REVIEW OF RESEARCH
IMPACT FACTOR : 5.7631(UIF) UGC APPROVED JOURNAL NO. 48514 ISSN: 2249-894X

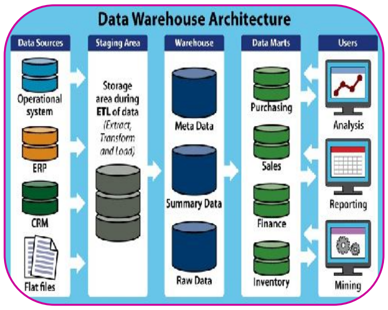
VOLUME - 7 | ISSUE - 11 | AUGUST - 2018

DATA WAREHOUSE OF QUALITY META MODELS AND IT'S ARCHITECTURE

Jayashrei Shivajirao Kolthe
Research Scholar

ABSTRACT

The computerization of our society has greatly enhanced our capabilities for both creating and collecting data from a variety of sources. Data on every element of our lives has grown enormously. There is a need to change the vast amount of data and useful information. This created a promising and flourishing frontier in computer science called data mining. Data mining is the automated or convenient processing of specimens representing knowledge stored or captured in large databases, data warehouses, the web, other massive information stores, or data streams. Data mining can be applied to any type of data as long as the data is meaningful for the target application. In this paper, we discuss in detail the data warehouse and data warehouse data, which is an almost basic type of data for data mining applications. We also present to you a specific framework of data warehouse and data pre-processing techniques.



KEY WORD: *society , promising and flourishing frontier.*

INTRODUCTION

This tool enables you to create an enterprise model based on metadata derived from different systems such as your data warehouse, data integration tools, data modelling tools, etc. For example will be able to resolve conflicts based on attribute names and types. You will be able to create custom metadata types "stitch" metadata between the two systems. A fully-built metadata management model provides a 360-degree view on how different systems are integrated into your organization. This model can be the starting point of any new data governance initiative. Data modelers will now have a place 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 initiatives, as the metadata in individual systems changes, the model needs to be updated according to the SDLC methodology which 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 the perfect solution for complex business intelligent applications. The data warehouse failed to meet expectations because of 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. Data needs to be analyzed and cleared before data can be used effectively in the data warehouse. To take advantage of Business Intelligence, it is imperative to implement ongoing data cleaning processes and procedures and to know the level of data quality over time. The main area for data warehouse failure is the move of data from a variety of sources into an easily accessible repository, i.e. integration of data. Understanding the complete data in the data warehouse begins with data quality. The purpose of the development of a data warehouse for a business application for the integrated development of information systems is to prioritize different processes. When a data warehouse is defined, the main task is to define the different levels of granularity and 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 largely overrun their budgets as a result of misunderstandings about source data and data interpretation. Similar surveys conducted by the Gartner Group point to the quality of the data and lead to failed projects.

CLASSIFICATION OF DATA QUALITY ISSUES:

Different data sources have different issues related to them, such as data from legacy data sources that do not even have the 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 of the data comes from text files, part of MS Excel files, and some data sources have direct Open Data Base Connectivity (ODBC) connection to the database. Some files are the result of manual consolidation of multiple files so that data quality can be compromised at any stage. Below are some reasons for a data quality problem with data sources.

- Insufficient selection of candidate data sources.
- Inadequate knowledge of interdependencies in data sources.
- Lack of validation routines on resources.
- Unexpected changes to source systems.
- Multiple data sources generate meaningful heterogeneity that can lead to quality problems.
- Presents conflicting information in data sources.
- Incompatible / incorrect data formatting.
- Multiple sources for the same data.
- Inconsistent use of special characters of data.

The data warehousing reputation is growing as organizations become aware of the benefits of a decision-oriented and business-oriented data base. However, there is an important obstacle to the rapid development and implementation of a quality data warehouse, especially because of warehouse data quality issues at different stages of data warehousing. Cells, especially with quality data, cause problems. Over the period of time many researchers have contributed to the problem of data quality, but all the causes of the data quality problem have not been consolidated into any research according to all stages of data warehousing. 1) Data Source 2) Data Collection and Data Profiling, 3) Data Staging and ETL 4) Data Warehouse Modeling and Schema Design. The ultimate purpose of the paper is to identify the reasons behind data shortages, unavailability or capacity problems at all stages of data warehousing, and to create a descriptive classification of these causes. We have explored possible causes of data quality issues through extensive literature review and as per 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 moving to quality decision-making and data integration and data warehouse solutions for business intelligence applications.

