Business Intelligence

ADDIS ABABA SCIENCE AND TECHNOLOGY UNIVERSITY
UNIVERSITY FOR INDUSTRY

Week 4

Data Integration and Preparation

- Need for Data Preprocessing
- Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration

Tilahun Melak(PhD)

April, 2026

Objectives

At the end of this lecture students will be able to :

- Explain the need for data preprocessing and preparation
- Discuss the major tasks involved in data preprocessing
- Explain data cleaning
- Explain data integration

Need for Data Preprocessing

- Quality decisions must be based on quality data
- Real world data is often incomplete, inconsistent, and noisy
  - **incomplete**:
    - lacking attribute values that is vital for decision making so they have to be added,
    - lacking certain attributes of interest in certain dimension and should be again added with the required value,
    - containing only aggregate data so that the primary source of the aggregation should be included

Inmon, W. H. (2005). Building the Data Warehouse (4th ed.). Wiley.

Need for Data Preprocessing…

- Real world data is often incomplete, inconsistent, and noisy
  - **noisy**: containing errors or outliers that deviate from the expected
  - **inconsistent**: containing discrepancies in codes or names of the organization or domain
  - etc

Inmon, W. H. (2005). Building the Data Warehouse (4th ed.). Wiley.

Need for Data Preprocessing…

- Incomplete, noisy and inconsistent data are commonplace properties of large real world databases and data sources
- Data cleaning is a routine work to clean such problems so that results can be accepted

Inmon, W. H. (2005). Building the Data Warehouse (4th ed.). Wiley.

Need for Data Preprocessing…

- Before starting data preprocessing, it will be advisable to have overall picture of the data we have so that it tell us high level summary such as
  - General property of the data
  - Which data values should be considered as noise or outliers
- This can be done with the help of ***descriptive data summarization***

Inmon, W. H. (2005). Building the Data Warehouse (4th ed.). Wiley.

Major Tasks in Data Preprocessing

- Data pre-processing in data analytics/business intelligence activity refers to the processing of the various data elements to prepare for the mining operation.
- Any activity performed prior to mining the data to get knowledge out of it is called **data pre-processing**
- This involves:
  - Data cleaning
  - Data integration
  - Data transformation
  - Data reduction
  - Data Discretization and concept hierarchy generation

Inmon, W. H. (2005). Building the Data Warehouse (4th ed.). Wiley.

Data Cleaning

- Refers to the process of improving data quality by:
  - Handling missing values (e.g., imputation or removal)
  - Smoothing noisy data to reduce variability
  - Identifying and removing outliers
  - Resolving inconsistencies and errors in the dataset
- Ensures data is accurate, consistent, and ready for analysis

Inmon, W. H. (2005). Building the Data Warehouse (4th ed.). Wiley.

Data Cleaning: Missing Data

- Data is not always be available (or missed)
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - not entered into the database due to misunderstanding
  - Some data may not be considered important at the time of entry
- Missing data may need to be inferred.

Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.

Data Cleaning: How to Handle Missing Data?

- **Ignore the tuple:** usually done when class label is missing (assuming the tasks is classification)
  - not effective when the percentage of missing values per attribute is significantly large.
- **Fill in the missing value manually:** tedious and infeasible

Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.

Data Cleaning: How to Handle Missing Data?

- **Use a global constant to fill in the missing value:**

  e.g., “unknown”, a new class?!

  – Simple but not recommended as this constant may form some interesting pattern for the data mining task which mislead decision process

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Cleaning: How to Handle Missing Data?

- **Use the attribute mean**
  - for all samples belonging to the same class fill in the missing value with the mean
- **Use the most probable value** to fill in the missing value:
  - inference-based such as Bayesian formula (probability) or decision tree (information theory)

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Cleaning: How to Handle Missing Data?

- Except handling the missing value using by ignore the tuple and filling in the missing value manually, the filled values are incorrect

- Using the attribute mean and Use the most probable value are the most commonly used technique to fill missing data

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Cleaning: Noisy Data

- **Noise:** is random error or variance in a measured variable
- Incorrect attribute values may be due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
  - duplicate records
  - incomplete data per field

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Cleaning: How to Handle Noisy Data?

