Importance and Quality Evaluation of Metadata

Arushi Jain, Aakansha Bansal, Palak Jain, Neha Sharma

Abstract —Metadata usually called data about data represents information about data to be stored in a data warehouse. Importance of metadata arises from the fact that it is needed to map data from source systems to data warehouse with the help of Extraction Transformation Loading (ETL) tools. Metadata helps in developing consistency in data collection and usage and moreover provides the foundation for Data Change Management. It facilitates user to have faster and more accurate access to the data that is needed. Metadata also plays an important role in real world environment as seen in case of legal system nowadays that focuses on preservation of data about its records i.e. metadata so that the validity and admissibility of evidences can be ensured. This paper intends to find a better and more efficient way to determine the importance of metadata in the data warehouse and thus performs its quality evaluation.

Keywords—Consistency, Data Change Management, ETL, Metadata.

I. INTRODUCTION

Internet is expanding day by day and this rapid expansion of internet is leading to a demand for systems that can satisfy various user requirements such as storing, managing, searching etc. The value added information known as metadata documents the important characteristics that include administrative, descriptive, usage history and many more associated with the data. The metadata that is used to describe highly complex objects on the internet is relatively complex than the one used to describe library database. Not only the complexity but the cost associated with generating and exploiting such metadata is also very high. But still it cannot decline the fact that the metadata is required in all the fields.

Cory Doctorow [1] believes that the vision of an internet in which everyone describes their goods, services or information using concise, accurate and common or standardized metadata which is universally understood by both machines and humans is a "pipe-dream, founded on self-delusion, nerd hubris and hysterically inflated marked opportunities". Metadata is an important component of data warehouse. Ralph Kimball describes metadata as the DNA of the data warehouse as metadata defines the elements of the data warehouse and how they work together. Metadata is referred as "technical "or "business focused" [2].

Manuscript received December, 2013.

Arushi Jain, Department of Information Technology Northern India Engineering College, New Delhi, India.

Aakansha Bansal, Department of Information Technology Northern India Engineering College, New Delhi, India.

Palak Jain, Department of Information Technology Northern India Engineering College, New Delhi, India.

Neha Sharma, Department of Information Technology Northern India Engineering College, New Delhi, India.

Metadata is often called encyclopedia of data warehouse. Metadata contains information about ETL transformation which is responsible for mapping of data from operational environment to the DW environment.

Metadata has a positive impact on data quality in data warehouse. Purpose of data warehouse is to give data to the user as and when required. However poor quality of data prohibits successful operation of data warehouse. Poor quality of data may occur because of data coming from different operational systems is difficult to integrate or redundant and inaccurate data that is accumulated in data warehouse. By introducing metadata, Metadata helps to index data, to facilitate access to data and to determine source of the data**.** As such, metadata can be stored and managed in a database, often called a "Metadata Registry" or "Metadata Repository" [3].

Section 2 in the paper describes the important research work done in the area of metadata.

Section 3 describes the research methodologies whereas section 4 explains the role of metadata, examines the need of metadata, types of metadata, life cycle of metadata, components of metadata and important function of metadata in data warehouse.

Section 5 describes the advantages and disadvantages of metadata whereas section 6 gives the overall conclusion on metadata.

II. LITERATURE REVIEW

Eric klovning proposed about the currently growing and emerging practices in the metadata management and also gave an overview about the role of metadata in data warehouse development [4].

Martin Staudt, Anca Vaduva and Thomas Vetterli proposed about the role of metadata particularly in the case of data warehouse since the popularity of data warehousing is growing exponentially [5].

Arun Sen proposed the use of metadata in developing information systems and how it helps in decision support by providing a description about 40-year chronological development of metadata concept [6].

Thanh N. Huynh, Oscar MangiSengi and A Min Tjoa proposed object-relational data warehouse architecture with new metadata layer and described the design and implementation on new kind of metadata [7].

Pete Johnston proposed about the vital role of Extensible Markup Language (XML) in the sharing of structured data across applications and sharing of metadata to facilitate resource discovery [8].

Jane Hunter proposed about various metadata research efforts that are going on nowadays and that are expected to improve the ability of searching, retrieving and assimilating information on the internet [9].

