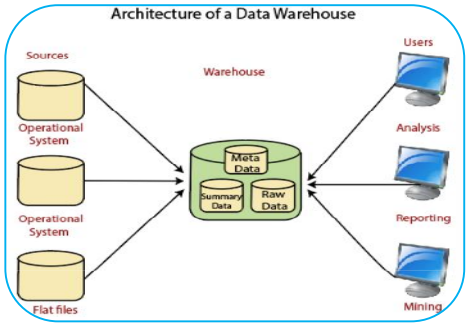
REVIEW OF RESEARCH
ISSN: 2249-894X
IMPACT FACTOR : 5.7631(UIF)
VOLUME - 11 | ISSUE - 11 | AUGUST - 2022

ARCHITECTURE OF QUALITY DATA MODELS IN DATA WAREHOUSE

Dr. M. D. Acharya[1] and Dr. Rohini Shinde[2]
[1]Yogeshwari Mahavidyalaya, Ambajogai
[2] Dayanand Science College,Latur.

ABSTRACTS

The computerization of our society has greatly increased our ability to both create and collect data from a variety of sources. Data on every aspect of our lives has grown exponentially. A large amount of data and useful information needs to be exchanged. This led to a promising and thriving frontier in computer science called data mining. Data mining is the automated or convenient processing of patterns representing knowledge stored or captured in large databases, data warehouses, the web, other large information stores, or data streams. Data mining can be applied to any type of data as long as the data is meaningful to the target application. In this paper, we discuss data warehouse and data warehouse data in detail, which is almost fundamental data for data mining applications. We also present to you a specific framework of data warehouse and data pre-processing techniques.



KEY WORDS : *computerization, data warehouse and data pre-processing techniques.*

INTRODUCTION

This tool allows you to integrate your data warehouse, data integration tools, data modeling tools, etc. Enables the creation of enterprise models based on metadata obtained from different systems. For example it will be able to resolve conflicts based on attribute names and types. You will be able to create custom metadata type "stitch" metadata between the two systems. A fully-fledged metadata management model provides a 360-degree view of how different systems are integrated into your organization. This model can be the starting point for any new data governance initiative. Data modelers will now have room to find specific properties and use them in their own data models. This model is also the foundation of the 'database' we talked about in the previous section. As with any other data governance initiative, as metadata in individual systems changes, the model needs to be updated according to the SDLC methodology that includes versioning, workflow, and approval. Access to the metadata model must also be managed by creating roles, privileges and policies. Implementing a data warehouse is a great solution for complex business intelligence applications. Data warehouses fail to meet expectations due to poor data quality. Once the data warehouse is built, the issue of data quality does not matter. Data quality is a critical issue in the implementation and management of data warehouses. Before the data can be used effectively in a data warehouse, data needs to be analyzed and cleaned. To leverage business intelligence, it is imperative to implement ongoing data cleaning processes and procedures and to know the level of data quality over time. A key area of data warehouse

failure is moving data from various sources into an easily accessible repository, i.e. data consolidation. Understanding the complete data in a data warehouse starts with data quality.

The purpose of data warehouse development for business application is to prioritize various processes for integrated development of information systems. When a data warehouse is defined, the main task is to define the granularity and different levels of abstraction of the data in the data warehouse that will support the best decision-making process. According to a recent study by the Standish Group, 83% of data warehouse projects significantly exceed their budget as a result of misunderstandings about source data and data interpretation. Similar surveys by the Gartner Group point to poor data quality as a contributing factor to failed projects.

CLASSIFICATION OF DATA QUALITY ISSUES:

Different data sources have different problems associated with them, such as data from legacy data sources that don't even have metadata that describes them. Sources of dirty data include data entry errors by human or computer systems, data update errors by human or computer systems. Some data comes from text files, is part of MS Excel files, and some data sources have a direct Open Data Base Connectivity (ODBC) connection to the database. Some files are the result of manual merging of multiple files so data quality can be compromised at any point.

Data warehousing is growing in popularity as organizations realize the benefits of decision-centric and business-centric data bases. However, there is a significant obstacle in the rapid development and implementation of a quality data warehouse, especially due to warehouse data quality issues at different stages of data warehousing. Cells, especially with quality data, pose a problem. Over time, many researchers have contributed to data quality problems, but no research has integrated all the causes of data quality problems according to all stages of data warehousing. 1) Data sources 2) Data collection and data profiling, 3) Data staging and ETL 4) Data warehouse modeling and schema design.

The final objective of the paper is to identify the reasons behind data scarcity, unavailability or capacity problems at all stages of data warehousing and to create a descriptive taxonomy of these reasons. We have explored the possible causes of data quality issues through an extensive literature review and with the advice of data warehouse practitioners working in reputed IT giants in India. We hope this will help developers and warehouse implementers examine and analyze these issues before making quality decisions and moving towards data integration and data warehouse solutions for business intelligence applications.

ABOUT DATA QUALITY:

The mere existence of data does not guarantee that all managerial tasks and decisions can be carried out easily. One definition of data quality is that it is about bad data – data that is missing or incorrect or, in some cases, invalid. The broad definition is that data quality is achieved when an organization uses data that is comprehensive, understandable, consistent, relevant and timely. Understanding the important dimensions of data quality is the first step to improving data quality. To be able to process and interpret effectively and efficiently, data must meet a set of quality criteria. Data that meet those quality criteria are said to be of high quality. Enormous efforts have been made to define and quantify data quality. Dimensions of data quality typically include accuracy, reliability, significance, consistency, accuracy, timeliness, accuracy, comprehensibility, conciseness, and applicability.

