RAMA UNIVERSITY

www.ramauniversity.ac.in

FACULTY OF EGINEERING

DATA MINING & WAREHOUSEING
LECTURE-07

MR. DHIRENDRA
ASSISTANT PROFESSOR
RAMA UNIVERSITY

OUTLINE

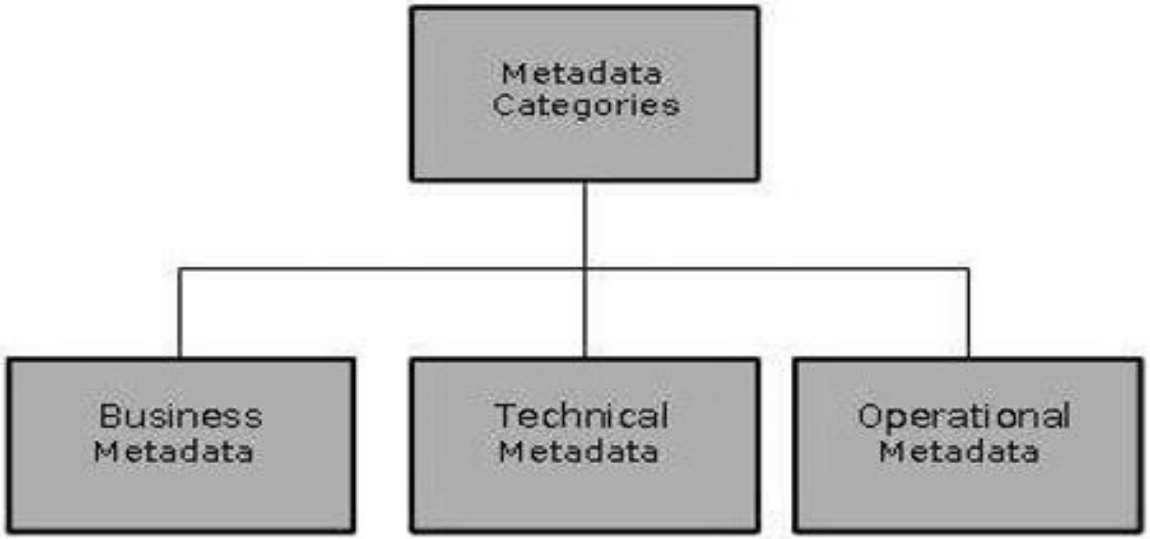
- WHAT IS METADATA
- CATEGORIES OF METADATA
- ROLE OF METADATA
- METADATA REPOSITORY
- CHALLENGES FOR METADATA MANAGEMENT
- MCQ
- REFERENCES

What is Metadata

- Metadata is simply defined as data about data. The data that is used to represent other data is known as metadata. For example, the index of a book serves as a metadata for the contents in the book. In other words, we can say that metadata is the summarized data that leads us to detailed data. In terms of data warehouse, we can define metadata as follows.
- Metadata is the road-map to a data warehouse.
- Metadata in a data warehouse defines the warehouse objects.
- Metadata acts as a directory. This directory helps the decision support system to locate the contents of a data warehouse.

