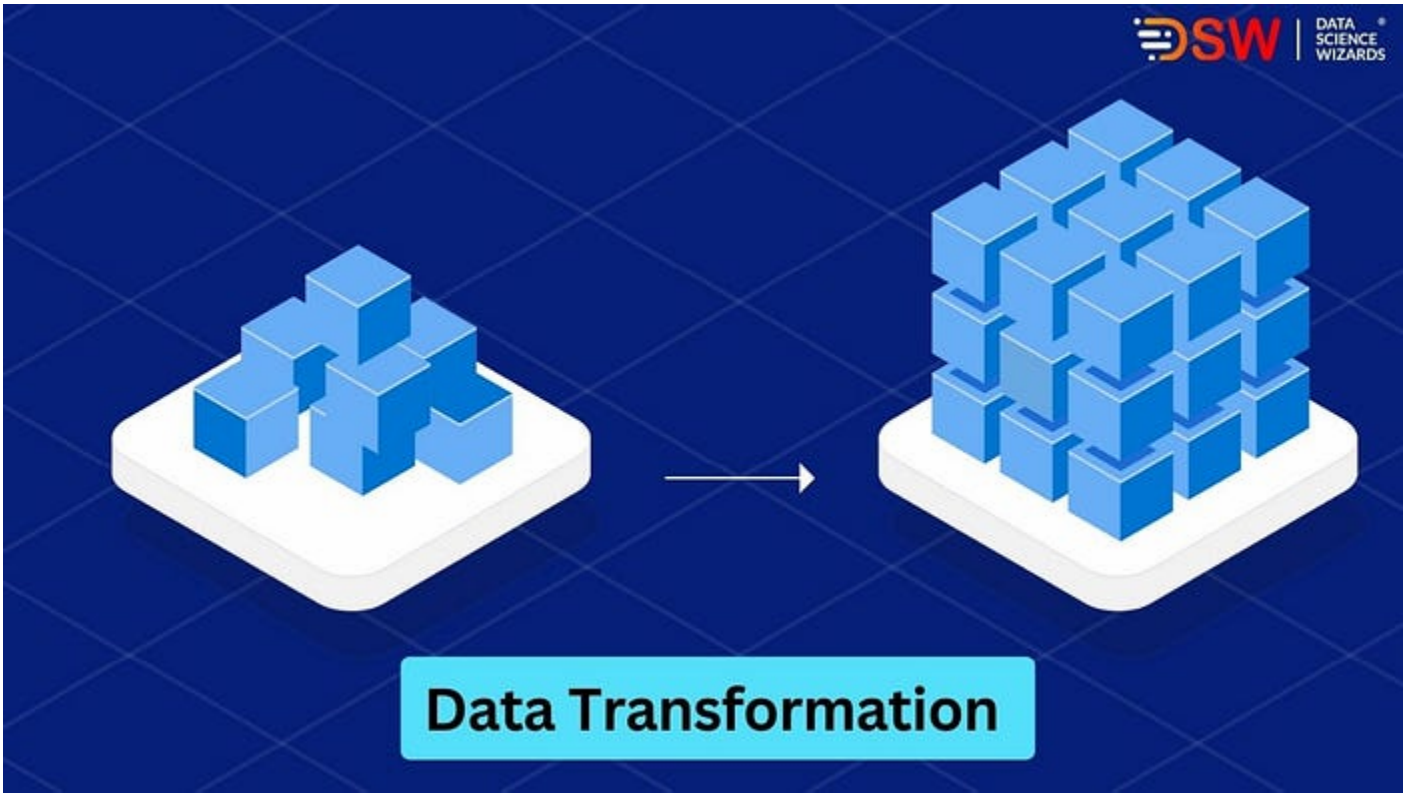
Introduction to Data Transformation

Data Science Wizards
.
Follow
8 min read
.
Mar 31, 2023



When applying data science, machine learning and artificial intelligence to different use cases, one should always take care of one fact raw data is difficult to understand and trace. Here, the need for data processing comes in front of us so that critical, accurate and valuable information can be retrieved. Data transformation is one of the techniques that we use in between data processing. This technique lets us convert the raw data into a required format so that the next procedures of data processing and data modelling can be performed efficiently.

