Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 5, May 2015, pg.1025 – 1029

RESEARCH ARTICLE

A NEW DATABASE FOR RETAIL DOMAIN USING ETL SYSTEM

Sagar Bhujbal, Dhanesh Gite, Yadnesh Kadam, Bhushan Narkhede
Computer Engg & University of Pune, India
Computer Engg & University of Pune, India
Computer Engg & University of Pune, India
Sagar.bhujbal1@gmail.com, ygk369@gmail.com, bmn8060@gmail.com

Abstract -- The software processes that facilitate the original loading and the periodic refreshment of the data warehouse contents are commonly known as Extraction-Transformation-Loading (ETL) processes. The intention of this survey is to present the research work in the field of ETL technology in a structured way. To this end, we organize the coverage of the field as follows:
(a) First, we cover the conceptual and logical modeling of ETL processes, along with some design methods
(b) We visit each stage of the E-T-L triplet, and examine problems that fall within each of these stages
(c) We discuss problems that pertain to the entirety of an ETL process, and
(d) We review some research prototypes of academic origin.
Keywords: extraction, transformation and loading, data warehouses, data mart, online analytical processing, online transaction protocol

I. INTRODUCTION

To Built a secondary "**New Orders Data Base**" for a retail company or organization having a separate ETL system using Informatica which will help to update the database by daily data. The software processes that facilitate the original loading and the periodic refreshment of the data warehouse contents are commonly known as Extraction-Transformation-Loading (ETL) processes. The intention of this survey is to present the research work in the field of ETL technology in a structured way.

ETL Process:

During the ETL process, data is extracted from an OLTP database, transformed to match the data warehouse schema, and loaded into the data warehouse database. Many data warehouses also incorporate data from non-OLTP systems, such as text files, legacy systems, and spreadsheets; such data also requires extraction, transformation, and loading.

In its simplest form, ETL is the process of copying data from one database to another. This simplicity is rarely, if ever, found in data warehouse implementations; in reality, ETL is often a complex combination of process and technology that consumes a significant portion of the data warehouse development efforts and requires the skills of business analysts, database designers, and application developers.

When defining ETL for a data warehouse, it is important to think of ETL as a process, not a physical implementation. ETL systems vary from data warehouse to data warehouse and even between department data marts within a data warehouse. A monolithic application, regardless of whether it is implemented in Transact-SQL or a traditional programming language, does not provide the flexibility for change necessary in ETL systems. A mixture of tools and technologies should be used to develop applications that each performs a specific ETL task.

The ETL process is not a one-time event; new data is added to a data warehouse periodically. Typical periodicity may be monthly, weekly, daily, or even hourly, depending on the purpose of the data warehouse and the type of business it serves.

Because ETL is an integral, ongoing, and recurring part of a data warehouse, ETL processes must be automated and operational procedures documented. ETL also changes and evolves as the data warehouse evolves, so ETL processes must be designed for ease of modification. A solid, well-designed, and documented ETL system is necessary for the success of a data warehouse project.

Data warehouses evolve to improve their service to the business and to adapt to changes in business processes and requirements. Business rules change as the business reacts to market influences—the data warehouse must respond in order to maintain its value as a tool for decision makers. The ETL implementation must adapt as the data warehouse evolves. Microsoft® SQL Server™ 2000 provides significant enhancements to existing performance and capabilities, and introduces new features that make the development, deployment, and maintenance of ETL processes easier and simpler, and its performance faster.

