DATAWAREHOUSING AND ETL PROCESSES: An Explanatory Research

Priyanshu Gupta
ETL Software Developer
United Health Group

Abstract- In this paper, the author has focused on explaining Data Warehousing and the involved ETL processing. The focus has also been laid on the importance and significance of Data Warehousing for an organization. This paper attempts to describe an approach used for the migration of historical and current data of an organization to a DW product. So, this study will be substantially fruitful for understanding the concept of DW.

Index Terms- Data Warehousing, Data Integration, OLTP, ETL, Datastage

I. INTRODUCTION:

As a warehouse is a place to store objects, thereby a Data Warehouse is a place to store data. It is the central point of data integration for **business intelligence** i.e. for delivering a common view of enterprise data. Almost a decade of research has been spent on the study of data warehouses, especially for its design and exploitation for decision-making purposes. The author of this paper has reviewed all the processes involved and explained the things.

According to W.H. Inmon [1], a leading architect in the construction of data warehouse systems, "A data warehouse is a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management's decision making process". As the words are self- explanatory, DW in nothing more than a collection of data objects.

II. NEED OF A DATA WAREHOUSE

We all know that it is an era of intense competition. It is not just enough to take smart decisions but those strategic decisions should be taken on time. If the business miss taking an important decision on time, somebody else will already be there posing on them.

Taking into account a real time example, whenever you go to a bank, every transaction is recorded into OLTP systems. Imagine the number of transactions taking place in a day considering all the bank branches. And the transactions performed in an ATM are also stored in the same OLTP system. Now, if different end-users want to hit queries on such kind of system, imagine the load that will be put on the system. So, this system can't be used for querying purposes, it can only be used for transactional purposes. And here comes the need of a Data Warehouse which can be used for analytical purposes.

So, this **analytical data model** is business friendly, and a single source of origination of data, leading to better faith in data by the business. Moreover, operational and analytical data got segregated, thus maintaining high performance structure at both the ends.

III. TO TRANSFOR OLTP TO OLAP

This challenging transformation from OLTP to OLAP systems requires adhering to the following processes on a big scale:

A. Data Merging
Since data is store in multiple desperate sources, data from multiple OLTP systems are merged into a single OLAP system. The merge process must resolve differences in encoding between the different OLTP systems.

B. Data Scrubbing
This includes removing all the inconsistencies in the data from multiple sources, before it can be loaded into the data warehouse for use by the OLAP system. This transformation gives you an opportunity to scrub data.

C. Data Aggregation
Since OLTP systems record all the transactional details and OLAP typically need summary data, so, data is aggregated in some fashion that depends on a number of business factors, such as the speed requirements of your OLAP queries and the level of granularity required for your analysis.

D. Data Storing
While moving the data into an OLAP system, it must be transformed into an organization that better supports decision support analysis. The process of building a DW involves reorganizing OLTP data stored in relational tables into OLAP data stored in multidimensional cubes.

