Advanced techniques of Data Warehousing for Business Intelligence Using State-of-the-Art ETL Tool

*** Prof. Babali Rai**

Abstract

Data warehouse is a physical repository where relational data are organized to provide enterprise-wise, cleansed data in a standardized format. In Data Warehouse (DW) environment, Extraction-Transformation-Loading (ETL) processes constitute the integration layer which aims to pull data from data sources to targets, via a set of transformations. ETL is responsible for the extraction of data, their cleaning, conforming and loading into the target. Data warehouses are traditionally refreshed in a periodic manner, most often on a daily basis. Thus, there is some delay between a business transaction and its appearance in the data warehouse. The most recent data is trapped in the operational sources where it is unavailable for analysis. The intention of this survey is to present the research work in the field of ETL technology in a structured way to create and maintain a Data Warehouse. The paper shows that (1) the conceptual and logical modeling of ETL processes, along with some design methods (2) review of open source and commercial ETL tools, along with some ETL prototypes coming from academic world. (3) Proposes to avoid refreshment anomalies through various ETL Tools (like Informatica,Talend etc.),which will help to accelerate the pace of development for future data marts.

Key words: Business Intelligence, Data Mart, Data Warehousing Concept, OLAP, OLTP, ETL Tools.

Introduction

The explosive growth in volume of information with over 85% of the information in unstructured format, the complexity arising from disparate information sources and the nature of information present severe information management challenges. It is no longer access to information that determines winners; rather it is the application of information to solve business problems. Despite substantial investments, businesses are unable to leverage information fully. DW/BI (Data WareHouse/Business Intelligence) encompasses a wide range of tools, technologies and platforms, such as analytics, data mining and intelligent search, to enable organizations to share, use and uncover the value hidden in their information assets. Used effectively, DW/ BI can improve the quality of decision making by presenting relevant information in easy to understand reports, dashboards and scorecards.

Business Intelligence

Business intelligence (BI) is the set of techniques and tools for the transformation of raw data into meaningful and useful information for business analysis purposes. Getting the right information to the right people at the right time – that's still the goal of everyday business intelligence. At the same time, the term business intelligence (BI) represents the tools and systems that play a key role in the strategic planning process of the corporation. These systems allow a company to gather, store, access and analyze corporate data to aid in decision-making.Data Warehouse is known as a tool/ Business Intelligence Software, designed with the primary goal of extracting important data from an organization's raw data to reveal insights to help a business make faster and more accurate decisions. The software typically integrates data from across the enterprise and provides end-users with self-service reporting and analysis. BI software uses a number of analytics features including statistics, data and text mining and predictive analytics to reveal patterns and turn information into insights.

Business Intelligence that has three key advantages

1. Business intelligence systems not only support the latest information technologies, but also provide prepackaged application solutions.
2. Business intelligence systems focus on the access and delivery of business information to end

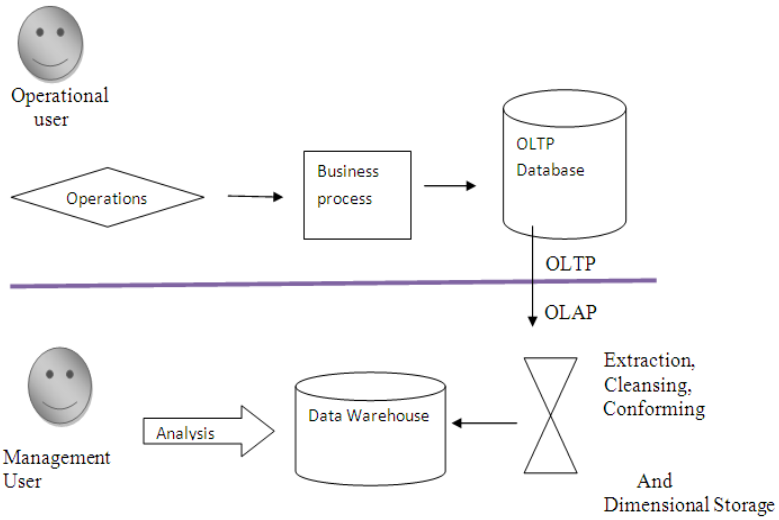
users, and support both information providers and information consumers.
3. Business intelligence systems support access to all forms of business Information, and not just the information stored in a data warehouse.