Categories of Metadata

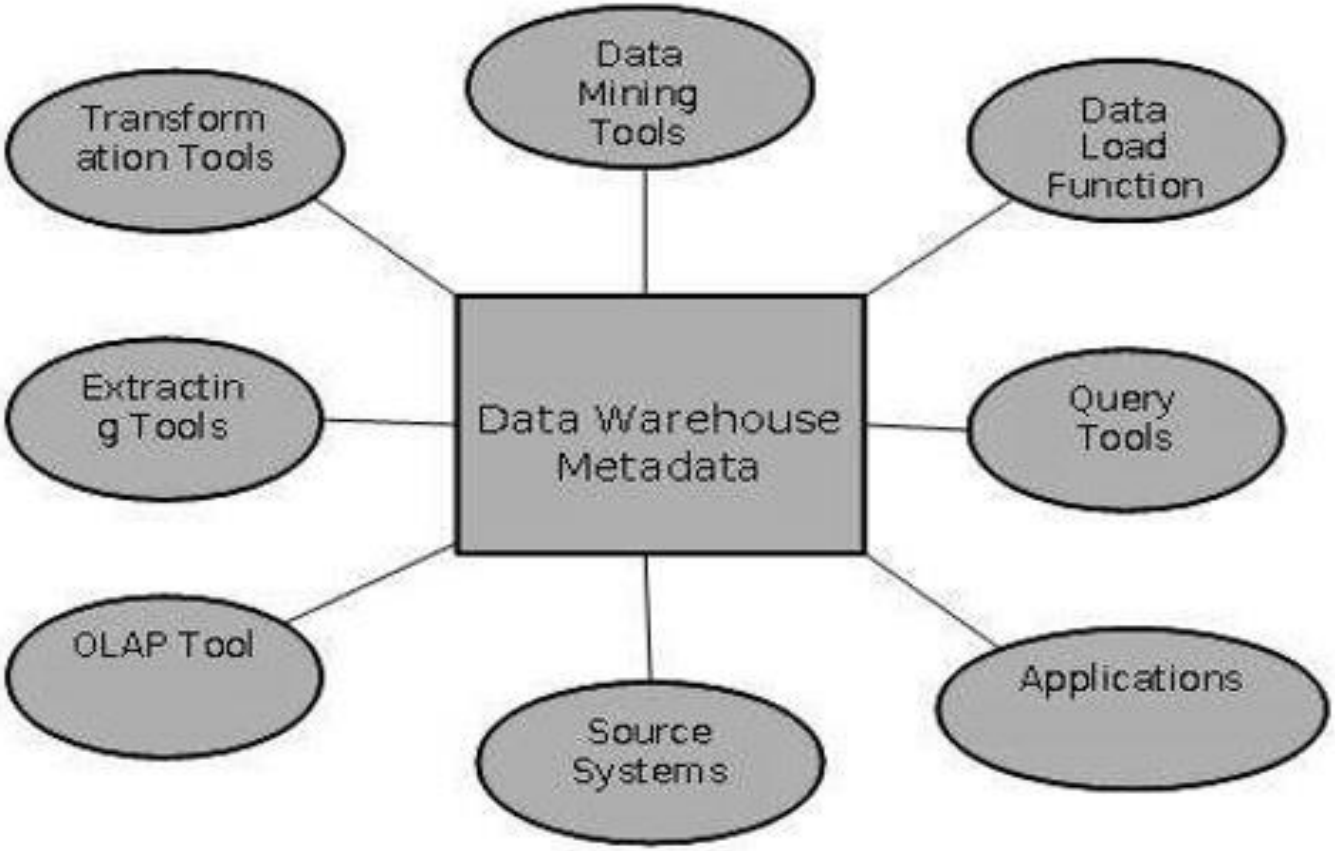
- **Business Metadata** – It has the data ownership information, business definition, and changing policies.
- **Technical Metadata –** It includes database system names, table and column names and sizes, data types and allowed values. Technical metadata also includes structural information such as primary and foreign key attributes and indices.
- **Operational Metadata** – It includes currency of data and data lineage. Currency of data means whether the data is active, archived, or purged. Lineage of data means the history of data migrated and transformation applied on it.



Role of Metadata

- Metadata acts as a directory.
- This directory helps the decision support system to locate the contents of the data warehouse.
- Metadata helps in decision support system for mapping of data when data is transformed from operational environment to data warehouse environment.
- Metadata helps in summarization between current detailed data and highly summarized data.
- Metadata also helps in summarization between lightly detailed data and highly summarized data.
- Metadata is used for query tools.
- Metadata is used in extraction and cleansing tools.
- Metadata is used in reporting tools.
- Metadata is used in transformation tools.
- Metadata plays an important role in loading functions.

Role of Metadata



Metadata Repository

- **Definition of data warehouse** – It includes the description of structure of data warehouse. The description is defined by schema, view, hierarchies, derived data definitions, and data mart locations and contents.
- **Business metadata** – It contains has the data ownership information, business definition, and changing policies.
- **Operational Metadata** – It includes currency of data and data lineage. Currency of data means whether the data is active, archived, or purged. Lineage of data means the history of data migrated and transformation applied on it.
- **Data for mapping from operational environment to data warehouse** – It includes the source databases and their contents, data extraction, data partition cleaning, transformation rules, data refresh and purging rules.
- **Algorithms for summarization** – It includes dimension algorithms, data on granularity, aggregation, summarizing, etc.

Challenges for Metadata Management

- Metadata in a big organization is scattered across the organization. This metadata is spread in spreadsheets, databases, and applications.
- Metadata could be present in text files or multimedia files. To use this data for information management solutions, it has to be correctly defined.
- There are no industry-wide accepted standards. Data management solution vendors have narrow focus.
- There are no easy and accepted methods of passing metadata.

Multiple Choice Question

1. An operational system is _____.
a) used to run the business in real time and is based on historical data.
b) used to run the business in real time and is based on current data.
c) used to support decision making and is based on current data.
d) used to support decision making and is based on historical data.

2.. The generic two-level data warehouse architecture includes _____.
a) at least one data mart.
b) data that can extracted from numerous internal and external sources.
c) near real-time updates.
d) far real-time updates.

3. The active data warehouse architecture includes _____
a) at least one data mart.
b) data that can extracted from numerous internal and external sources.
c) near real-time updates.
d) all of the above.

4. Reconciled data is _____.
a) data stored in the various operational systems throughout the organization.
b) current data intended to be the single source for all decision support systems.
c) data stored in one operational system in the organization.
d) data that has been selected and formatted for end-user support applications.

5. Transient data is _____.
a) data in which changes to existing records cause the previous version of the records to be eliminated.
b) data in which changes to existing records do not cause the previous version of the records to be eliminated.
c) data that are never altered or deleted once they have been added.
d) data that are never deleted once they have been added.

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