Noisy data can be handled using the following techniques:

- **Binning (Simple Discretization)**
  - Smooths data by grouping values into bins
- **Clustering**
  - Identifies and removes outliers based on group similarity
- **Regression**
  - Fits data to a function to smooth noise
- **Computer and Human Inspection**
  - Automatically detects suspicious values
  - Human experts verify and correct anomalies

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Cleaning: Binning Method for Handling Noisy Data

- Sort the data and partition it into bins
- Bins can be created using:
  - Equal-width binning
  - Equal-depth (equal-frequency) binning
- Apply a smoothing technique to each bin
- Common smoothing methods include:
  - Smoothing by bin means
  - Smoothing by bin medians
  - Smoothing by bin boundaries

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Cleaning: Binning Method for Handling Noisy Data

- Equal-width (distance) partitioning:
  - It divides the range into *N* intervals of equal size: uniform grid
  - if *A* and *B* are the lowest and highest values of the attribute, the width of intervals will be: $W = (B-A)/N$.
  - This approach leads into bins with non uniform distribution of data elements per bin
  - The most straightforward approach
  - But outliers may dominate presentation
  - Skewed data is not handled well.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Cleaning: Binning Method for Handling Noisy Data

- Equal-width (distance) partitioning:
  - Given the dataset (say 24, 21, 28, 8, **4**, 26, **34**, 21, 29, 15, 9, 25)
    - Determine the number of bins : N (say 3)
    - Determine the range R = Max – Min
    - Divide the range into N equal width where the ith bin is [Xi-1, Xi) where X0=Min and XN=Max and Xi = Xi-1 + R/N
    - For the above data R = 30, R/3 = 10, X1 = 14, X2 = 24, and X3 = 34
    - First sort the data as 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
    - Therefore in our case Bin 1 = 4,8,9 Bin 2 = 15, 21, 21 Bin3 = 24, 25, 26, 28, 29, 34

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Cleaning: Binning Method for Handling Noisy Data

- Equal-depth (frequency) partitioning:
  - It divides the range into *N* intervals, each containing approximately same number of samples
  - Given the data set (say 24, 21, 28, 8, 4, 26, 34, 21, 29, 15, 9, 25)
    - Determine the number of bins : N (say 3)
    - Determine the frequency F of the data set (12)
    - Determine the number of sample per bin F/N (12/3 = 4)
    - Sort the data as 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34 and place F/N element in order into a bin
    - Therefore in our case Bin1 = 4,8,9 ,15 Bin2 = 21, 21,24, 25 Bin3 = 26, 28, 29, 34
  - Good data scaling
  - Managing categorical attributes can be tricky.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Cleaning: Binning Method for Handling Noisy Data

Smoothing algorithm

- Given the data set in bins as say:
  - **Bin 1:** 4, 8, 9, 15
  - **Bin 2:** 21, 21, 24, 25
  - **Bin 3:** 26, 28, 29, 34

- Smoothing by bin means:
  - Find the mean in each bin and replace all the element by the bin mean
  - **Bin 1:** 9, 9, 9, 9
  - **Bin 2:** 23, 23, 23, 23
  - **Bin 3**: 29, 29, 29, 29

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Cleaning: Binning Method for Handling Noisy Data