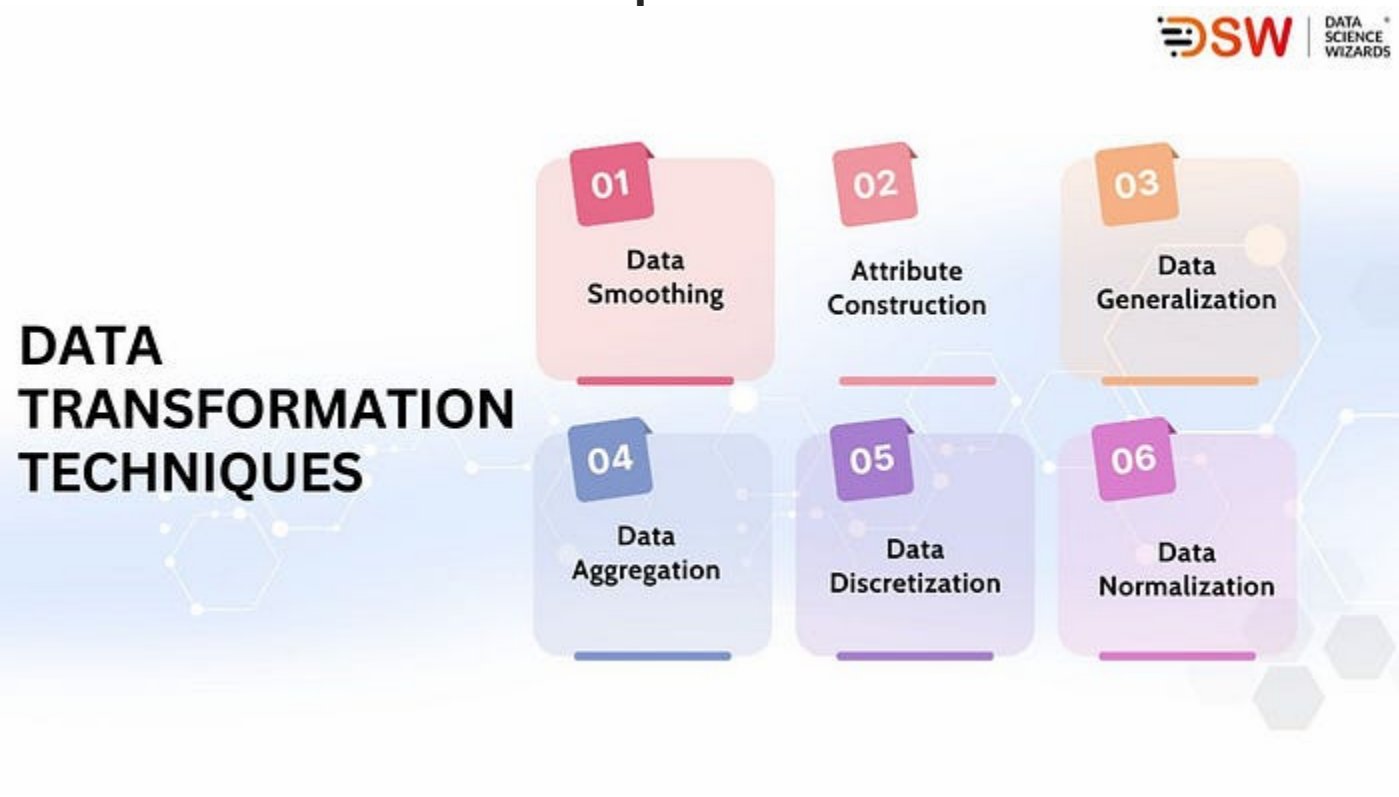
Technically. Data transformation changes the data structure, format and value and makes it clean and usable for the next processes. There can be two stages of data transformation processes because many organisations use data warehouses arranged in the ETL process, where data transformation is an in-between process. On the other hand, nowadays, many organisations rely on cloud-based data warehouses, which makes them capable of loading raw data and transforming the data in query time.

However, data transformation is an integral part of the entire data processing pipeline because it can be found right from data integration to the final stages of data wrangling. In a general sense, we can find the following type of data transformation:

- **Constructive:** The process of adding, copying or replicating data.
- **Destructive:** deleting the records or field.