UNDERSTANDING OF DATA QUALITY:

The mere existence of data does not guarantee that all managerial functions and decisions can be easily carried out. One definition of data quality is that it's about bad data - data is missing or inaccurate or invalid in some respects. The broad definition is that data quality is achieved when an organization uses comprehensive, understandable, consistent, relevant and timely data. Understanding the key data quality dimensions is the first step in improving data quality. To be able to process and interpret in an effective and efficient manner, a set of quality criteria must be met for the data. Data that meets those quality criteria are said to be of high quality. Huge efforts have been made to define the quality of the data and to identify its dimensions. Data quality dimensions typically include accuracy, reliability, importance, consistency, accuracy, timeliness, accuracy, comprehensibility, concision and applicability.

- Completeness: Are all the necessary information available to make sure? Missing some data values or being useless?
- Consistency: Different instances of the same data event agree with each other or provide conflicting information. Are values consistent across data sets?
- Validity: Refers to the correctness and reasonableness of data
- Conformity: Are there any expectations that the data values are consistent with the specified format? If so, are all values consistent with that format? It is important to keep a specific look.
- Accuracy: Do data objects accurately represent the "real world" values expected from the model? Misspellings of product or person names, addresses, and timely or recent data may affect operational and analytical applications.
- Integrity: Which data key is missing the link? The inability to link related entries together can detect the reality of duplication on your system.

QUALITY OF DATA:

Depending on how the data is received entered, integrated, maintained, and processed and loaded, the quality of the data may be compromised. Data is impacted by the numerous processes that bring data into your data environment, most of which affect its quality. All these stages of data warehousing are responsible for the data quality in the data warehouse. Despite all the effort, a few percent of the dirty data still exists. These residual deficit data should be reported, stating the reasons for failure to clear the data. Data quality issues can come in many different ways. The most common include.....

- Poor data handling procedures and processes.
- Failure to stick on to data entry and maintenance procedures.
- Errors in the migration process from one system to another.
- External and third-party data that may not fit with your company data standards or may otherwise be of unconvinced quality.

The assumption taken is that data quality questions can arise at any stage of data warehousing, in data sources in data integration and profiling, in data staging in ETL and database modelling. The following model describes potential states that are vulnerable to data quality problems. The quality concept in this article is understood as a set of measurable and vast features of a product that a customer needs 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 issues where you get different stages in the product cycle of architecture products.

CAUSES OF DATA QUALITY ISSUES AT DATA PROFILING STAGE:

Whenever possible, candidate data sources are identified and final data profiling comes into play immediately. Data profiling is an investigation and evaluation of the data quality, integrity and consistency of your source system, sometimes called source system analysis. Data profiling is a fundamental one, but it is often overlooked or undermined so that the data quality of the data warehouse is compromised. At the beginning of the data warehouse project, as soon as a candidate's data source is identified, a quick data profiling assessment should be made to decide on the project going forward.

The data profiling (DP) function is known as one of the undertakings of the data quality process. The purpose of the data profiling process is to regularly detect the existence of errors, inconsistencies, redundancies, and incomplete information in the data and related metadata. After analyzing the data, a set of reports is generated in the DP process, which contains information about the data state. These reports enable analysts.

- Evaluate whether the metadata accurately describes the actual values of the database. For example, a field can be defined as alphanumeric when it should be defined as numeric;
- There is a clear idea of data quality. The information in the report can be used to find out if the quality of the data is being monitored. For example: the existence of too many empty records can mean that the field or table is not complete (dimension completion);
- Correct problem data with the Data Cleaning Tool (Data Cleaning). Most DP tools create rules to solve some of the problems found, in addition to pointing out a set of errors that need to be resolved. It is advisable to follow the cycle 'Detect, rake' till the results of the DP reports.
- Equal to the change in application requirements. When analyzing the data, if the errors found are not understood by the analyst, then the application may not be well designed, that is, the application is incompatible with the data in the DB. If the application is presenting a problem, it is important to review the required items and related business processes and make changes if necessary;

CONCLUSION:

In a data warehouse, data always goes from the operating system to the warehouse through the data storage area. Data extracted, integrated, refined, modified, and loaded in a data warehouse means that the quality of the data is always compromised. All of these stages are potential sources for data quality problems, and for this reason, all of these steps should be considered for potential data quality problems. In the model, there is a so-called quality objective for each stakeholder. These quality targets are abstract requirements defined on data warehouse objects and are documented for the purpose for which stakeholders are interested. The model has quality dimensions used to abstract different aspects of quality. Quality targets are related to one or more quality questions. Quality query defines whether or not to reach a goal. Quality query is defined as a quality metric that reflects quality measurements and is defined for a specific data warehouse object. The quality metric also defines the interval of expected values in the domain and includes the actual value of a given point in time. Simple software agent calculates 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 main aspect of the system and the metadata of conversions, processes and data schemes is the most important. The most important part of the concept is the integrated metadata management component that stores all the information related to data quality.

REFERENCES:

1. 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.
2. 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
3. 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.
4. 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.

5. 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.
6. 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.
7. Bellahsene, Z. (2002). Schema evolution in data warehouses. Knowledge and Information Systems, Vol. 4(3), pp. 283-304.