Markus H¨anse, Min-Yen Kan and Achem P. Karduck proposed about an end-to-end scholarly paper metadata harvester named kairos that can be used to find and extract

paper metadata from seed URLs and proactive focused crawls [10].

Gad Kother discussed about the data warehouse based on SAS regarding metadata and specific details needed for SAS end users were highlighted [11].

Hao Fan and Alexandra Poulovassilis proposed about the use of AutoMed Metadata in data warehousing environment to express the data schemas, data cleansing, and transformation and integration process. HDM-low level common data model was adopted to transform and integrate data from multiple heterogeneous data source [12].

Nayem Rahman, Jessica Marz and Shameen Akhter showed that through ETL metadata model, data warehouse refreshes can be made standardized and therefore enabling resource savings in the data warehouse [13].

III. RESEARCH METHODOLOGIES

Research papers are collected from the following journals

- Google scholars
- ACM
- IEEE
- Oxford journal
- Journal of digital information
- IBM system journal

IV.METADATA

The term Metadata is called "data about data" or more precisely we can say "data about container of data". For example music catalog.

It is often called *Meta contents* or "contents about contents, the explanation about the contents is maintained by metadata. Data quality is increased if metadata is properly managed in data warehouse. Metadata contains information about data creation, purpose of data, time and date of creation, author of data and so on.

Metadata not only contain information about important characteristics of data and processes that takes place in the data in data warehouse but also information regarding security and authentication which shows the appropriate working of the data warehouse.

Sample list of definition:

- Data about the data
- Table of contents for the data
- Catalog for the data
- Data warehouse atlas
- Data warehouse roadmap
- Data warehouse directory
- Glue that holds the data warehouse contents together

A. Examining the need for metadata in broad terms

Meta data is important for using, building, and administering data warehouse. Metadata contains the answers to the questions regarding the data in the data warehouse. Metadata is the internal view of the data warehouse showing the inner details of data.

- *For using the data warehouse*

For the users to know about the data in the data warehouse, they need to know about the metadata for browsing and examining the contents of the data warehouse & preventing them from drawing wrong conclusion from their analysis through their ignorance about the exact meanings

- *For building the data warehouse*

For building the data extraction and data transformation component of the data warehouse, we need metadata about the source systems, source-to-target mappings, and data transformation rules.

- *For administering the data warehouse*

To administer the data warehouse we need metadata because of the complexities and enormous sizes of modern data warehouse.

TABLE I: TECHNIQUES USED IN VARIOUS RESEARCH PAPERS

S.No	METHOD/TOOL/TECHNIQUE	APPLICATION	SOURCES
1	*Metadata management in the support of data warehouse development*	Overview about the role of metadata in data warehouse development.	[4]
2	*The role of metadata for data warehousing*	The role of metadata particularly in the case of data warehouse since the popularity of data warehousing is growing exponentially	[5]
3	*Metadata management: past, present and future*	The use of metadata in developing information systems	[6]
4	*Metadata for object-relational data warehouse*	Proposed an object-relational data warehouse architecture with new metadata layer	[7]
5	*Metadata sharing and XML*	Proposed about the vital role of XML(Extensible Markup Language)	[8]
6	*Working towards MetaUtopia*	Improve the ability of searching, retrieving and assimilating information on the internet	[9]
7	*Kairos: proactive harvesting of research paper metadata from scientific conference web sites*	Kairos that can be used to find and extract paper metadata from seed URLs and proactive focused crawls	[10]
8	*Metadata for End-Users*	Data warehouse based on SAS	[11]
9	*AutoMed Metadata in data Warehousing environment*	The use of AutoMed Metadata in data warehousing environment	[12]
10	*ETL Metadata Model for data warehousing*	Through ETL metadata model, data warehouse refreshes can be made standardized	[13]

B. Who needs metadata?

Purpose of data warehouse is to give data to the user as when required. However poor quality of data prohibits successful operation of data warehouse. Poor quality of data may occur because of data coming from different operational systems is difficult to integrate or because of inaccurate and redundant data that is accumulated in data warehouse. By introducing metadata, understanding of end users about the data increases. Metadata helps to index data, to facilitate access to data and to determine source of the data.