- **Aesthetic:** standardising the data to make it valuable.
- **Structural**: reorganising the data by moving, merging and renaming the columns.

Data Transformation Techniques



There are various types of data transformation techniques that we use before applying data to any process or storing it in a data warehouse. Let's take a look at various general data transformation techniques:

1. **Data smoothing**

We can think of data smoothing as the process of removing noise from the data that can also involve some algorithms. This transformation makes the important data feature more visible and helps predict the pattern.

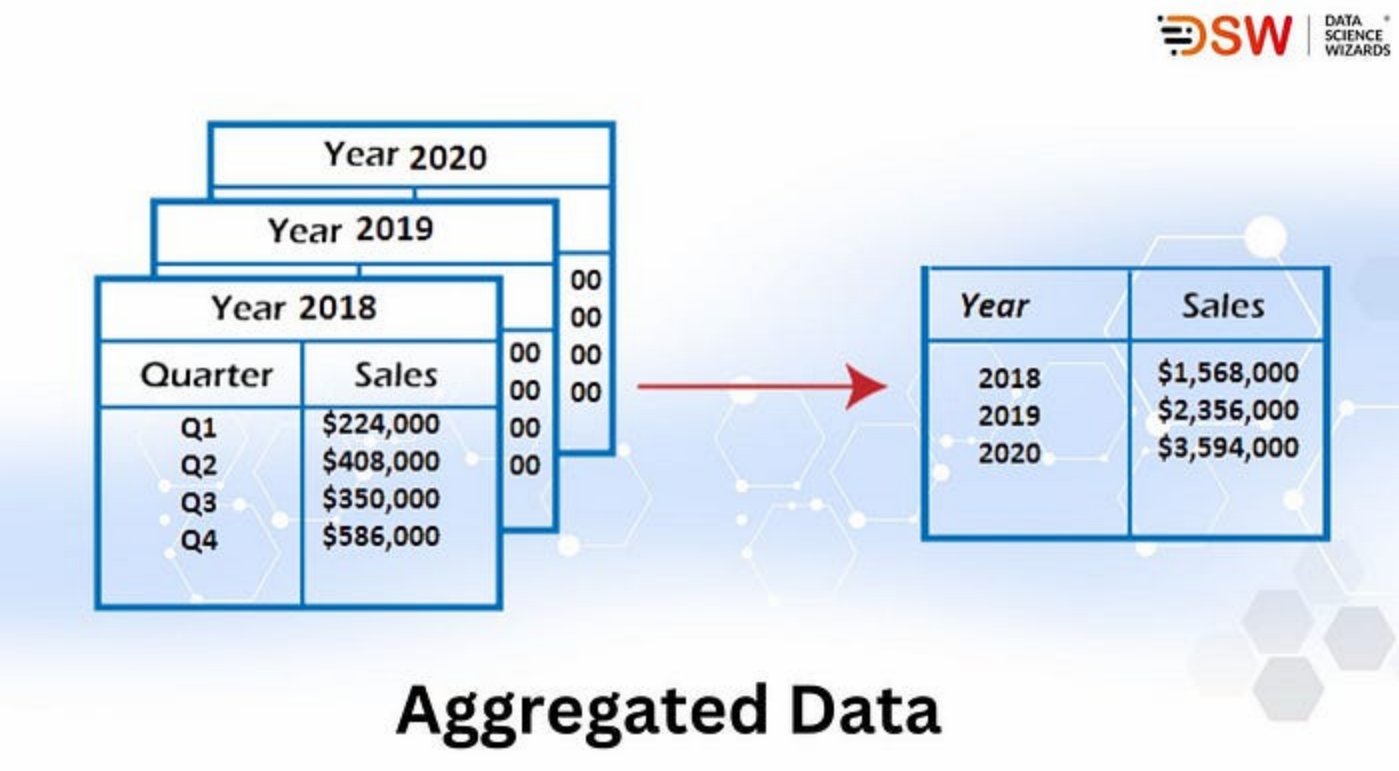
The concept behind this data transformation technique is to identify the simple changes and predict simple changes in trends and patterns of data. Various analyst performs this to make patterns finding process easy. There are various techniques, such as binning, regression and clustering, that help in removing the noise from the data.

