

Data transformation steps to ensure reliable data

Previously, you learned that data transformation is the process of taking raw data and converting it into a usable format. In a data pipeline, transformation is the stage of taking data, cleaning it, and putting it into a standard format to be stored. In this reading, you'll learn more about the role of transformation in a data pipeline, and the steps analysts take to transform data.

The essence of data transformation

The purpose of data transformation is to prepare raw data for analysis. With the rise of technology integrations for businesses, like data collection and tracking software, there's an increasing variety and velocity of data to be analyzed. Data transformations range from simple data type conversions, to complex aggregations depending on the dataset and business needs. Data transformation is crucial to ensuring the accuracy and reliability of data, and the analysis it's used for. The work done in the transformation phase determines how the data will be stored and used moving forward.

Steps in data transformation

Analysts take various steps to ensure that data is both accurate and fit for analysis. Remember, not all datasets will require every transformation step. The dataset's initial quality, and the business needs for analysis, will determine whether a step is needed. Data transformation steps include profiling, cleansing, conversion, and aggregation. Let's explore each of these steps in further detail.

Profiling

Data profiling is the process of exploring data to identify quality issues. Profiling data is like taking inventory of a store's items. In this step, analysts gather information about the data's structure, format, values, and relationships. Analysts might find issues like missing or null values, outliers, duplicate records, inaccurate data, and inconsistent data formats while profiling. Once issues are identified, analysts determine their plan for transformation.

Cleansing

The errors found in the data profiling process drive the next step: data cleansing. Data cleansing is the process of fixing or removing data quality issues. Data cleansing ensures that the data is accurate, consistent, and complete. As an analyst, you want to make sure no errors

are present in the data to ensure accurate analysis. For example, during the data profiling phase, you could notice inconsistent formatting in a column containing numeric data. Some entries are whole numbers, while others are floating numbers. During the cleansing process, it's your responsibility to implement a consistent format based on the needs and use of the data.

Conversion

Data conversion is the process of changing the format of data to improve compatibility, readability, or make data more secure. In this step, analysts convert data to make it compatible with different systems or applications, to save storage space, or make it easier to understand and use. For example, timestamp data might need to be changed from one time zone to another, or categorical data might need to be converted into numerical data to perform specific analysis.

Aggregation

Data aggregation involves summarizing data. When analysts deal with large datasets, it's essential to derive high-level insights. For example, daily sales data for five years could be aggregated monthly, quarterly, or even yearly for a general view. Creating a data summary is effective for a brief overview of the data including averages, minimums, and maximums. If you wanted to quickly gather insights on sales data, you could aggregate the sales to display the minimum and maximum sales for the month, quarter, or year.

Key takeaways

Data transformation prepares raw data for analysis. Transformation processes vary according to business needs and dataset constraints. Some steps data analysts take to transform data include profiling, cleansing, conversion, and aggregation. Intentional transformation processes are necessary for accurate and reliable data analysis.

Resources for more information

This resource provides more information on the role of data transformation and methods for data cleansing:

- A comprehensive guide detailing the importance and methods of effective data cleaning: [Guide to Data Cleaning: Definition, Benefits, Components, and How to Clean Your Data](#)