Data Mart

Data Mart is a specialized subset of Data Warehouse. It Often holds only one subject area- for example, Finance, or Sales. Data marts are often built and controlled by a single department within an organization. Given their single-subject focus, data marts usually draw data from only a few sources. The sources could be internal operational systems, a central data warehouse, or external data.Data Mart concept can be better explained with the *fig 1*.

Data Warehousing

A Data warehouse is a composite and collaborated data model that captures the entire data of an organization. It brings together data from heterogeneous sources into one single destination.
Inmon; Father of Data warehousing defines a Data warehouse as subject oriented, integrated, non-volatile & time variant collection of data in support of Management Decision.Generally, architecture for systems based on DW involves the integration of current and Historical data. It is a huge repository of data that does not tell us much by itself; like in the operational databases, we need auxiliary tools to query and analyze data stored. Data is Extracted, Transformed and Loaded (ETL) into the Data warehouse with a tool, called ETL tool. But from where does the data warehouse get its data? The data is derived from the operational system that supports the basic business processes of the organization .In between the operational systems and the data warehouse; there is a data staging area. In this staging area, the operational data is cleansed and transformed into a form suitable for placement in the data warehouse for easy retrieval. DM/DW can be accessed by OLAP (Online Analytical Processing) or Data Mining tools and/or DSS (Decision Support Systems). These tools make the data navigation possible, as well as the managerial analysis and the knowledge discovery.



OLTP Vs. OLAP System

On Line Transaction Processing describes processing of short and simple transaction data at operational sites i.e. day to day operations in the Source systems. The Database is designed as Application-oriented (E-R based) i.e Highly Normalized so as to efficiently support INSERT And UPDATE operations. Data stored in these systems are raw Current (Up-to-date) and Isolated Data, in a much Detailed level in flat relational tables. Generally OLTP systems are designed as normalized where every column in a tuple is related to the unique identifier and only the unique identifier.
systems use the primary-secondary key relationship to relate entities (tables) to each other. OLTP systems are usually created for a specific use such as order processing, ticket tracking, or personnel file systems.

On Line Analytical Processing describes processing at the Centralized, Integrated and Consistent Data Warehouse. It acts as the Decision Support System for the Business End Users. The Database is designed as Subject-oriented (Star/Snowflake Schema) i.e. highly denormalized to support the SELECT operations. Data in these systems are generally Consolidated, Summarized and Historical Data in nature.O LAP Queries will give aggregated information about the things happened in the past over a period of time and this will help the management in strategic decision making. An overview of the BI concepts and technologies with regard to their process-oriented aim and their orientation can be seen in *Fig 2*.

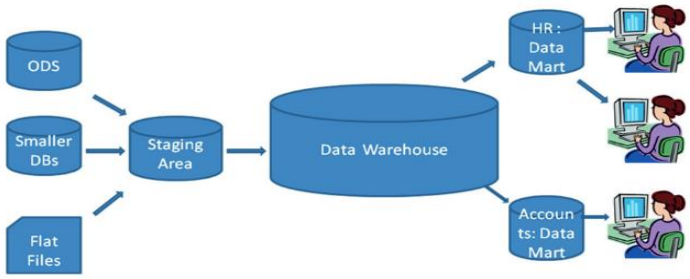


Figure 1: Data warehouse – The big picture

Fig 2: An Overview of Business Intelligence Process steps.

