CDS DS 594

Data Transformation: A Comprehensive Guide to Benefits, Challenges, and Tools

1. What is Data Transformation?
    a. Analyzing information requires structured and accessible data for best results
    b. Data transformation enables organizations to alter the structure and format of raw data as needed.
2. Defining Data Transformation and its Role in Data Management
    a. Businesses run on data that is used to inform decision making in every realm of organization
    b. For data to be useful, it has to be changed from its raw data source form into an easy format for applications and systems to use
    c. To achieve this, companies use data transformation to convert the data into the needed format
3. The Importance of Transforming Raw Data for Analysis and Visualization
    a. Make data usable for analysis and visualization, key components of business intelligence and data-driven decision making
    b. Businesses generate and collect vast amounts of data, but until it is transformed, its value cannot be leveraged
    c. Raw data is often stored in data warehouses or data lakes, where it waits to be selected and used for analysis
4. How Data Transformation Fits into the ETL/ELT Process
    a. To obtain the data from its repository, businesses use related data transformation processes called extract/transform/load (ETL) and extract/load/transform (ELT).
    b. For data stored in on-premises data warehouses, ETL extracts the data from the repository, transforms it into the required format, then loads it into an application or system. There it can be used for business intelligence, data analysis, and other purposes.
    c. For cloud-based data warehouses, the ELT process is used. The scalability of the cloud platform lets organizations skip preload transformations and load raw data into the data warehouse, then transform it at query time.
    d. 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).
5. The Benefits and Challenges of Data Transformation
    a. Transforming data yields several benefits:
        i. Data is transformed to make it better organized. Transformed data may be easier for both humans and computers to use.

        ii. 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.
        iii. Data transformation facilitates compatibility between applications, systems, and types of data. Data used for multiple purposes may need to be transformed in different ways.
    b. Challenges
        i. Data transformation can be expensive. The cost is dependent on the specific infrastructure, software, and tools used to process data. Expenses may include software licensing, computing resources, and the time spent on task by the needed personnel.
        ii. Data transformation processes can be resource-intensive. Performing transformations before loading into a data warehouse, or transforming data before feeding it into applications can create a computational burden that slows down other operations.
        iii. Lack of expertise and carelessness can introduce problems during transformation. Data analysts without appropriate subject matter expertise are less likely to notice typos or incorrect data because they are less familiar with the range of accurate and permissible values.
        iv. Enterprises can perform transformations that don't suit their needs. A business might change information to a specific format for one application only to then need to revert the information back to its prior format for a different application.
6. Techniques for Data Transformation
    a. Data transformation can increase the efficiency of analytic and business processes and enable better data-driven decision-making.
    b. The first phase of data transformations should include things like data type conversion and flattening of hierarchical data.
    c. These operations shape data to increase compatibility with analytics systems
7. Extraction and Parsing: Accessing Data from Different Sources
    a. Data ingestion begins with extracting information from a data source, followed by copying the data to its destination.
    b. Initial transformations are focused on shaping the format and structure of data to ensure its compatibility with both the destination system and the data already there.
    c. Parsing fields out of comma-delimited log data for loading to a relational database is an example of this type of data transformation.
    d. Must replicate it to a data warehouse architected for analytics.
    e. Most organizations today choose a cloud data warehouse, allowing them to take full advantage of ELT.

8. Translation and Mapping: Converting Data Formats and Structures
    a. Some of the most basic data transformations involve the mapping and translation of data
    b. Translation converts data from formats used in one system to formats appropriate for a different system.
    c. Even after parsing, web data might arrive in the form of hierarchical JSON or XML files, but need to be translated into row and column data for inclusion in a relational database.
9. Filtering, Aggregation, and Summarization: Reducing and Generalizing Data
    a. Data transformation is often concerned with whittling data down and making it more manageable.
    b. Data may be consolidated by filtering out unnecessary fields, columns, and records.
    c. Omitted data might include numerical indexes in data intended for graphs and dashboards or records from business regions that aren't of interest in a particular study.
    d. Data might also be aggregated or summarized by, for instance, transforming a time series of customer transactions to hourly or daily sales counts.
10. Enrichment and Imputation: Handling Missing Values and Enhancing the Dataset
    a. Data from different sources can be merged to create denormalized, enriched information.
    b. A customer's transactions can be rolled up into a grand total and added into a customer information table for quicker reference or for use by customer analytics systems.
    c. Long or freeform fields may be split into multiple columns, and missing values can be imputed or corrupted data replaced as a result of these kinds of transformations.
11. Indexing and Ordering: Organizing Data for Optimal Retrieval
    a. Data can be transformed so that it's ordered logically or to suit a data storage schema.
    b. In relational database management systems, for example, creating indexes can improve performance or improve the management of relationships between different tables.
12. Anonymization and Encryption: Protecting Sensitive Data
    a. Data containing personally identifiable information, or other information that could compromise privacy or security, should be anonymized before propagation.
    b. Encryption of private data is a requirement in many industries, and systems can perform encryption at multiple levels, from individual database cells to entire records or fields.

13. Modeling, Typecasting, Formatting, and Renaming: Preparing Data for Analysis
    a. A whole set of transformations can reshape data without changing content.
    b. This includes casting and converting data types for compatibility, adjusting dates and times with offsets and format localization, and renaming schemas, tables, and columns for clarity.
14. Data Transformation Tools and Technologies
    a. Businesses have multiple options for data transformation tools and technologies, depending on size of organization, budget, and a company's data management strategy.
    b. ETL Tools for Data Transformation
        i. Enterprise-grade
            1. Sold by commercial organizations and often deliver the most mature solutions.
            2. Geared for businesses that do not have the time or resources to devote to staffing an in-house team to build their own solutions.
            3. Come with pre-defined pipelines, easy to use interfaces, and are built with the IT and line of business user in mind.
        ii. Open-Source
            1. Teams that can develop, build, and maintain their own ETL process often use open-source ETL tools to do so.
            2. Many are free and allow businesses to access the tool's source code to study its technical infrastructure and extensibility.
        iii. Cloud-Based Platform Tools
            1. Integration platform-as-a-service providers often bake ETL tools into their offerings.
            2. Platform-based tools offer high latency, availability, and elasticity, enabling organizations to scale their data transformation to the volume and speed the business needs.
        iv. Custom ETL Tools
            1. Some businesses prefer to develop their own custom ETL tools so they can tailor a solution to their organization's unique infrastructure or priorities.
            2. ETL tools are often built with SQL, Python, and Java programming languages.
            3. This approach requires intensive internal resources to build, test, maintain, and update the tool.
            4. It also requires in-house documentation and training to enable new users.