C. Metadata is like a nerve center

Various processes during the building and administering of the data warehouse generate parts of the data warehouse metadata. Parts of metadata generated by one process are used by another. In the data warehouse, metadata assumes a key position and enables communication among various processes. It acts like a nerve center in the data warehouse.

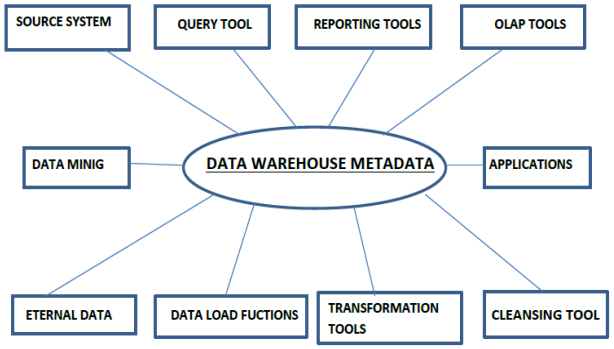

Fig1: Metadata as a nerve center

TABLE II: PURPOSE OF METADATA IN THE DATA WAREHOUSE

Purpose	IT Professionals	Power Users	Casual Users
Information Discovery	Databases, Tables, Columns, Server, Platforms	Databases, Tables, Columns	List of predefined queries and reports, business views
Meaning of Data	Data Structure, Data Definition, Data Mapping, Data Cleansing Functions, Transformation Rules	Buisness Terms, Data Definitions, Data Mapping, Cleansing Functions, Transformation Rules	Buisness Terms, Data Definitions, Filters, Data Sources, Conversion, Data Owners
Information Access	Program Code in SQL,3GL,4GL,front-end applications, Security	Query Toolsets, Database access for complex analysis	Authorization Requests, Information retrieval into Desktop Applications such as spreadsheets

D. Various Types of Metadata

Different types of metadata that are defined are given in the tabular form below:

TABLE III: TYPES OF METADATA

S.No.	Category	*Description*
1.	Operational Metadata	Contains all information about the operational data sources. When one runs a datastage, operational metadata describes the events and processes that occur and the objects that are affected.
2.	Extraction and Transformation Metadata	Contains information about all the data transformation that take place in the data staging area
3.	End-user Metadata	It enables the end-users to find information from the data warehouse
4.	Descriptive Metadata	Used for describing and identifying information resources. Elements that can be included are title, abstract, author and keywords.
5.	Structural Metadata	It is used so that navigation and presentation of electronic resources can be facilitated for the users. It also describes how compound objects are compiled together as pages are compiled in a specific order to form a book.
6.	Administrative Metadata	It is mainly used for the management (either short term or long term) and processing of digital collections. Two types of subsets are listed as: - Rights Management Metadata - Preservation Metadata
7.	Technical Metadata	It describes data structures like tables, fields, data types. It is also used for the documentation of hardware and software.
8.	Business Metadata	It stresses on the information which is important from business point of view. In the day-to-day conduct of business it is highly useful for the business person.
9.	Process Metadata	It is the third category added by Ralph Kimball after technical and business metadata. It is used to explain consequences of various operations performed in data warehouse.

E. Life Cycle of Metadata

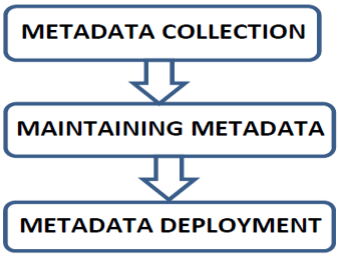

Fig2: Life Cycle of Metadata

- *Metadata collection*
  Metadata is identified and captured in the central repository
- *Maintaining metadata*
  After the collection of metadata is over, maintenance phase begins. To maintain high quality of data, automation is required. Data warehouse queries require tables and various data structures dynamically, so new information should be periodically added without changing the existing information.
- *Metadata deployment*
  Metadata is of utmost importance to the end users to explore the warehouse to fetch information regarding what is present in the warehouse and which queries are present that can be reused.

