Google Cloud

Data transformation techniques that achieve different business needs

So far, you've learned that data transformations are a step in the data journey to creating meaningful analysis. Data transformation techniques are directly connected with businesses' needs for accurate and actionable data. Desired outcomes for the data determined in the collection phase drive transformations. In this reading, you'll learn more about data transformation techniques and their on-the-job applications.

Considerations for data transformation

Data transformation involves several processes to convert raw data into a structured and usable format for analysis. As an analyst, you must consider the data's initial state, the desired outcome, and tools available before choosing a transformation method. Tools like Python and SQL are available to connect to many data sources, enabling you to locate information and manipulate data how you choose.

Not all transformation techniques are applicable to every dataset. Each transformation technique has its own benefits and potential challenges. The selection process is crucial for the accuracy of the final analysis. Always start with a thorough data assessment before deciding on a transformation technique to help identify the unique challenges and needs of the data. There are many methods to transform data, like aggregation, deduplication, data derivation, normalization, and tokenization. Let's explore these methods in more detail.

Aggregation

Aggregation is when you summarize entries of data. As an analyst, you'll handle data from a variety of sources that need to be combined. For example, if you have sales data spanning multiple years, you can aggregate or combine this data to present the average sales for each year. Aggregating data is useful for summarizing data at a high-level. Businesses use summary data to make informed decisions about marketing strategies, product pricing, and operational structure.

Deduplication

Deduplication means removing redundant data entries. Duplicate data entries skew the outcome of the final analysis. As an analyst, it's your responsibility to locate and handle duplicate data to ensure you're working with the most accurate information. Data duplication is a common process for fixing or removing data quality issues. Consider a social media app that seeks to track user engagement on their platform. Analysts must account for usage across all

devices like phones, tablets, and desktop computers. Each device must be attached to the respective account to ensure multiple devices aren't counted as multiple users. Otherwise, the data will look like there's much more engagement than there really is.

Data derivation

Data derivation is when you create, or derive, new data from existing data points. Analysts rely on derivation techniques like creating new measures or columns to better understand the data. For example, an e-commerce business wants to determine their most and least successful products within the last year to make decisions about restock. The business gives you their sales data for the last year containing product codes, inventory amounts, and sales per day. You want to know the average and total sales per product. While the business didn't give you this information directly, you can extract this information by creating new columns based on product codes, and summarizing the sales data for each product.

Normalization

Normalization is the process of reorganizing data in a database to make it easier to find, group, and analyze. Normalization can make data more consistent, accurate, and complete. For example, imagine an e-commerce company with sales data initially stored in a single table with inefficiencies and inconsistencies. By normalizing the data, it's reorganized into separate tables for Customers, Orders, Products, and Shipping. This restructuring addresses data redundancy and information is stored in one place instead of being repeated across multiple orders. Analysts can easily perform analysis like evaluating sales trends or assessing product performance. Normalization creates a more efficient, accurate, and user-friendly database system.

Tokenization

Tokenization is replacing sensitive data with unique symbols. Tokenization is necessary to help protect sensitive information like PII and PHI. Data including, but not limited to names, birthdates, addresses, social security numbers, credit card information, and medical information are considered sensitive data. As an analyst, you might find yourself working in healthcare or finance. You'll work with this information on a need-to-know basis, meaning you'll handle only the data needed to complete your job. For example, if you're working for a major bank, and they want to analyze credit card usage for their newest travel card, names and credit card numbers would be tokenized to help protect user information.

Key takeaways

Data transformation techniques are necessary for converting raw data into a structured format for analysis. Before performing transformation techniques, a data assessment will be useful for determining errors and inconsistencies within a dataset. Not every transformation technique

will be applicable to your data. The choice of transformation technique aligns with the specific needs and desired outcomes for the data.