Smoothing algorithm

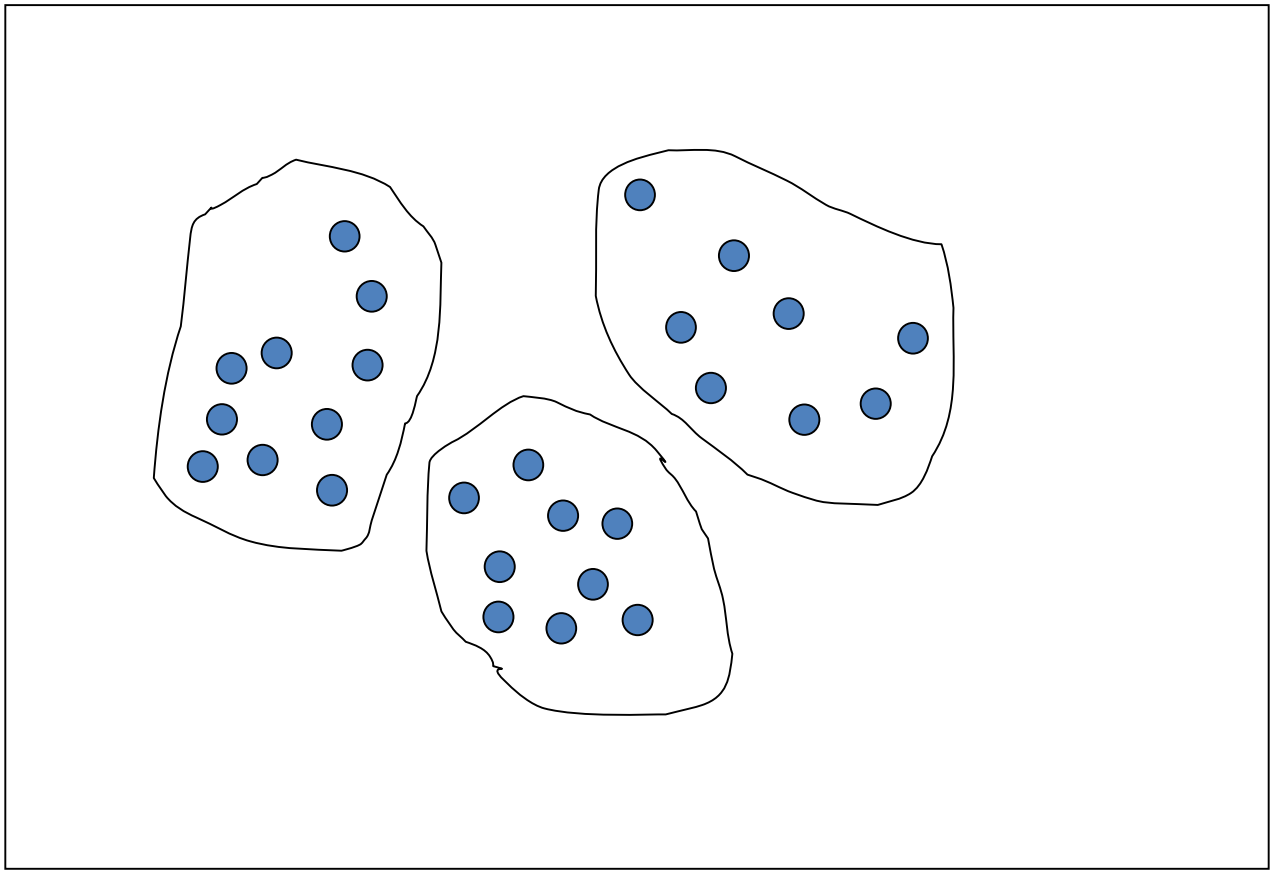
- Smoothing by bin median:
  - Find the median in each bin and replace all the element by the bin median
  - **Bin 1:** 8.5, 8.5, 8.5, 8.5
  - **Bin 2:** 22.5, 22.5, 22.5, 22.5
  - **Bin 3**: 28.5, 28.5, 28.5, 28.5

- Smoothing by bin boundaries:
  - Replace each element by the bin min or bin max which ever is the nearest
  - Distance is measured just as the absolute of the difference
  - **Bin 1:** 4, 4, 4, 15
  - **Bin 2:** 21, 21, 25, 25
  - **Bin 3:** 26, 26, 26, 34

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Cleaning: Cluster Analysis Method for Handling Noisy Data