F. Metadata objectives

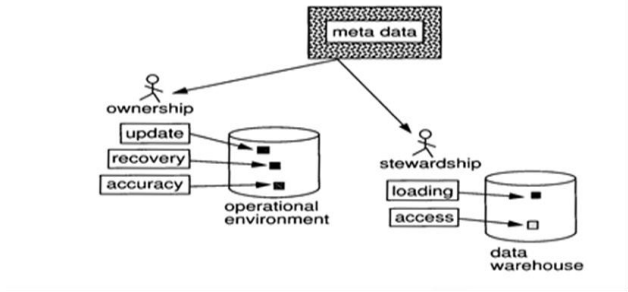

Fig3: Using the data warehouse-W.H. Inmon

- *Source of data in data warehouse* includes data from the operational system in an organization.
- *Transformation of data from operational system to data warehouse* is required to make data best for analysis.
- *The actual description of data in data warehouse* includes information about the tables present, fields present in the tables and physical characteristics of the fields.
- *Versioning to the storage of essential metadata over time* implies that metadata needs to be maintained overtime as data in the data warehouse is added in a continuous manner.
- *Data model information* helps the end users analysts for finding the useful data present in the data warehouse.
- *Development* is the constant process in data warehouse. Thus, information regarding track of work done in metadata is useful to the end users.
- *Security* is the essential part of data warehouse. End users must know how to access the protected data.
- *Loading schedule* includes information regarding time when the data was last refreshed which is essential to the end users analysts.

G. Components of Metadata

Keys and Attributes, Mapping information, Extract history are the key components of metadata.

- *Keys and Attributes*
  Data warehouse metadata comprises of the tables (keys and attributes) present in the data warehouse. Incase information regarding the keys and attributes of the tables is wrong, then it would result in data inaccuracy problems.
- *Mapping information*
  When the data is transferred from the from operational system to data warehouse ,we need to update information in data warehouse metadata respository.The mapping information comprises information about fields ,attributes, physical characteristics conversions, key changes etc. which must be updated in the data warehouse metadata store.
- *Extract history*
  It is the information about when the data was last refreshed in the data warehouse.

H. Metadata Repository

Metadata repository is like a general-purpose information directory that includes several enhancing functions.

Storage of metadata can be either

- Internal (i.e. in the same file)
- External (i.e. in the separate file).

Internal storage allows transferring metadata along with data which may cause easy manipulation of metadata.

External storage requires the use of URLs for linking metadata to their data.

Metadata may be stored in human readable format such as XML which is easier for the users to understand and edit it. However this format requires lots of storage space. Thus storing metadata in binary format increases transfer speed and saves space.

I. Important functions of metadata in data warehouse

- *Metadata is in data virtualization*

Metadata is often used in data virtualization servers and also database and application server. Metadata in these servers is stored as data repositories and explains the business objectives in various applications.

- *Granularity of metadata*

Granularity refers to the level of detail to which metadata can be structured. High level of granularity in metadata allows deeper level of technical manipulations whereas metadata with lower level of granularity will not provide detailed information. Granularity of metadata is responsible for both creation and maintenance.

V. ADVANTAGES AND LIMITATIONS OF METADATA

A. Metadata Advantages

Metadata provides various *advantages* to users as well as enterprise such as:

- *Consistency of definitions*

It resolves difference between terminology such as "clients" and "customers", "revenue" and "sales" thus bringing consistency of definitions.

- *Clarity of relationships*

When the relationship between entities that are stored throughout data environment has to be determined then metadata helps in resolving corresponding ambiguity and inconsistencies.

An example can be seen as: when a client declares a "beneficiary" in one application and this beneficiary is called a "participant" in any other application then for clarifying this situation help can be taken from metadata definition.

- *Metadata Lineage*

Data lineage is all about demonstrating the fact that where the data comes from, where it flows to and how it is transformed as it travels through an enterprise.

- *Metadata ensures accessibility*

Metadata can also be defined as the key that ensures that the resources will survive and continue to be accessible in the future.