ETL-(Extract/Transform/Load)

Suppose an organization having several decades of history & operations in multiple domains and industries. Over a period, the company went through various changes in both internal operations and external customer interactions. Moreover, to achieve these changes, the company acquired several specialized tools and systems from different technological background and vendors. This introduces the challenge of creating a holistic view of company as the data to calculate various performance metrics reside in discrete and disconnected systems. There are two ways to address this challenge:

- Pull data from all these systems manually, then analyze via manual means. This method is good only if the analysis is needed once in a while, however if some reports are required every day at close of business, every week, every month etc.then for all practical purposes manual analysis is not a workable solution.

- Pull data from all these systems via automated means, convert data in a format that supports analysis, and save it for future references/reuse.

The second solution above is what we call **Extract**, **Transform** and **Load** (ETL).

An organization needs to load data warehouse regularly so that it can serve its purpose of facilitating business analysis. To do this, data from one or more operational systems needs to be extracted and copied into the data warehouse. The challenge in data warehouse environments is to integrate, rearrange and consolidate large volumes of data over many systems, thereby providing a new unified information base for business intelligence.

There are so many ETL tools available to save the time and make the whole process more reliable.

ETL Processes: Let us briefly describe each step of the ETL process with the fig. given Here.

Extract: The main objective of the extract step is to retrieve all the required data from the source system with as little resources as possible. The extract step should be designed in a way that it does not negatively affect the source system in terms or performance, response time or any kind of locking.

Clean: The cleaning step is one of the most important as it ensures the quality of the data in the data warehouse. Cleaning should perform basic data unification rules, such as:

- Making identifiers unique (sex categories Male/Female/Unknown, M/F/null, Man/Woman/Not Available are translated to standard Male/Female/Unknown)
- Convert null values into standardized Not Available/Not Provided value
- Convert phone numbers, ZIP codes to a standardized form
- Validate address fields, convert them into proper naming, e.g. Street/St/St./Str./Str
- Validate address fields against each other (State/Country, City/State, City/ZIP code, City/Street).
- Transform: This includes converting any measured data to the same dimension (i.e. conformed dimension) using the same units so that they can later be joined. The transformation step also requires joining data from several sources, generating aggregates, generating surrogate keys, sorting, deriving new calculated values, and applying advanced validation rules.

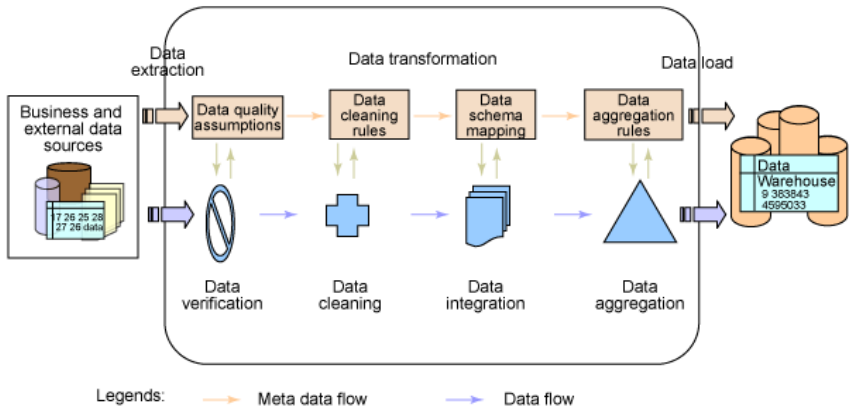
Load: During the load step, it is necessary to ensure that the load is performed correctly and with as little resources as possible. The target of the Load process is often a database.

ETL Tool Implementation

Designing and maintaining the ETL process is often considered one of the most difficult and resource-intensive portions of a data warehouse project. Many data warehousing projects use ETL tools to manage this process.

Besides the support of extraction, transformation, and loading, there are some other tasks that are important for a successful ETL implementation as part of the daily operations of the data warehouse and its support for further enhancements.

For ex. Oracle is not an ETL tool and does not provide a complete solution for ETL. However, Oracle does provide a rich set of capabilities that can be used by both ETL tools and customized ETL solutions. Oracle offers techniques for transporting data between Oracle databases, for



transforming large volumes of data, and for quickly loading new data into a data warehouse.