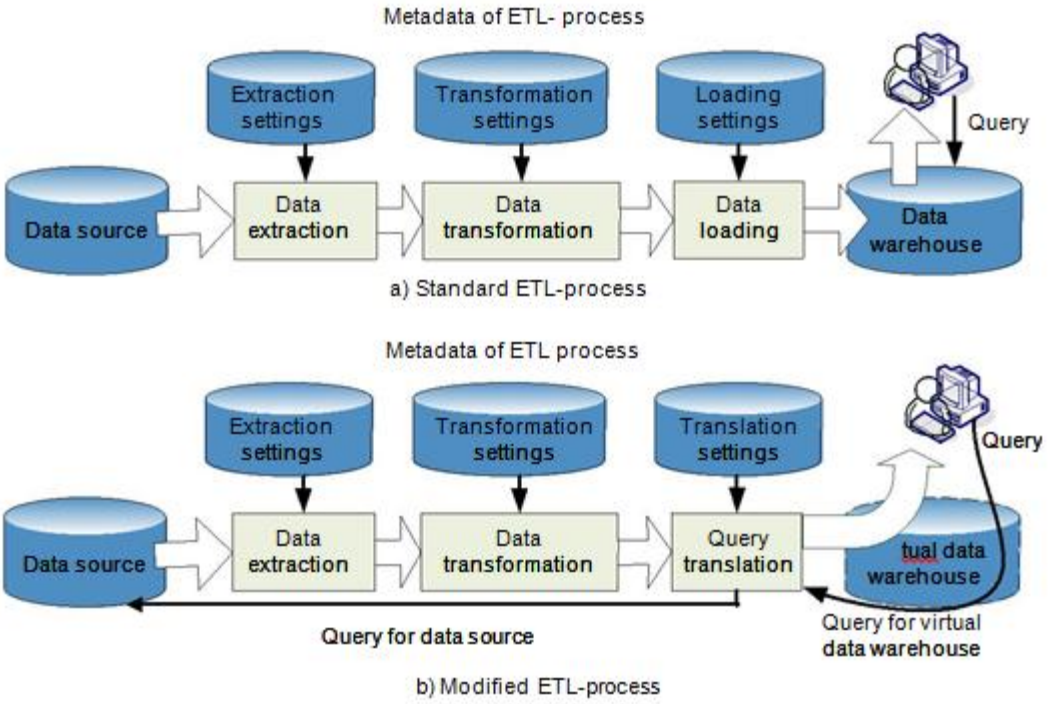

Figure 1.1 ETL Process

IV. BACKGROUND ETL PROCESS

A good Data Warehouse design substantially reduces the report processing time, but it requires an arduous effort in **ETL design and implementation**. ETL i.e. extraction, transformation and loading process is responsible for the extraction of data from various heterogeneous data sources which are currently operational, their transformation (which involves cleaning, format conversion, and business logics etc.) and then finally loading the completely transformed data into the target DWs.

The design of an ETL process has various steps involved viz:-

A. Source Selection: All the operational sources required for DW creation are recognized
B. Source Transformation: Data in multiple sources is standardized into a common one for which data cleansing is performed.
C. Source Joining: Desperate sources are combined together to load to a specific target location.
D. Target Selection: The DW in which the data needs to be loaded is recognized
E. Objects Mapping: The mapping of source attributes is performed with target attributes.
F. Loading data: The transformed data is loaded into the target location.

Earlier this ETL processing was done manually using Java/ SQL code which was a highly complex and inefficient task. In order to overcome these difficulties, various powerful ETL tools are launched in the application market.

Some of the **well-known ETL tools** available are mentioned below:

Tool	**Company**
Datastage	IBM
Informatica	Informatica Corporation
Ab Initio	Ab Initio Software Corporation
MS Sql Server Integration	Microsoft
Talend	Talend

Figure 1.2 ETL Tools

According to [2], about 30% of DW project is spent on the expensive ETL tools and the entire ETL efforts take around 60% to 80% of Data Warehouse implementation. So, the said figures represent the importance of an ETL tool in DW project.

So, to conclude, the end result i.e. the target DW is available in a ready to access format. And there are a number of reporting tools available in the market to produce fruitful results for better decision making.

V. PROJECT RELATED WORK

One of the India's largest service providers of mobile and broadband services wanted to create a reporting solution taking into consideration the need of business users at various levels in the organizational hierarchy.

As the requirement objectives were very clear, a proper development methodology was followed in order to implement the migration of data from non-OLTP systems, such as text files, legacy systems, and spreadsheets to target data warehouse, which can be further used for the business analytical purposes.

