



Proposed work on ETL

Satkaur^[1]

Research scholar, S.K.I.E.T
Kurukshetra, Haryana, India

Anuj Mehta^[2]

Asst.Prof., S.K.I.E.T
Kurukshetra, Haryana, India

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: first, we cover the conceptual and logical modeling of ETL processes, along with some design method we visit each stage of the E-T-L triplet, and examine problems that fall within each of these stages we discuss problems that pertain to the entirety of an ETL process, and we review some research prototypes of academic origin.*

Keywords: *extraction, transformation and loading, data warehouses, datamart, online analytical processing, online transaction protocol*

I. Introduction of ETL

You need to load your 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 warehouse. The process of extracting data from source systems and bringing it into the data warehouse is commonly called ETL, which stands for extraction, transformation, and loading. The acronym ETL is perhaps too simplistic, because it omits the transportation phase and implies that each of the other phases of the process is distinct. We refer to the entire process, including data loading, as ETL. You should understand that ETL refers to a broad process, and not three well-defined steps. The methodology and tasks of ETL have been well known for many years, and are not necessarily unique to data warehouse environments: a wide variety of proprietary applications and database systems are the IT backbone of any enterprise. Data has to be shared between applications or systems, trying to integrate them, giving at least two applications the same picture of the world. This data sharing was mostly addressed by mechanisms similar to what we now call ETL.

II. ETL Tools

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. Oracle Warehouse Builder (OWB), for example, provides ETL capabilities and takes advantage of inherent database abilities. Other data warehouse builders create their own ETL tools and processes, either inside or outside the database. 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. Besides the support for designing a data warehouse and the data flow, these tasks are typically addressed by ETL tools such as OWB. Oracle9i is not an ETL tool and does not provide a complete solution for ETL. However, Oracle9i does provide a rich set of capabilities that can be used by both ETL tools and customized ETL solutions. Oracle9i offers techniques for transporting data between Oracle databases, for transforming large volumes of data, and for quickly loading new data into a data warehouse.

(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.)

III. ETL Process

During extraction, the desired data is identified and extracted from many different sources, including database systems and applications. Very often, it is not possible to identify the specific subset of interest, therefore more data than necessary has to be extracted, so the identification of the relevant data will be done at a later point in time. Depending on the source system's capabilities (for example, operating system resources), some transformations may take place during this extraction process. The size of the extracted data varies from hundreds of kilobytes up to gigabytes, depending on the source system and the business situation. The same is true for the time delta between two (logically) identical extractions: the time span may vary between days/hours and minutes to near real-time. Web server log files for example can easily become hundreds of megabytes in a very short period of time. After extracting data, it has to be physically transported to the target system or an intermediate system for further processing. Depending on the chosen way of transportation, some transformations can be done during this process, too. For example, a SQL statement which directly accesses a remote

target through a gateway can concatenate two columns as part of the SELECT statement. The emphasis in many of the examples in this section is scalability. Many long-time users of Oracle are experts in programming complex data transformation logic using PL/SQL. These chapters suggest alternatives for many such data manipulation operations, with a particular emphasis on implementations that take advantage of Oracle's new SQL functionality, especially for ETL and the parallel query infrastructure.