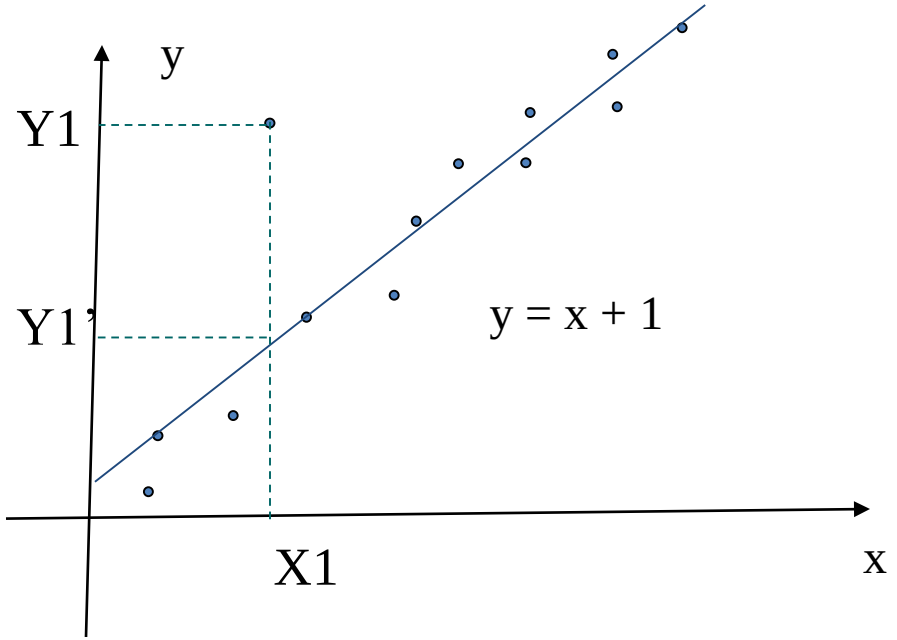
- Detect and remove outliers



Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Cleaning: Handling Noisy Data by Regression

- smooth by fitting the data into regression functions
- Finding a fitting function for two dimensional case so that the value of the second variable can be estimated using the value of the first variable



- It can be extended to N dimensional case in which regression finds multidimensional space so that value of one of the dimension can be predicted from the rest N-1 values

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Cleaning: Identifying and Removing Outliers

- Can be done by
  - cluster analysis
  - The box plot techniques

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Cleaning: Resolve Inconsistencies

- Involve write the same information in many ways such as Bahrdar vs Bahirdar, AAU vs A.A.U., $X vs Y Birr

- Very difficult and most important task which require manual investigation

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Data Integration

- **Data integration:**
  - Combines data from multiple sources (databases, data cubes, or files) into a coherent store

- There are a number of issues to consider during data integration

- Some of these are
  - Schema integration issue
  - Entity identification issue
  - Data value conflict issue
  - Avoiding redundancy issue

Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.

Data Integration…

- Schema integration
  - Schema refers to the design of an entity and its relation in the data source
  - Integrate metadata from different sources

Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.

Data Integration…

- Entity identification problem:
  - identify real world entities from multiple data sources which are identical so that they can be integrated properly
  - As data source for data analytics differ, the same entity will have different representation in the different sources
  - How can equivalent real world entities from multiple sources can be matched up?
  - e.g., A.cust-id ≡ B.cust-#

Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.

Data Integration…

- **Data value conflict issue**
  - Involves detecting and resolving data value conflicts
  - For the same real world entity, attribute values from different sources may be different
  - Possible reasons: different representations, different scales, measurement unit used

Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.

Data Integration…

- **Avoiding redundancy data issue**
  - Redundant data occur often during integration of multiple databases
    - The same attribute may have different names in different databases
    - One attribute may be a "derived" attribute in another table, e.g., annual revenue from monthly revenue
  - Redundant data may be able to be detected by ***correlation analysis for numeric data***

$$r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \bar{A})(b_i - \bar{B})}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^{N}(a_i b_i) N\bar{A}\bar{B}}{N\sigma_A\sigma_B}$$

Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.

Data Integration…

- **Avoiding redundancy data issue**
  - The correlation between two attribute A and B ($r_{A,B}$) is always in the range from -1 to +1.
  - -1 is to mean negatively correlated, 0 to mean uncorrelated and +1 is perfectly correlated
  - Careful integration of the data from multiple sources may help to reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.

Summary

- In todays lecture we have discussed about;
  - Importance of data preprocessing and preparation
  - Key tasks involved in data preprocessing
  - Data cleaning
  - Data integration

References

- Inmon, W. H. (2005). Building the Data Warehouse (4th ed.). Wiley.
- Jiawei Han, Micheline Kamber, & Jian Pei. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.