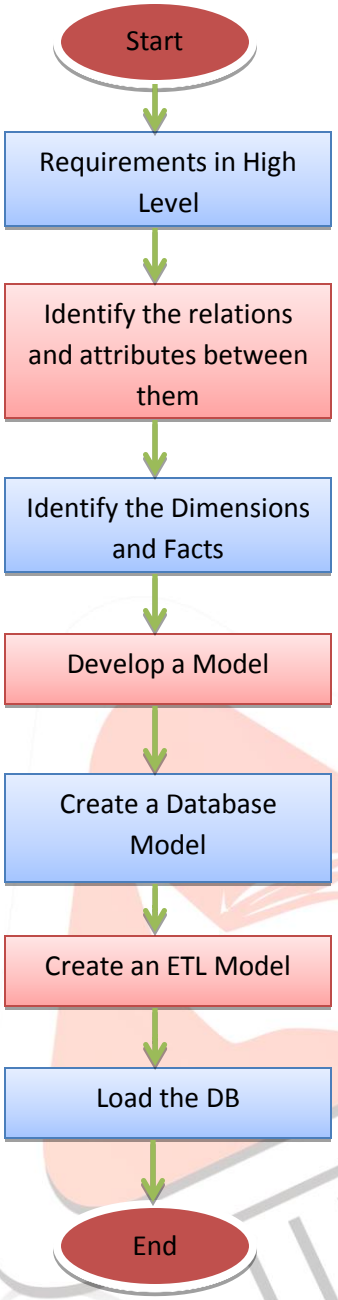

Figure 1.3 Datawarehousing Flow Diagram

The ETL implementation had four distinct phases:

A. Recognizing the Source to Target mapping
B. Designing and building the DW
C. ETL Development
D. Testing

And for the ETL implementation, Datastage was primarily chosen due to the following reason:

A. Proven capabilities for large data migration at customer location [3]

B. Materialized view features
C. Easy access of required data to business leaders

Project Results:

Some of the benefits realized during the course of the project:

A. On-Time Delivery: Team was able to successfully create the DW as per the business' expectations despite difficulties in scope identification and lack of any prior documentation. This was successfully implemented on schedule using an ETL tool. This can be attributed to the chosen transformation methodology and program management discipline.
B. Low Post-Production Defects: Usually in DW projects, the number of PPDs is very high. And this happens not due to the data transformation but due to incorrect mapping rules. Due to the meticulous mapping sessions and inclusion of the right stakeholders at the right time, these errors were limited to very minimal.
C. Better Decision making: The reporting solution i.e. Cognos lying on top of the Data Warehouse comprises of executive dashboards which are having rich interactive BI interface coving advanced visualizations, which can be used by the business for better and on-time decision making in the organization.

VI. CONCLUSION AND FUTURE WORK

So, to conclude, Data warehouse is a special form of database that is intended for deep analysis and strategic planning. It is a repository of integrated information from various OLTP and legacy systems. And since the implementation of a DW is a very expensive process, objectives should be very clear before-hand.

Secondly, for the success of a DW Project, a solid, well-designed, and documented ETL system is very necessary. If it is not planned effectively, then high chances of cost wastage are there. But effective implementation usually leads to curtail down the manual efforts by scheduling the ETL jobs to do the same procedure on other data with periodically fixed time interval.

The proposed methodology is clear, lucid and the application components are easy to configure.

Since, the results in the project are fruitful and encouraging, we are currently working on:

A. Adding other data sources to extend the functionality
B. MIS Sunset of some of the old reports which are still operational
C. ETL tool- Datastage upgradation from V8.7 to V11.3 to move to faster and better functionality
D. To provide mobile solutions to customer i.e. reports are delivered on mobile phones and can be viewed in offline mode as well once downloaded to device

REFERENCES:

[1] Inmon W. Building the Data Warehouse. – New York: John Willey & Sons, 1992.

[2] 1. E. Malinowski, E. Zimanyi, "A Conceptual Model For Temporal Data Warehouses And Its Transformation To The ER And The Object-Relational Models".

[3] Kar Sitikantha, "Large scale data migration using ETL tool", submitted for Software Engineering 2010 Conference, Innsbruck, Austria.