optimize the performance of a database. It will also help you to locate and access the required data in a database quickly.

6. Anonymization and encryption: Data anonymization refers to any piece of data that cannot be reversibly transformed. It is done to protect the identification of a particular set of information or individual. Now, the level of competition among organizations has become tough and calls for the encryption of private data. You can encrypt data at multiple levels, ranging from individual databases to entire records.

7. Modeling, typecasting, formatting, and renaming: A whole bunch of transformations that help you reshape your data into the desired format without changing the content. It makes your data compatible by casting and converting data types, renaming columns, tables, and schemas for better clarity, and adjusting times and dates with format localization.

8. Refining the data transformation process: Before transforming the data, it's important you replicate it to a data warehouse built for analytics. If you want to make the most out of your ELT solution, it's better to opt for a cloud data warehouse.

Benefits of data transformation

- Data is transformed to make it better-organized. Transformed data may be easier for both humans and computers to use.

- Properly formatted and validated data improves data quality and protects applications from potential landmines such as null values, unexpected duplicates, incorrect indexing, and incompatible formats.

- Data transformation facilitates compatibility between applications, systems, and types of data. Data used for multiple purposes may need to be transformed in different ways.

Challenges in Data Transformation

- Slow: The extraction and transformation of large volumes of data are difficult to be processed in one go and can become a burden on your system. Therefore, the same is carried in batches, which means that the next batch has to wait for hours until the first one is entirely transformed. This thing can delay the making of crucial business decisions and result in missing growth opportunities.

- Time-consuming: Cleansing of unstructured data can take a lot of time before it becomes ready for a transformation. This is one of the biggest complaints of data scientists or analysts working with unstructured data.

- Expensive: The size of your infrastructure will impact your data transformation requirements. With a bigger infrastructure, you will require a team of data experts to manage the data, resulting in more expense.