Data Transformation

- Data Transformation is a technique of conversion as well as mapping of data from one format to another. The tools and techniques used for data transformation depend on the format, complexity, structure and volume of the data.
- It enables a developer to translate between XML, non-XML, and Java data formats, for rapid integration of heterogeneous applications regardless of the format used to represent data.

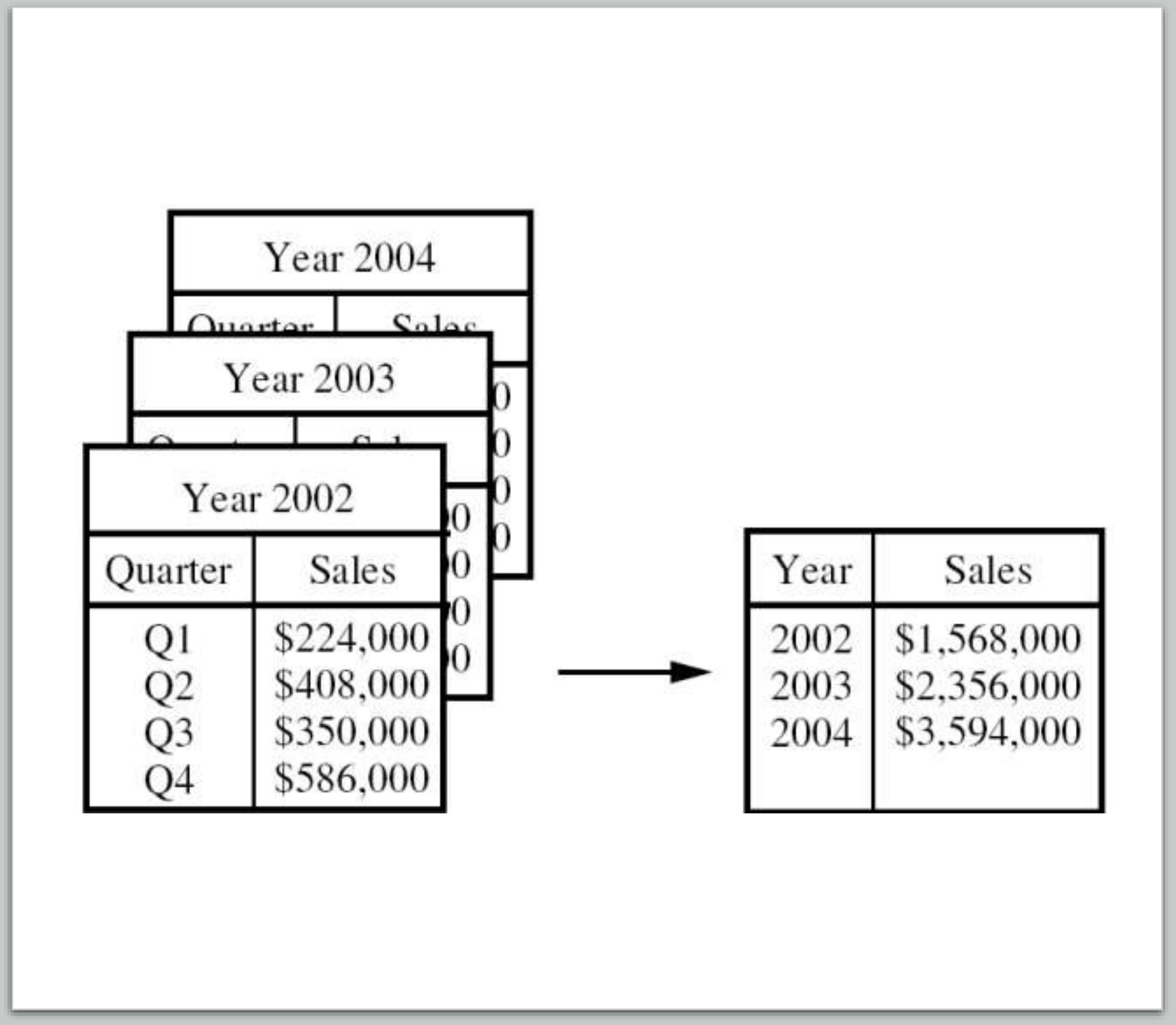
Data Transformation Methods

1. Aggregation
2. Attribute Construction
3. Discretization
4. Generalization
5. Integration
6. Manipulation
7. Normalization
8. Smoothing

1) Aggregation

- Data aggregation is the method where raw data is gathered and expressed in a summary form for statistical analysis. For instance, raw data can be aggregated over a given time period to provide statistics such as average, minimum, maximum, sum, and count.
- After the data is aggregated and written as a report, you can analyze the aggregated data to gain insights about particular resources or resource groups. There are two types of data aggregation: time aggregation and spatial aggregation.
- For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts.

- Sales data for a given branch of AllElectronics for the years 2002 to 2004.
- On the left, the sales are shown per quarter. On the right, the data are aggregated to provide the annual sales



Year 2004

Year 2003

Year 2002

Quarter	Sales
Q1	$224,000
Q2	$408,000
Q3	$350,000
Q4	$586,000

Year	Sales
2002	$1,568,000
2003	$2,356,000
2004	$3,594,000

2) Attribute Construction

- This method helps create an efficient_data mining process. In attribute construction or feature construction of data transformation, new attributes are constructed and added from the given set of attributes to help the mining process.
- new attributes are constructed from the given attributes and added in order to help improve the accuracy and understanding of structure in high-dimensional data.
- Example – we may wish to add the attribute area based on the Data Transformation attributes height and width.
- By attribute construction can discover missing information.

3) Discretization

- Data discretization is the process of converting continuous data attribute values into a finite set of intervals and associating with each interval some specific data value. There are a wide variety of discretization methods starting with naive methods such as equal-width and equal-frequency to much more sophisticated methods such as MDLP.
- Dividing the range of a continuous attribute into intervals – For example, values for numerical attributes, like age, may be mapped to higher-level concepts, like youth, middleaged, and senior.

4) Generalization

- Data Generalization is the method of generating successive layers of summary data in an evaluational database to get a more comprehensive view of a problem or situation.
- Data generalization can help in Online Analytical Processing (OLAP). OLAP is mainly used for providing quick responses to the analytical queries which are multidimensional.
- The method is also beneficial in the implementation of Online transaction processing (OLTP). OLTP refers to a class system designed to manage and facilitate transaction-oriented applications, especially those involved with data entry and retrieval transaction processing.

5) Integration

- Data integration is a crucial step in data pre-processing that involves combining data residing in different sources and providing users with a unified view of these data. It includes multiple databases, data cubes or flat files and works by merging the data from various data sources.

6) Manipulation

- Data manipulation is the process of changing or altering data to make it more readable and organized. Data manipulation tools help identify patterns in the data and transform it into a usable form to generate insights on financial data, customer behavior etc.

7) Normalization

- the attribute data are scaled so as to fall within a small specified range, such as -1.0 to 1.0, 0.0 to 1.0
- Data normalization is a method to convert the source data into another format for effective processing. The primary purpose of data normalization is to minimize or even exclude duplicated data. It offers several advantages, such as making data mining algorithms more effective, faster data extraction, etc.

8) Smoothing

- Data smoothing is a technique for detecting trends in noisy data where the shape of the trend is unknown. The method can help identify trends in the economy, stocks, consumer sentiments etc.