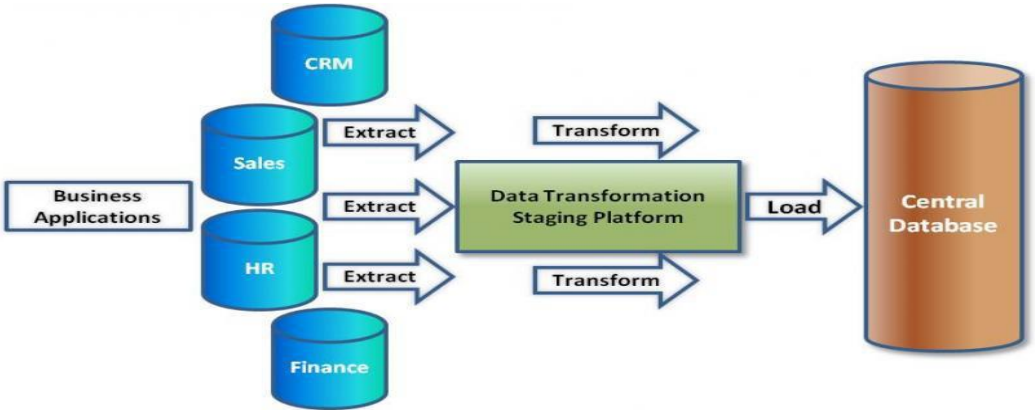


Fig No. 1.2.1 ETL Process

ETL TOOL:

INFORMATICA

Informatica Corporation is a software development company. Founded in 1993, it's headquartered in Redwood. Informatics' product is a portfolio focused on Data Integration: ETL, Information Lifecycle Management, B2B Data Exchange, Cloud Data Integration, Complex Event Processing, Data Masking, Data Quality, Data Replication, Data Virtualization, Master Data Management, Ultra Messaging; currently at version 9.6. These components form a toolset for establishing and maintaining enterprise-wide data warehouses. It has a customer base of over 5,000 companies.

II. PROBLEM DEFINITION

ETL systems vary from data warehouse to data warehouse and even between department data marts within a data warehouse. A mixture of tools and technologies should be used to develop applications that each performs a specific ETL task.

Models of ETL Process:

This section will navigate through the efforts done to conceptualize the ETL processes. Although the ETL processes are critical in building and maintaining the DW systems, there is a clear lack of a standard model that can be used to represent the ETL scenarios. After we build our model, we will make a comparison between this model and models discussed in this section. Research in the field of modeling ETL processes can be categorized into three main approaches:

1. Modeling based on mapping expressions and guidelines.

2. Modeling based on conceptual constructs.

Modeling ETL process using mapping expressions

We have defined a model covering different types of mapping expressions. They used this model to create an active ETL tool. In their approach, queries are used to achieve the warehousing process. Queries will be used to represent the mapping

between the source and the target data; thus, allowing DBMS to play an expanded role as a data transformation engine as well as a data store. This approach enables a complete interaction between mapping metadata and the warehousing tool. In addition, it addresses the efficiency of a query-based data warehousing ETL tool without suggesting any graphical models. It describes a query generator for reusable and more efficient data warehouse (DW) processing

ETL Conceptual Model

In this section, we focus on the conceptual part of the definition of the ETL process. For a detailed presentation of our conceptual model and formal foundations for the representation of ETL processes, we refer the interested reader to [29]. This model has a particular focus on (a) the interrelationships of attributes and concepts, and (b) the necessary transformations that need to take place during the loading of the warehouse. The latter part is directly captured in the proposed metamodel as a first class citizen; we employ *transformations* as a generic term for the restructuring of schema and values or for the selection and even transformation of data. Attribute

III. LITERATURE SURVEY

This chapter aims to present the literature review, a study about data warehouse technology, specifying concepts, characteristics and different types of data warehouses and data marts architectures.

Different researchers from different areas (database management, information system design, data and information integration) have come out with their own conclusions. As Mull (1983) observes:

"We must be prepared to learn more than we can understand."

Thus, there are many sources that could be quoted to illustrate the research methods used to understand data warehousing and integration concepts. The work summarized here is based on relevant literature review and on research performed in Norway and Mozambique.