- *Metadata in Spatial Data Infrastructure (SDI)*

Spatial data as its name specifies is the data that identifies the geographical location of features anywhere on earth including oceans, mountains etc. Users of spatial data obtain a great advantage of metadata as they are provided with the information about the purpose, quality and accuracy of spatial data sets and also perform important functions

B. Metadata disadvantages

- It is very labor intensive and time consuming to manage metadata for spatial data sets.
- Cost, reliability, subjectivity, lack of authentication and lack of interoperability with respect to syntax, semantics, vocabularies and languages are some of the major disadvantages of metadata.
- Certain tension and bridge has also been created by some of the metadata researches.
- The web metadata also brings about various disadvantages. Before a reasonable number of people will start using metadata to provide a better web classification a long time will pass. Also, no one can guarantee that a majority of web objects can properly be classified through metadata because metadata by its nature is an optional feature that makes writing of objects heavier and its usage cannot be imposed.
- Maintenance of embedded data with external metadata sources is quite difficult.
- When documentation of an interface is created using metadata then the interface is documented at quite a low level.
- It is not possible for the metadata to provide comprehensive descriptions which means that various search terms may return few or no results.

VI. CONCLUSION

The paper basically focuses on the metadata & its critical need for using, building, and administering the data warehouse. Metadata describes data warehouse from different point of view, Metadata is needed for finding the data sources, to understand the data extraction and transformation and to navigate through the contents and to retrieve information. This paper also explains about the role of metadata, examines the need of metadata, types of metadata, life cycle of metadata, objectives of metadata, components of metadata and important function of metadata in data warehouse. It also focuses on the advantages and disadvantages of metadata in data warehouse.

REFERENCES

[1] Cory Doctorow, "Metacrap: Putting the torch to seven straw-men of the meta-utopia", version 1.3: 26 August, 2001.
[2] Richard J Kachur,"Data warehouse Management Handbook", prentice Hall, 2000, pp-348.
[3] Kai M. Hüner, Borris Otto and Hubert Österle, "Collaborative management of business metadata", International Journal of Information Management, volume 31 issue 4, August, 2011, pp.366-373.
[4] Eric Klovning, "Metadata management in the support of data warehouse development", University of Wisconsin-Stout, July, 2008.
[5] Martin Staudt, Anca Vaduva and Thomas Vetterli, "The role of metadata for data warehousing", Technical Report, University of Zurich, 1999.
[6] Arun Sen, "Metadata management: past, present and future", Decision Support Systems, volume 37 issue 1, April 2004, pp. 151-173.
[7] Thanh N. Huynh, Oscar MangiSengi and A Min Tjoa, "Metadata for object-relational data warehouse", Data Mining and Data Warehousing 2000:3.
[8] Pete Johnston, "Metadata sharing and XML", *NOF-digitise Technical Advisory Service Information Paper*, January, 2001,
[9] Jane Hunter, "Working towards MetaUtopia- a survey of current metadata research", Library Trends, Organizing the internet, 52(2), fall 2003.
[10] Markus H¨anse, Min-Yen Kan and Achem P. Karduck, "Kairos: proactive harvesting of research paper metadata from scientific conference web sites", ICADL'10 Proceedings of the role of digital libraries in a time of global change, and 12th international conference on Asia-Pacific digital libraries, pp.226-235.
[11] Gad Kother, "Metadata for End-Users-A crucial part of your Data Warehouse quality policy", SAS Institute, January 2008.
[12] Hao Fan and Alexandra Poulovassilis, "Using AutoMed Metadata in data Warehousing environment", DOLAP '03 Proceedings of the 6th ACM international workshop on Data warehousing and OLAP, New York, NY, USA, 2003, pp. 86-93.
[13] Nayem Rahman, Jessica Marz and Shameen Akhter, "An ETL Metadata Model for data warehousing", Journal of Computing and Information Technology, volume 20, Number 2, 2012, pp. 95-111.

Arushi Jain is currently pursuing Bachelor of Technology in Information Technology from Northern India Engineering College, Guru Gobind Singh Indraprastha University, New Delhi, India.

Aakansha Bansal is currently pursuing Bachelor of Technology in Information Technology from Northern India Engineering College, Guru Gobind Singh Indraprastha University, New Delhi, India.

Palak Jain is currently pursuing Bachelor of Technology in Information Technology from Northern India Engineering College, Guru Gobind Singh Indraprastha University, New Delhi, India.

Neha Sharma is currently working as Assistant Professor in Northern India Engineering College, Guru Gobind Singh Indraprastha University, New Delhi, India. Her area of research is data warehousing and data mining.