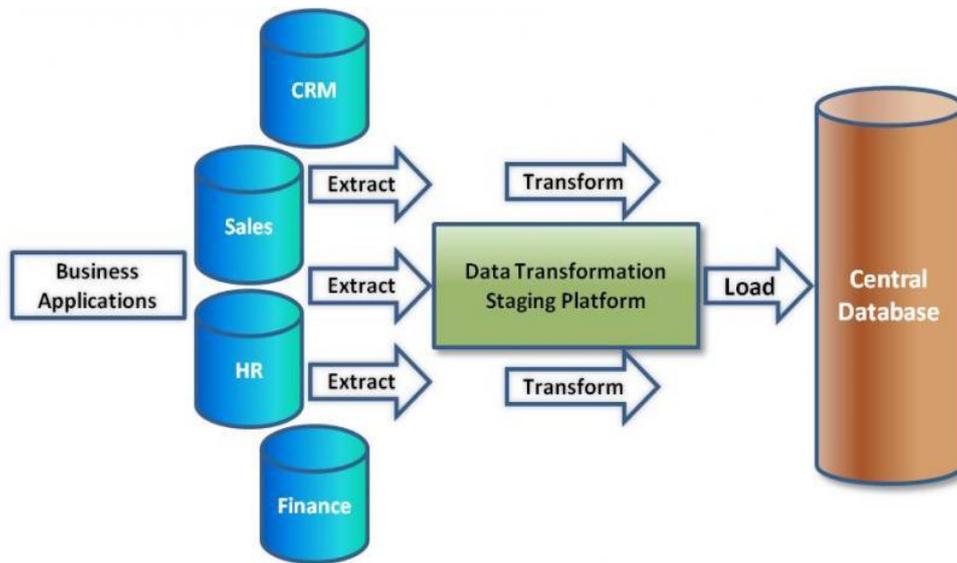


Fig.: ETL process

IV. Existing work on ETL

There are many ETL tools available in the market. However, there are few problems with them. Such tools are very expensive and do not support small size businesses. Configuration of such tools takes a lot of time. Tool customization is not possible and thus, sometimes does not support the scenario provided. These tools have often been seen having common problems of:

- a) Data dependencies.
- b) Complexity of source code.
- c) Poor Quality of data.

V. Proposed work on ETL

In this article, we have tried developing a prototype of Retail application to show case the easy implementation of ETL processes. This is just a prototype we have configured and developed. The aim here is to design and develop a simple and small application to carry out the ETL processes. Tools available in the market are quite expensive and are difficult to configure and use. This prototype is very easy to configure and use. It could also be configured in seconds. Customization would take time depending upon the size and effort proposed. It is a one-time effort. These tools are very easy to use and master. PL/SQL is the basic language used. Not much coding is required.

(PonasVassiliadis, "A survey of Extract-transform-load technology", International Journal of data warehousing & mining, 5(3), 1-27, July-September 2009.)

VI. Comparison Table between existing work and proposed work

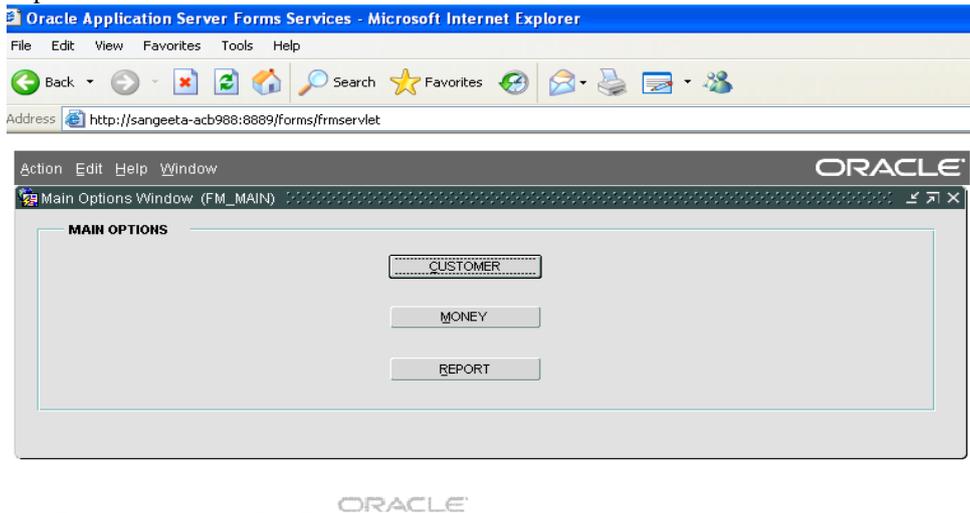
Table 3.1- Comparison between existing work and proposed work.

Serial no.		Prototype	ETL tools
(1)	Design	Simple	Complex to very complex
(2)	Configuration	Very simple	Very complex
(3)	Cost	Cheap	Expensive
(4)	Size	Depending upon Business requirement	Depending upon Business requirement.
(5)	Customization	Could be customized	ETL Tools can not be customized
(6)	Benefits	Prototype software is easily available. So could be used for Testing purpose.	ETL tools are not free to use.

VII. Concept of Loading and Transformation

Using the environment setup, Screen 1 has been built in Oracle Forms to load data warehouse. Underlying coding would transform the data and help in loading database concept CUSTOMER_MASTER.

Screen 1: Main Options Window.