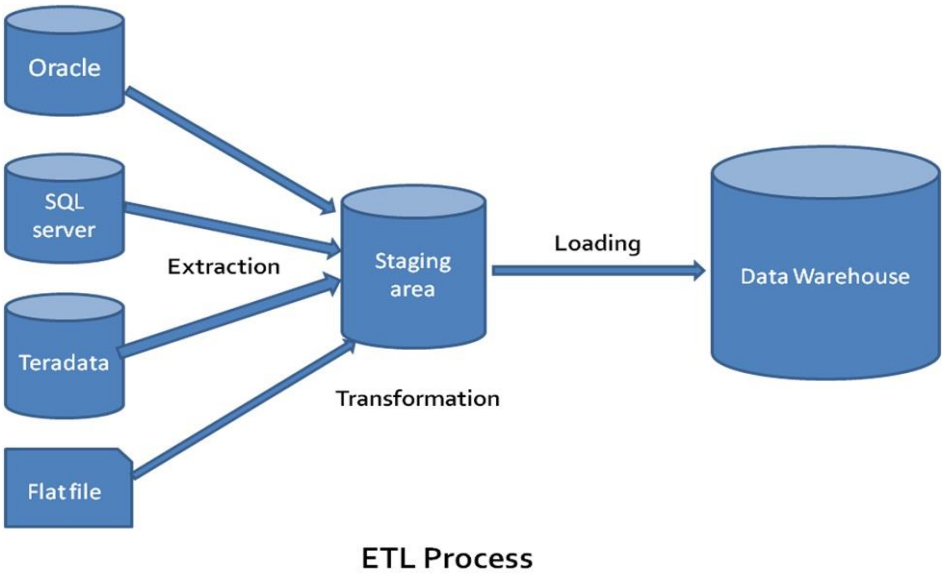


Fig 2.1: Data warehouse Environment

A data warehouse (DW, DWH), or an enterprise data warehouse (EDW), is a system used for reporting and data analysis. Integrating data from one or more disparate sources creates a central repository of data, a data warehouse (DW). Data warehouses store current and historical data and are used for creating trending reports for senior management reporting such as annual and quarterly comparisons.

The data stored in the warehouse is uploaded from the operational systems (such as marketing, sales, etc., shown in the figure to the right). The data may pass through an operational data store for additional operations before it is used in the DW for reporting

IV. OVERVIEW OF FRAMEWORK

During the ETL process, data is extracted from an OLTP database, transformed to match the data warehouse schema, and loaded into the data warehouse database. Many data warehouses also incorporate data from non-OLTP systems, such as text files, legacy systems, and spreadsheets; such data also requires extraction, transformation, and loading.

In its simplest form, ETL is the process of copying data from one database to another. This simplicity is rarely, if ever, found in data warehouse implementations; in reality, ETL is often a complex combination of process and technology that consumes a significant portion of the data warehouse development efforts and requires the skills of business analysts, database designers, and application developers.

When defining ETL for a data warehouse, it is important to think of ETL as a process, not a physical implementation. ETL systems vary from data warehouse to data warehouse and even between department data marts within a data warehouse. A monolithic application, regardless of whether it is implemented in Transact-SQL or a traditional programming language, does not provide the flexibility for change necessary in ETL systems. A mixture of tools and technologies should be used to develop applications that each performs a specific ETL task.

The ETL process is not a one-time event; new data is added to a data warehouse periodically. Typical periodicity may be monthly, weekly, daily, or even hourly, depending on the purpose of the data warehouse and the type of business it serves. Because ETL is an integral, ongoing, and recurring part of a data warehouse, ETL processes must be automated and operational procedures documented. ETL also changes and evolves as the data warehouse evolves, so ETL processes must be designed for ease of modification. A solid, well-designed, and documented ETL system is necessary for the success of a data warehouse project.

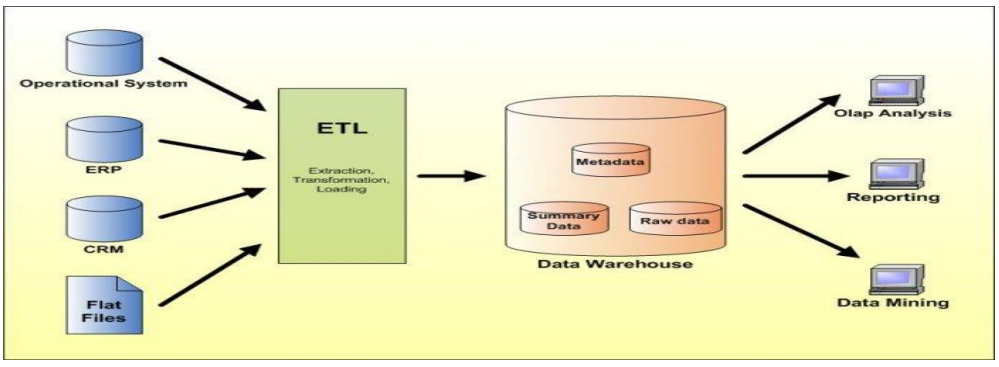
Data warehouses evolve to improve their service to the business and to adapt to changes in business processes and requirements. Business rules change as the business reacts to market influences—the data warehouse must respond in order to maintain its value as a tool for decision makers. The ETL implementation must adapt as the data warehouse evolves. Microsoft® SQL Server™ 2000 provides significant enhancements to existing performance and capabilities, and introduces new features that make the development, deployment, and maintenance of ETL processes easier and simpler, and its performance faster.