2. Attribute construction

In this data transformation technique, we add new attributes to the data based on the already existing attributes. This new attribute smoothens the other processes by simplifying the way to get accurate data.

Suppose any machine learning model is trained using the volume data, and in a database, we have height, width and length attributes. So by combining these three attributes, we can make the volume attributes which not only simplify the data ingestion process but also give us an attribute using which we can understand the relations among the other attributes in the data.

3. Data aggregation



Data aggregation or data collection is a technique of presenting data in a summary form. There is a huge chance of data coming from different sources, and integrating all the coming data into a description is data aggregation. This part of data processing is crucial because it depends on the quality and quantity of the data we use.

An example of this process is making an annual sales report based on the given quarterly or monthly data.

5. Data normalisation

Data normalisation is another technique of data transformation that involves scaling the data in a smaller range, such as scaling data

between 0 to 1 or -1 to 1. The main objective of data normalisation is to eliminate data redundancy and improve data consistency and accuracy. By organising data into smaller, more manageable ranges, it becomes easier to update and maintain the data. Normalisation also helps to reduce the likelihood of data inconsistencies, which can arise when the same information is stored in multiple locations.

There are various types of data normalisation techniques we use, such as Min-Max normalisation, Z-score normalisation, Decimal Scaling etc., to make the article precise. We are not explaining the terms here.

6. Data Discretization

We can think of data discretisation as the process of converting continuous data into a set of intervals so that small interval labels can substitute them. Just like normalisation, this technique makes the data more interpretable. If a data analysis work uses continuous values, then discrete forms of values can be replaced by constant quality attributes.

We can also call this technique the data reduction technique, as it transforms discrete data into a set of categorical data. We can classify this technique into supervised and unsupervised discretisation. Where supervised data discretisation uses class information, on the other hand, unsupervised data discretisation uses the processing direction of the involved data process.

For example, we can classify people in numeric age groups such as(0–5,5–10…) or in classes like kid, youth or adult.

7. Data Generalisation

This data transformation technique depends on the concept of hierarchy to transform low-level data attributes into high-level data attributes. Through this transformation, we get a clear picture of the data, and this process can be divided into two different approaches:

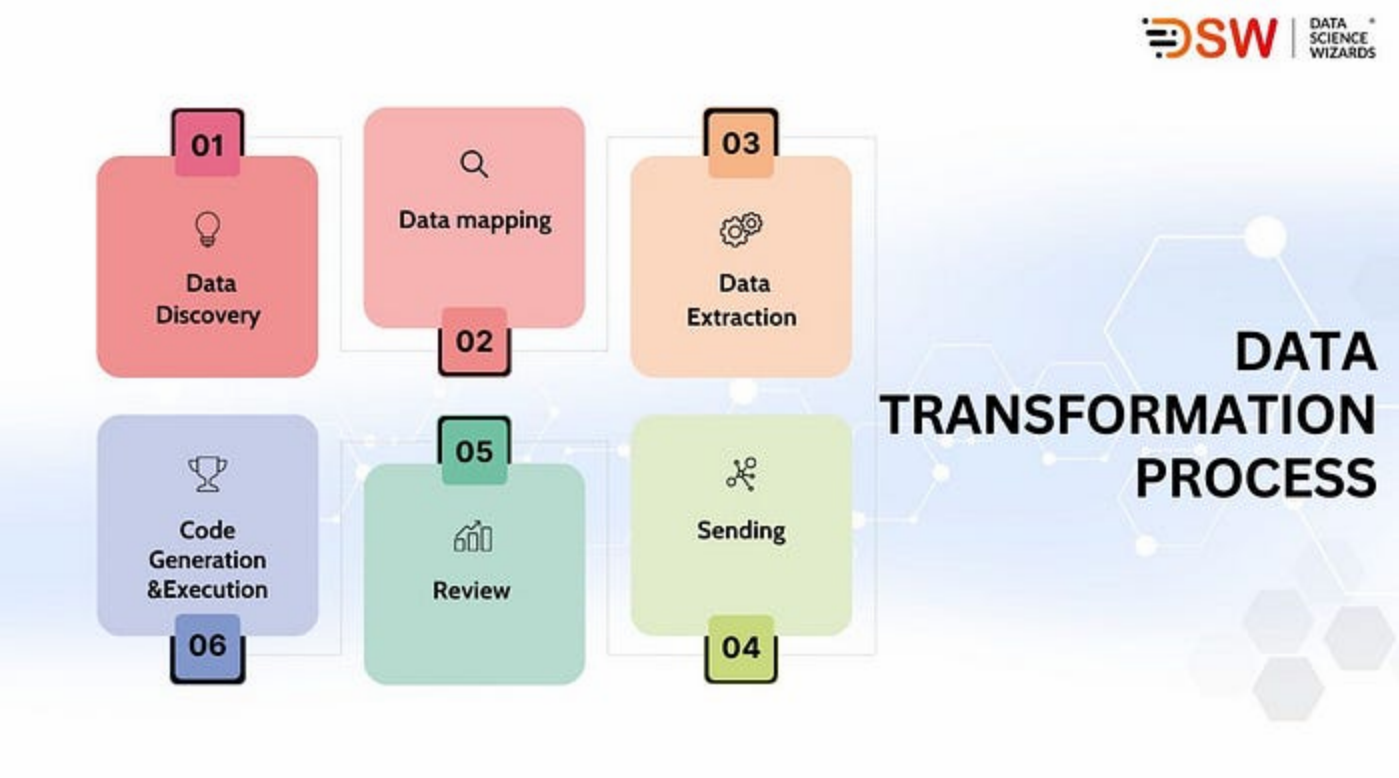
1. Data cube process (OLAP) approach.