Depending on the needs of customers there are several types of tools. One of them perform and supervise only selected stages of the ETL process like data migration tools(**EtL Tools** , *"small t"*tools) , data transformation tools(**eTl Tools** , *"capital T"*tools).Another are complete (**ETL Tools**) and have many functions that are intended for processing large amounts of data or more complicated ETL projects.

There are two more types. First called **code base tools** is a family of programming tools which allow working with many operating systems and programming languages. The second

one called **GUI base tools** remove the coding layer and allow to work without any knowledge (in theory) about coding languages.

Existing work on ETL

These are many ETL tools available in the market however this tool are expensive and does not support small size businesses. ETL tools don't provide any capability to enterprise apart from moving the data.

These tools also have some common problems like –**Data Dependency, Lack of good quality of Data and complexity of source code.**

Proposed work on ETL (ETL Tools – What problem they solve?)

These days there are so many ETL tools available in market. These are of two categories

1. Commercial Tools like-IBM Infosphere DataStage,Informatica PowerCenter,Oracle Warehouse Builder OWB.
2. Freeware (open-sources) ETL Tools like -Pentaho Data Integration (Kettle),Talend Integrator Suite,Clover ETL.

Commercial tools are too costly to purchase by small organizations. So here is a solution to use open source/freeware tools. Though open source ETL tools have some limitations hence they are less used in industrial world. Approx ten open source ETL tools are presented and among that the most notable are Talend and Kettle because of their large users' community their literature, their features and their inclusion in BI suites.

Here we will see the surprising effects of ETL Tools on enterprise IT

- **In multiple solutions one was ETL tools** because ETL Tools makes it easy to pull transform and share the data across the applications.
- Merger or acquisition: Merger or acquisition leads to a position where IT application data need to be integrated or disintegrated.ETL Tools can give data to both the application as per their needs without many efforts. Because of this both the application need not to be changed and application code can be re-used. New code development can be avoided. Moreover it gives accuracy as you have not touched the application core algorithms.
- Enterprise data warehouse: ETL tools are capable of reading different data from different sources to integrate (relate) them. These upstream system works as a data feeding system. This data is later used by business to come to some conclusion on different aspects. This is called reporting systems in data warehouse. So we can say now that reporting systems provides different aspects of data which helps to start any initiative which is worth for business.
- Forecasting application: You have a production unit and you want to forecast demand for the next quarter, so your vendor can supply the necessary stocks. Here ETL can pull data from different manufacturing units, stores, warehouses and after consolidation, it can give a rough picture about how much order needs to be placed for the next quarter.

Conclusion

This article examines the actual concepts of Data Mart and Data Warehouse. Though Data Warehouse is used since early 1990s, the design and implementation is still dependent on the architecture and size of application and organization due to the complexity and cost of ETL tools. ETL processes are expensive regarding time, money and effort. It consumes up to 70%

of resources. Due to its importance, in this paper I try to provide best of my knowledge about ETL tools, the first approach is to review on Freeware /open source ETL tools along with some ETL prototypes coming from academic world. Namely SIRIUS, ARKTOS, PYGMATEL, DWPP. Also, *Talend Open Studio* and Microsoft ETL solution (SSIS), respectively and second is how companies can get beneficial by using ETL tools to promote Business Intelligence.

References

- *Thomas J¨org and Stefan Dessloch:Near Real-Time Data Warehousing Using State-of-the-Art ETL Tools,2009*
  http://www.infosys.com/consulting/information- management/Documents/data-warehousing-business
- *Oscar Romero, Alberto Abelló: Multidimensional Design Methods for Data Warehousing*
- *Neeraj Sharma,Abhishek Iyer,Rajib Bhattacharya,Niraj Modi,Wagner Crivelini:GETTING STARTED W I T H Data Warehousing.*
- *https://docs.oracle.com/cd/B19306_01/server.102/b14223/ettover.htm*
- *http://en.wikipedia.org/wiki/Extract,_transform,_load*