RAMA UNIVERSITY

www.ramauniversity.ac.in

FACULTY OF EGINEERING

DATA MINING & WAREHOUSEING
LECTURE-26

MR. DHIRENDRA
ASSISTANT PROFESSOR
RAMA UNIVERSITY

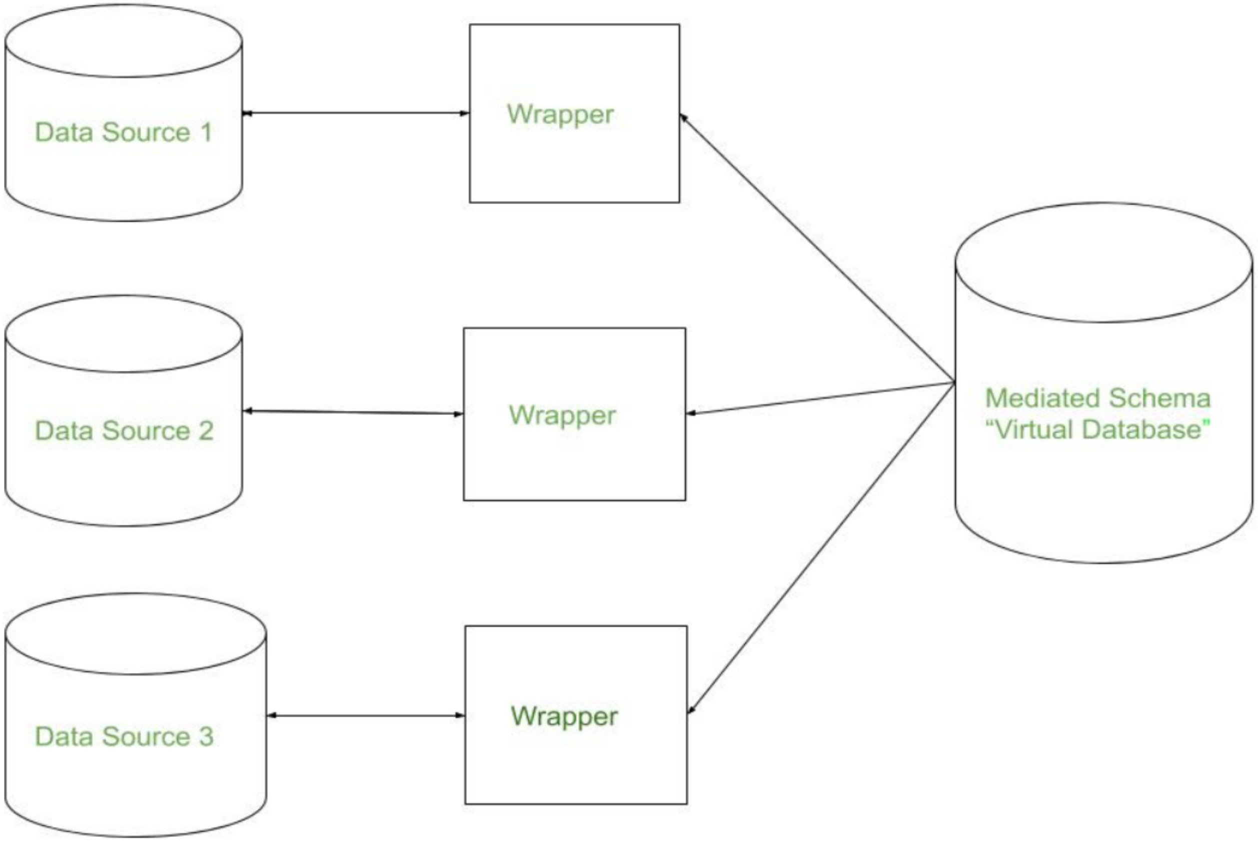
OUTLINE

- DATA INTEGRATION IN DATA MINING
- DATA INTEGRATION IN DATA MINING
- TIGHT COUPLING
- LOOSE COUPLING
- ISSUES IN DATA INTEGRATION
- MCQ
- REFERENCES

Data Integration in Data Mining

- Data Integration is a data preprocessing technique that involves combining data from multiple heterogeneous data sources into a coherent data store and provide a unified view of the data. These sources may include multiple data cubes, databases or flat files.
- The data integration approach are formally defined as triple <G, S, M> where,
- G stand for the global schema,
- S stand for heterogenous source of schema,
- M stand for mapping between the queries of source and global schema.