2. Attribute-oriented induction (AOI) approach.

A simple example of data generalisation is to convert age data from numerical to categories.

Now that we know about the general data transformation techniques, it becomes important to know how they can be performed. Here the data transformation process comes into mind.

Data transformation Process



Talking about the data transformation process, we can say it comes under the ETL process, where ETL is the abbreviation of Extract, Transform, and Load. by going through the process, data analysts can transform data into the required format. But the general steps he needs to take care of are mentioned below:

1. **Data Discovery**

Before applying ETL in any organisation, it is important to know and understand the data source. Usually, data profiling tools we use to do so. With the help of this step, one gets to know what ahead is required to get data in the required format.

2. Data Mapping

This step of the process is performed to determine how any field is mapped, modified, filtered, joined or aggregated.

3. Data Extraction

This step involves the extraction of data from its original source. Examples of sources can be databases or streaming sources such as sales log files from web applications.

4. Code Execution

The foremost step that involves actual data transformation is code generation and execution. Here code is generated to transform data in the required format.

5. Review

This step ensures that the code executed on data is transforming the data accurately.

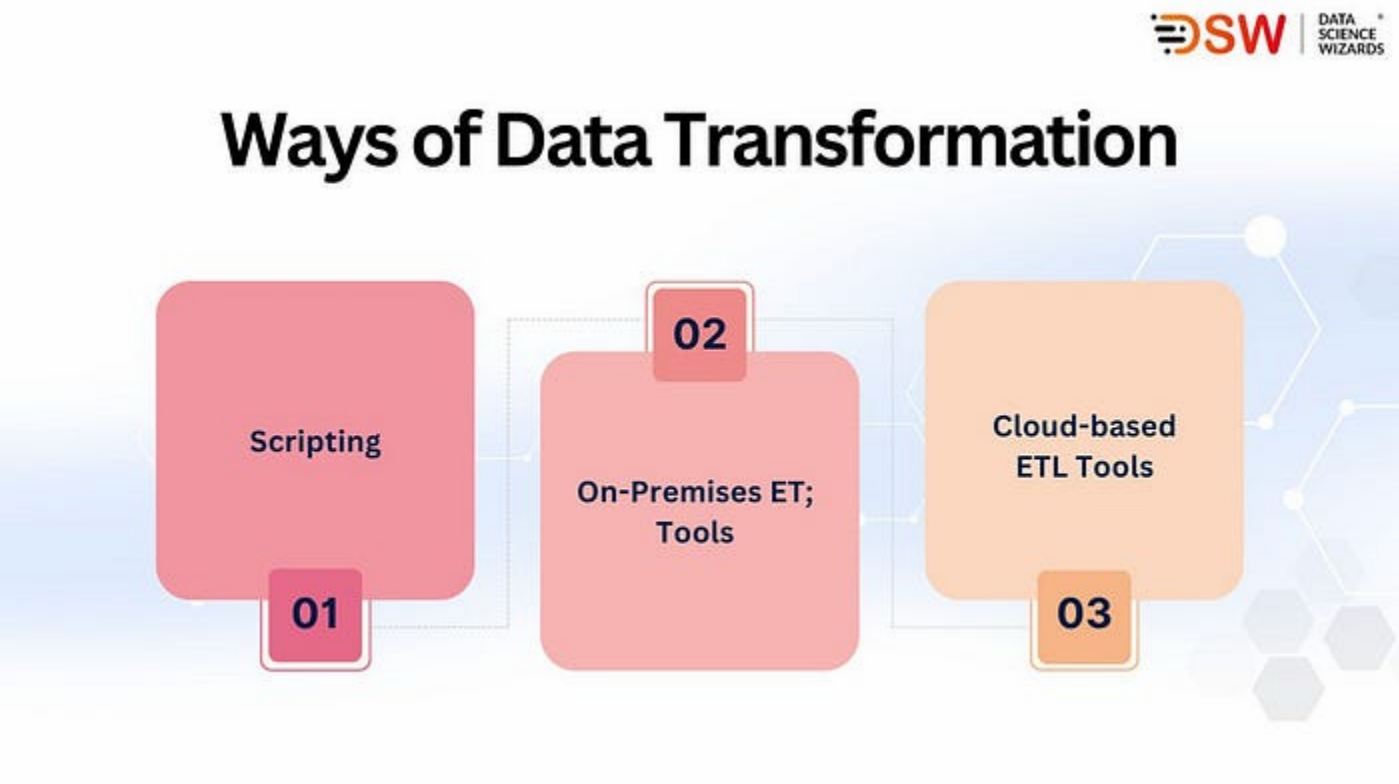
6. Sending

Till here, the data is transformed, but the last thing which remains to complete is to send the data to its target destination. Here the

destination can be a relational database or warehouse that handles both unstructured and structured data.

Here are the various processes that we follow while performing the data transformation. There can be many ways to transform data, but when we go deeper, we can categorise these ways into three categories.

Ways of data transformation



Three general ways of data transformation are as follows**.**

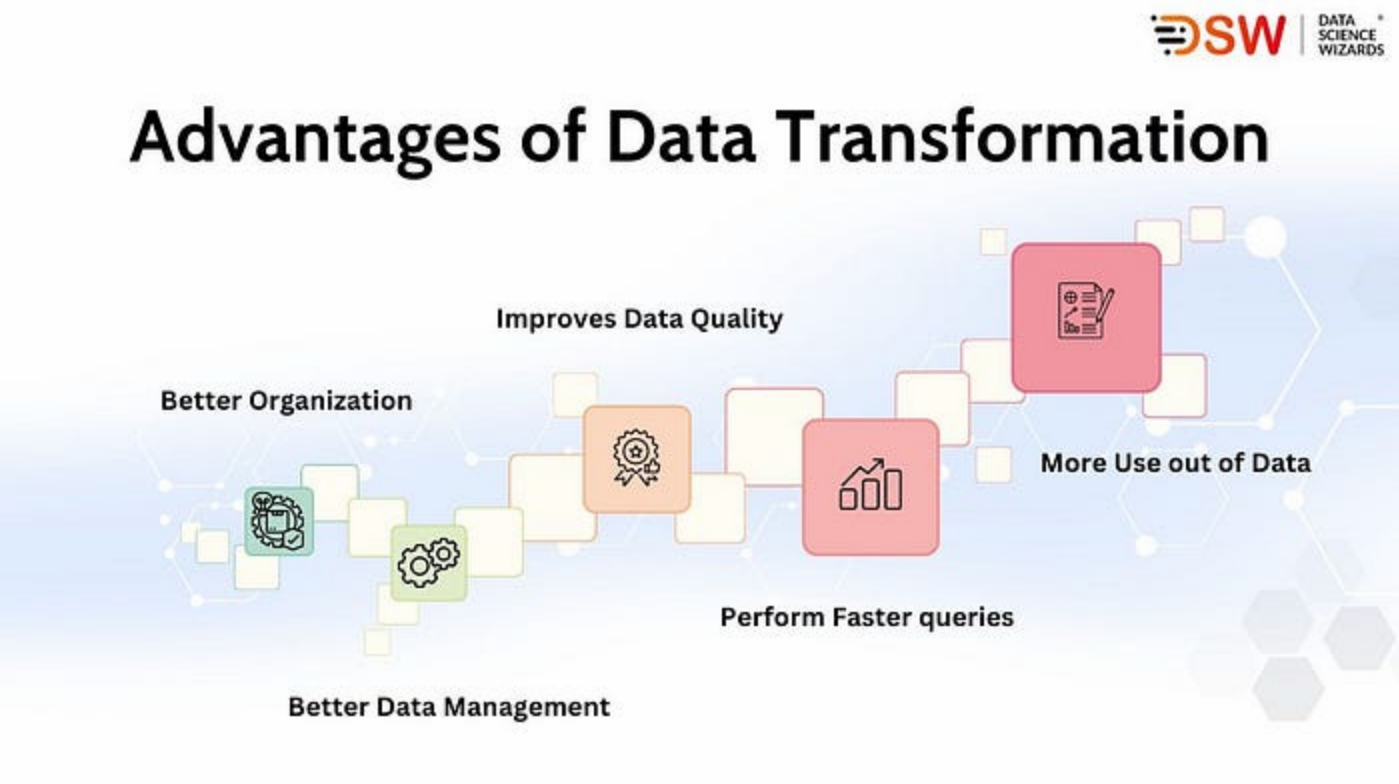
By scripting: this way, we write codes in python, SQL or other languages to query and transform the data. However, using python and SQL has its own benefits because they help automate various tasks, as