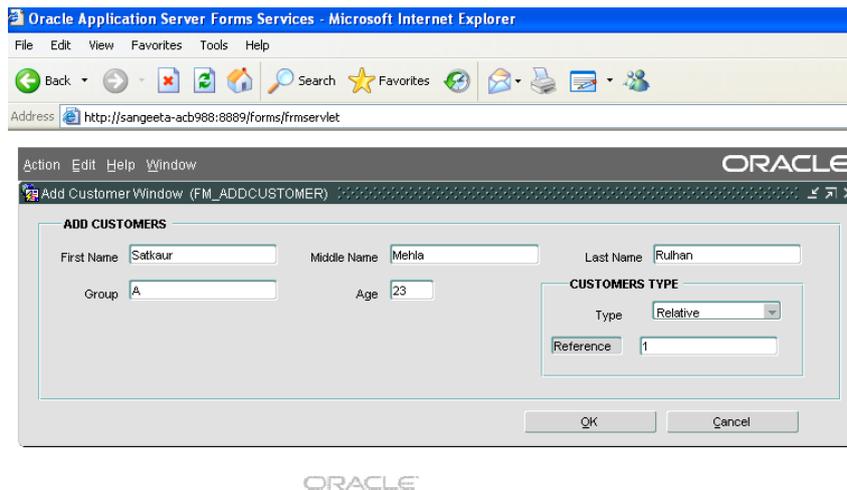
Press 'Customer' button would take user to Screen 2.

Screen 2: Customer Options Window.

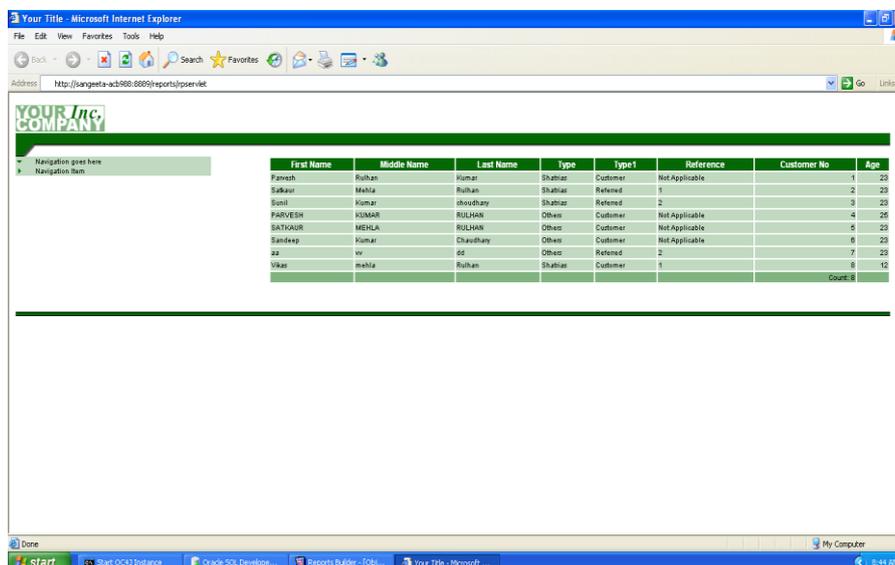
Data has been picked up from front end and has been transformed accordingly before loading into the warehouse database. For coding and to carry out the required transformation, PL/SQL language has been used and columns have been mapped properly to database concept CUSTOMER_MASTER.

Press button 'ADD Customer' in order to load new customer records.

Screen 3: Add Customer Window.