DATA QUALITY:

Depending on how data is entered, integrated, maintained and processed and loaded, data quality can be compromised. Data is affected by numerous processes that bring data into your data environment, most of which affect its quality. All these phases of data warehousing are responsible for the data quality in the data warehouse. Despite all efforts, some percentage of dirty data still exists. This residual deficit data should be reported with the reasons for failure to clean the data.

It is assumed that data quality issues can arise at any stage of data warehousing, from data sources to data integration and profiling, to data staging in ETL and database modeling. The following model describes possible states that are vulnerable to data quality issues. The concept of quality in this article is understood as a set of measurable and comprehensive characteristics of a product required by the customer for the development process of enterprise architecture. Product quality is monitored at all stages of its production, especially at so-called checkpoints. Control points in architecture development are points where you get different stages in the production cycle of architecture products.

CAUSES OF DATA QUALITY ISSUES AT DATA PROFILING STAGE:

Whenever possible, candidate data sources are identified and final data profiling is implemented immediately. Data profiling is the exploration and evaluation of the data quality, integrity, and consistency of your source system, sometimes called source system analysis. Data profiling is a fundamental task, but is often overlooked or minimized as the data quality of the data warehouse is compromised. At the start of a data warehouse project, once a candidate data source is identified, a quick data profiling assessment should be conducted to decide on the project going forward.

The data profiling (DP) function is recognized as one of the data quality process initiatives. The purpose of the data profiling process is to regularly detect the existence of errors, inconsistencies, redundancies and incomplete information in the data and associated metadata. After analyzing the data, the DP process generates a set of reports, which contain information about the data status. These reports enable analysts.

Evaluate whether the metadata accurately describes the actual values of the database. For example, a field may be defined as alphanumeric when it should be defined as numeric;

There is a clear idea of data quality. Information from the report can be used to find out whether data quality is being monitored. For example: the existence of many empty records may mean that the field or table is not complete (dimension complete);

Repair problem data with the Data Cleaning Tool (Data Cleaning). Most DP tools make rules to solve some problems, in addition to some errors that need to be fixed. It is advisable to follow the cycle of 'Detect, Rake' until the DP results are available.

Similar to a change in application requirements. While analyzing the data, if the detected errors are not understood by the analyst, then the application may not be properly designed, that is, the application is incompatible with the data in the DB. If the application presents problems, it is important to review the requirements and related business processes and make changes if necessary.

CONCLUSION:

In a data warehouse, data always moves from the operating system to the warehouse through the data storage area. Data extracted, integrated, refined, modified and loaded into a data warehouse means that data quality is always compromised. All of these steps are potential sources of data quality issues, and for this reason, all of these steps should be considered for potential data quality issues. In the model, each stakeholder has a so-called quality objective. These quality targets are abstract requirements defined on data warehouse objects and are documented for purposes of interest to stakeholders. The model consists of quality dimensions used to abstract different aspects of quality. Quality objectives relate to one or more quality questions. Quality defines whether a query goal is reached or not. A quality query is defined as a quality metric that reflects quality measurements and is defined for a specific data warehouse object. A quality metric also defines an interval of expected values in the domain and includes the actual value at a given point in time. A simple software agent measures the value of quality metrics. The system covers the entire data warehouse, from operating systems to analytical applications. Data quality is measured when data flows through a data warehouse system. Metadata is the core of the system, and the metadata of transformations, processes and data plans is the most important. The most important part of the concept is an integrated metadata management component that stores all information related to data quality.

REFERENCES:

1. Aravind Kumar (2016), 'Exploring the Application of Data Warehouse in Public Life', International Journal of Computer Science and Information Technologies, ISSN – 0975-9646, Vol-7, Issue-6, pp. 2544-2553
2. Abdullah A.S. Zina and Obaid Taleb A.S. (2016), 'Design and Implementation of Educational Data Warehouse Using OLAP', IJCSN International Journal of Computer Science and Network, ISSN 2277-5420, Volume 5, Issue 5, pp. 824-827.
3. Bianchi-Berthouze, N. and T. Hayashi. Subjective interpretation of complex data: requirements for supporting the mining process. In Systems, Man and Cybernetics, 2002 IEEE International Conference on 2002.
4. Bakar M.S.A., Taa A., Chit S.C., and Soid M.H.M. (2018), 'Data warehouse system for blended learning in institutions of higher education', Journal of e-Academy, Vol-6, Issue-2, pp. 144-155.
5. Banerjee, S., & Davis, K. C. (2009). Modeling data warehouse schema evolution over extended hierarchy semantics. In Journal on Data Semantics XIII, Springer, Berlin Heidelberg, pp. 72-96.
6. Bellahsene, Z. (2002). Schema evolution in data warehouses. Knowledge and Information Systems, Vol. 4(3), pp. 283-304.
7. Vishes S., Srinath Manu, Kumar K.C., and Nandan A.S. (2017), 'Data Warehousing Architecture and Pre-Processing', International Journal of Advanced Research in Computer and Communication Engineering', ISSN 2278-1021, Vol-6, Issue-5, pp. 13-18.