Data Integration in Data Mining



Data Integration in Data Mining

•There are mainly 2 major approaches for data integration – one is “tight coupling approach” and another is “loose coupling approach”.

Tight Coupling

Loose Coupling

Tight Coupling

- Here, a data warehouse is treated as an information retrieval component.
- In this coupling, data is combined from different sources into a single physical location through the process of ETL – Extraction, Transformation and Loading.

Loose Coupling

- Here, an interface is provided that takes the query from the user, transforms it in a way the source database can understand and then sends the query directly to the source databases to obtain the result.
- And the data only remains in the actual source databases.

Issues in Data Integration:

Issues in Data Integration:

There are no of issues to consider during data integration: Schema Integration, Redundancy, Detection and resolution of data value conflicts. These are explained in brief as following below.

Schema Integration:

- Integrate metadata from different sources.
- The real world entities from multiple source be matched referred to as the entity identification problem.
- For example, How can the data analyst and computer be sure that customer id in one data base and customer number in another reference to the same attribute.

Redundancy:

- An attribute may be redundant if it can be derived or obtaining from another attribute or set of attribute.
- Inconsistencies in attribute can also cause redundanciesin the resulting data set.
- Some redundancies can be detected by correlation analysis.

Detection and resolution of data value conflicts:

- This is the third important issues in data integration.
- Attribute values from another different sources may differ for the same real world entity.
- An attribute in one system may be recorded at a lower level abstraction then the “same” attribute in another.

Multiple Choice Question

1. Various visualization techniques are used in ____________ step of KDD.

a) selection
b) transformation
c) data mining.
d) interpretation.

2. Extreme values that occur infrequently are called as __________.

a) outliers
b) rare values.
c) dimensionality reduction.
d) All of the above.

3. Box plot and scatter diagram techniques are _______.

a) Graphical
b) Geometric
c) Icon-based.
d) Pixel-based.

4. ___________ is used to proceed from very specific knowledge to more general information.

a) Induction
b) Compression.
c) Approximation.
d) Substitution.

5. Describing some characteristics of a set of data by a general model is viewed as _____________

a) Induction
b) Compression
c) Approximation
d) Summarization

REFERENCES

- https://www.tutorialspoint.com/dwh/dwh_overview.htm
- http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf DATA MINING BOOK WRITTEN BY Micheline Kamber
- https://www.javatpoint.com/three-tier-data-warehouse-architecture
- M.H. Dunham, " Data Mining: Introductory & Advanced Topics" Pearson Education
- Jiawei Han, Micheline Kamber, " Data Mining Concepts & Techniques" Elsevier
- Sam Anahory, Denniss Murray," data warehousing in the Real World: A Practical Guide for Building Decision Support Systems, " Pearson Education
- Mallach," Data Warehousing System", TMH
- R. Agrawal, A. Gupta, and S. Sarawagi. Modeling multidimensional databases. ICDE'97 S. Chaudhuri and U. Dayal. An overview of data warehousing and OLAP technology. ACM SIGMOD Record, 26:65-74, 1997
- S. Agarwal, R. Agrawal, P. M. Deshpande, A. Gupta, J. F. Naughton, R. Ramakrishnan, and S. Sarawagi. On the computation of multidimensional aggregates. VLDB'96 D. Agrawal, A. E. Abbadi, A. Singh, and T. Yurek. Efficient view maintenance in data warehouses. SIGMOD'97
- E. F. Codd, S. B. Codd, and C. T. Salley. Beyond decision support. Computer World, 27, July 1993.
- J. Gray, et al. Data cube: A relational aggregation operator generalizing group-by, cross-tab and sub-totals. Data Mining and Knowledge Discovery, 1:29-54, 1997.