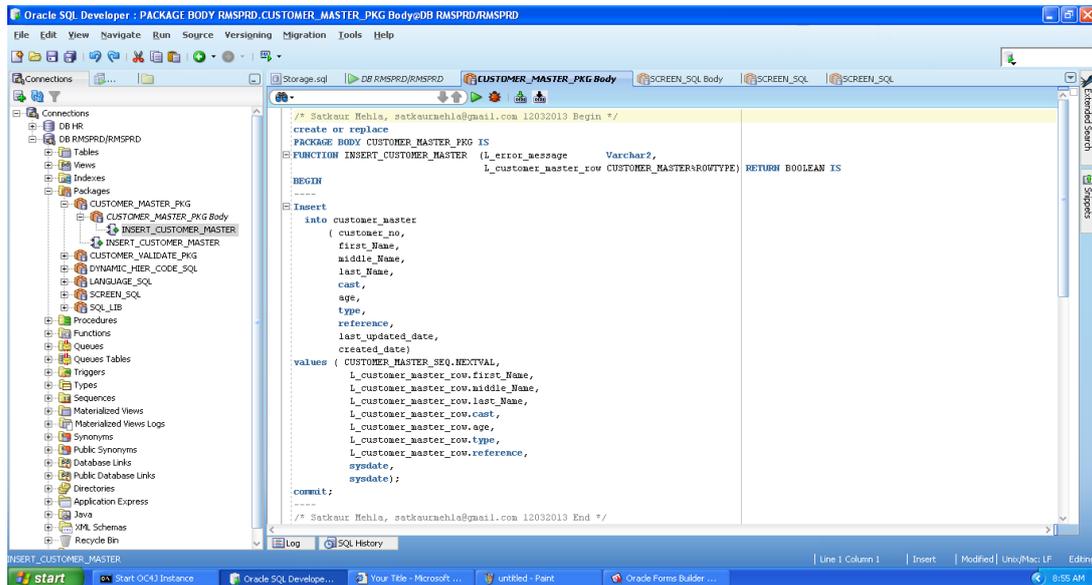


Fill all required values and press 'OK' to load data to warehouse database.



VIII. Concept of Transformation

Data has been picked up from front end and has been transformed accordingly before loading into the warehouse database. For coding and to carry out the required transformation, PL/SQL language has been used and columns have been mapped properly to database concept CUSTOMER_MASTER



IX. CONCLUSION

This article examined issues and approaches in transforming and loading data into the data warehouse. Also, we have tried to cover up the basic step by step installation and configuration of a prototype which carries out all the ETL process however quite easily and time saving manner.

The ETL process is a development process like any other. To create or purchase a viable ETL solution, it takes an understanding of the requirements, source and target data and data structures, and the technical and political environment and it is always good and wise to have an application which could be customized as per changing needs.

X. Future scope

1. Security to the data is one of the major challenges and area of concern in today's world. Current approaches for the modeling of ETL do not address the security issues in the ETL modeling.
2. Next generation Data integration Tools should support this extraction process to speed up the extraction process.
3. This Extraction process limits itself with the complex data types.
4. Further work can be carried out with the implementation of transformation and loading process to improve the entire ETL process.
5. CQL based on Unix can be used to improve the efficiency of Query Processing
6. Maximum use of parallelism: to load data into two databases, one can run the loads in parallel (instead of loading into 1st - and then replicating into the 2nd).
7. A common source of problems in ETL is a big number of dependencies among ETL jobs

REFERENCES

- [1] PonasVassiliadis, "A survey of Extract-transform-load technology", International Journal of datawarehousing& mining, 5(3), 1-27, July September 2009 1.
- [2] Vishal Gaur 1, Dr. S.S. Sarangdevot 2, Govind Singh Tanwar 3, Anand Sharma 4, " Improve performance of Extract, transform and Load ", Vishal Gaur et.al/ International Journal on Computer Science and Engineering Vol. 02, No.-03 ,2010 786-789.
- [3] Shaker H. Ali EI- Sappagh*a, Abdeltawab H. Ahmed Hendawi*b, Ali HamedEI Bastawissy*b, " A Proposed Model For datawarehouse ETL Process", Journal of King Saud University-Computer and information Sciences (2011)23, 91-104.
- [4] Radha Krishna author1 and Sreekanth Author 2, "An Object Oriented Modelling and Implementation of web based ETL process", International Journal of computer science and network security, Vol. 10 No. 2, February 2010.
- [5] 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.
- [6] Nestor Rodruetz, Kent Lawson, Eddie Molina, "Data Warehouses tool Evaluation. ETL Focused", Univesity of Texas-Pan American 1201 W. University Drive, Edinburg, TX (956) 665-UTPA.

- [7] Fundulaki [1], Alex Averbuch [2], Eva Daskalaki [3], “ Overview and analysis of Existing benchmark framework, LDBC Cooperative Project FP7-317548.
- [8] Thomas Van Raalte, “ Introduction to Oracle Retail data model Implementation and Operation Guide” , Release 11.3.2 E20363-03, January 2013.
- [9] AlkisSimitsis 1, PanosVasiliadis 2, “A Methodology for the Conceptual Modelling of ETL process”, National Technical University of Athens, Dept. of Electrical and Computer Engineering , Computer Science Division , IronPolytechniou. 9, 15773, Athens, Greece asimi@ dbnet.ece.ntua.gr, University of Ioannina, Dept. of Computer Science, 45110,Ioannina, Greece. Pvassil@ cs.uoi.gu.
- [10] ThomasJorg[1], Stefan De Bloch, “Towards generating ETL process for incremental Loading” University of Kaiserslanterm, 67.653 Kaiserslam term, Germany.
- [11] [http:// dbs.uni-leipzig.de](http://dbs.uni-leipzig.de). Erhard Rahm*, Hang Hai Do, University of Leipzig, Germany, “Data Cleaning :Problem and Current Approaches”.