well as they require less coding to perform than a traditional programming language requires.

Using ETL tools: these tools are designed to make the data extraction and transformation work easier than scripting on-premises. These tools can be hosted on the organisation's server and can save a lot of time. But these tools often require expertise in using tools and significant infrastructure costs.

Cloud-based ETL tools: these tools are based on a relevantly new technology where it can be hosted in the cloud. These tools are designed to provide the easiest way to extract and transform data even a non-technical person can perform operations on it. Using these tools, we can store the data in any cloud source, and when required easily, we can load the data in a data warehouse. With the help of such ETL tools, we can schedule data pulling while monitoring the usage.

Advantages of Data Transformation



There are various advantages of data transformation to any business, such as:

- **Improved data quality:** we all know of the risks and costs associated with low-quality data, so data transformation helps organisations eliminate data quality issues.
- **Faster Queries:** it becomes easy and quick to extract transformed data from its storage or location.
- **Efficient Data Management:** as data is non-stop generating in different sources, it becomes challenging to build and understand metadata. Data transformation helps in refining the metadata, and this is how data management becomes more efficient.

- **Better organisation:** Data transformation makes the process easy because it is more interpretable for both humans and computers.
- **More use of data:** In the process of generating and collecting data continuously, a lot of data gets unanalysed. Data transformation lets the get the most out of data by standardising it and making it more usable.

Conclusion

Here in this article, we have discussed different aspects of data transformation, including techniques of data transformation, how the process works, and what are the general ways using which we perform data transformation. There are various advantages of data transformation, and in many organisations, it has become a compulsory process to perform because it lets the organisations use their data appropriately and unleash the full potential of their data. In our next articles, we will discuss more details about this topic, and we hope you get valuable insights from this article.

About DSW

DSW, specializing in Artificial Intelligence and Data Science, provides platforms and solutions for leveraging data through AI and advanced analytics. With offices located in Mumbai, India, and Dublin, Ireland, the company serves a broad range of customers across the globe.

Our mission is to democratize AI and Data Science, empowering customers with informed decision-making. Through fostering the AI ecosystem with data-driven, open-source technology solutions, we aim to benefit businesses, customers, and stakeholders and make AI available for everyone.

Our flagship platform ‘UnifyAI’ aims to streamline the data engineering process, provide a unified pipeline, and integrate AI capabilities to support businesses in transitioning from experimentation to full-scale production, ultimately enhancing operational efficiency and driving growth.