V. DATA WAREHOUSE ETL MODEL

Data Model

In the literature we can find different approaches about data modeling. Although exists more than one model to construct a data warehouse with success, dimensional modeling becomes more effective for a data warehouse project (Domenico, 2001). In this research, I present in summarized way, Kimball dimensional model of data.

Complex questions that involve organization business analysis usually require a vision of the data from different perspectives. Answers to this type of questions can lead to correct or wrong decisions. Tools based on Structured Query Language (SQL) help in the data search related to this type of queries

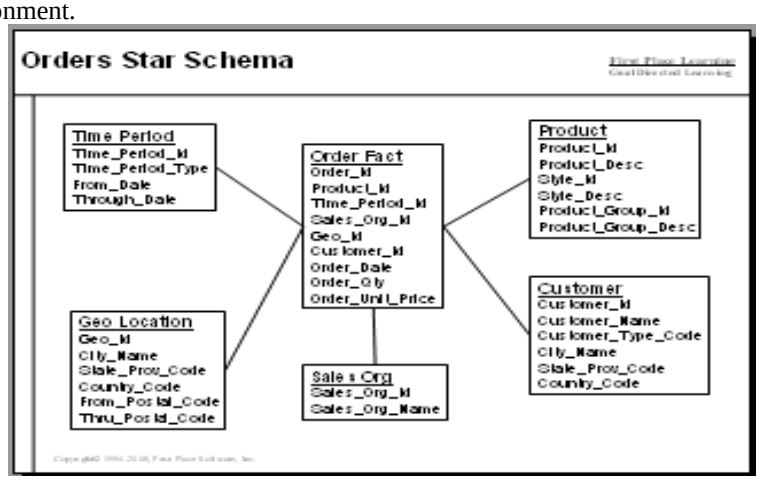


2.4.1 Data Model

Star Model

Valente (1996), in his research, states that, traditionally, data model of relational databases presents tables with complex relationships and with multiple unions. For most users using tools to compose their queries, it is necessary that the access to the database is simple to facilitate the direct access to the database.

The main type of sales & org model is called Star Model, where a dominant table exists at the center of the model. The table in the center is called order fact. This table has multiple junctions connecting with other tables called product, customer tables. Each secondary table has only one junction with a order fact table.
The star model presented in the Figure 3.7 has the advantage of being simple and Intuitive. For Kimball (1996), entity-relationship model is not adjusted to data analysis in the management environment. The dimensional model is the most appropriate for this Environment.



2.4.2 Star Model

CONCLUSION

1) We have built a secondary "**New Orders Data Base**" for Retail Organization having a separate ETL system which will help to update the database by history data and daily data.
2) It gives backup of history data for main system.
3) If any disaster occurs we can recover data from backup.
4) High availability.
5) There is no overriding data.

REFERENCES

- Shaker H. Ali EI- Sappagh*a, Abdeltawab H. Ahmed Hendawi*b, Ali Hamed EI Bastawissy*b, " A Proposed Model For datawarehouse ETL Process", Journal of King Saud University-Computer and information Sciences (2011)23, 91-104
- Qin Hanlin, Jin Xianzhen, Zhang Xianrong, "Research on Extract, Transform and load in Land and Resources Star Schema Data Warehouses", Computational Intelligence and Design (ISC10), 2012 fifth International Symposium on (Volume 1), 28-29 Oct.2012,Pages 120-123
- Thomas Van Raalte, " Introduction to Oracle Retail data model Implementation and Operation Guide" , Release 11.3.2 E20363-03, January 2013
- Ponas Vassiliadis, "A survey of Extract-transform-load technology", International Journal of data warehousing & mining, 5(3), 1-27, July September 2009 1.

