Unit II

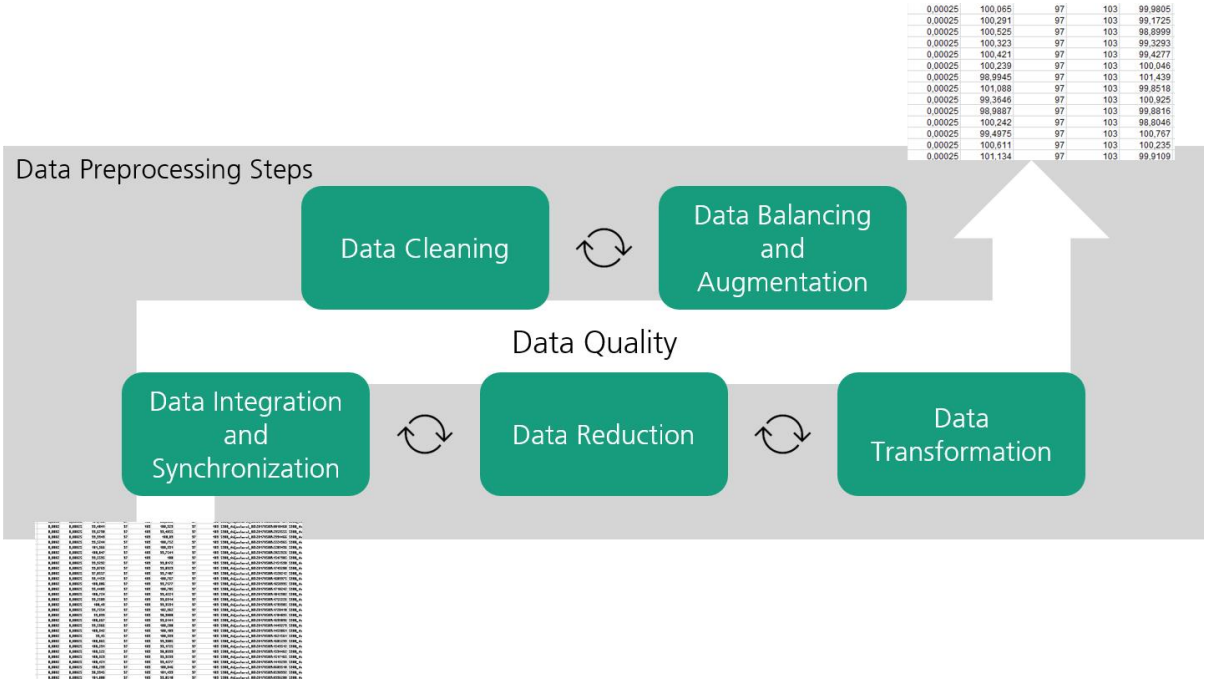
6. Data Quality and Preprocessing

Data Quality and Preprocessing are critical steps in data analytics that focus on ensuring that the data used for analysis is accurate, complete, and relevant. These steps involve identifying and rectifying any issues with the data to ensure that it is suitable for analysis and modeling.

Data Quality:

Data quality refers to the degree to which data is accurate, reliable, and fit for its intended purpose. Poor data quality can lead to misleading or erroneous results, impacting the validity and reliability of any analysis. Some common aspects of data quality include:

1. **Accuracy**: Data should be free from errors and inaccuracies, ensuring that the values recorded are correct and precise.
2. **Completeness:** Data should contain all necessary information without missing values or gaps, ensuring that no essential information is omitted.
3. **Consistency**: Data should be consistent within and across datasets, ensuring that values and formats align correctly.
4. **Timeliness**: Data should be up-to-date and reflect the latest information relevant to the analysis.



Data Preprocessing:

- Data preprocessing involves the transformation and preparation of raw data before analysis. It is a crucial step in data analytics as it can significantly impact the accuracy and effectiveness of the subsequent analysis. Data preprocessing includes the following steps:
- Data Cleaning: This step involves identifying and correcting errors, inconsistencies, and missing values in the data. Techniques like imputation or removal of missing data points are used to address missing values.
- Data Transformation: Data transformation techniques are applied to normalize the data, making it suitable for analysis. Common transformations include scaling, log transformation, and normalization.
- Feature Selection: In some cases, not all features (variables) in the data are relevant for analysis. Feature selection involves choosing the most informative and relevant features to reduce dimensionality and improve model performance.
- Outlier Detection: Outliers are extreme data points that differ significantly from the rest of the data. Identifying and handling outliers is essential to avoid them unduly influencing the analysis.
- Data Integration: When working with multiple datasets, data integration ensures that data from different sources are combined seamlessly, enabling a comprehensive analysis.
- Data Reduction: In cases of high-dimensional data, data reduction techniques like Principal Component Analysis (PCA) can be used to reduce the number of variables while preserving important information.


Data Integration
Data Cleaning
Data Preprocessing
Data Transformation
Data Reduction or Dimension Reduction