ISSN: 2321-7782 (Online)
Volume 2, Issue 3, March 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Paper / Case Study
Available online at: www.ijarcsms.com

ETL Tools for Organised Data Warehouse Management - A Survey

P. Amuthabala[1]
Dept. of Information Science & Engg.,
Atria Institute of Technology
Bangalore, India

Sheba Jebakani[2]
Dept. of Information Science & Engg.,
Atria Institute of Technology
Bangalore, India

Abstract: ***Data Warehousing is a promising part in data extraction and transformation for data management system. The necessity to understand in the application of this system is increasingly appreciated when the present corporate management is growing with sophistication generating more data from multi-sources and multi-disciplines. A survey of the available studies in this application is therefore needed to make use of the technique for accurate decision making for successful management.***

Keywords: Database; ETL; Business Models; Forecasting; Data Warehouse.

I. INTRODUCTION

The success of a business management model depends on its decision making in either domestic or global marketing environment. The present business strategy generates vast amount of data not from a single source but multiple sources and multi-disciplines like finance, procurement, production centres, marketing and sales and of course customer satisfaction feed-back. The customer's feed-back decides the success of the business organization and needs to be attended to with prioritized strategies using the large amount of data generated from all sources without much compensation in its profits. These data should be properly channelized for correct decision making. This is an important primary requirement of a data management system and a number of research studies have been attempted to evolve necessary tools and decide on a model to manage various databases on a real time basis. Some of the critical results and reports from connected works are studied and presented below briefly.

II. RESEARCH ACTIVITIES

Many activities have been carried out from requirement based object oriented processes [6] to evolve suitable data management system for a real time Data Warehouse approach. A real time process is necessary as there are constant changes in data due to change in expectations of corporate decision making [5]. These also include functional behaviour during the construction of models to store and retrieve data quickly, efficiently and accurately for decision making [2]. Many more research activities have been reported in forecasting from data analysis for predicting future trends in business environment [9]. More research is recommended in this field:

- As data delivery from different data sources must be entered simultaneously into real time DW.
- As the Accuracy of data entered into the realtime DW is necessary to forecast future business trend(s) which is a difficult process.

III. DATA PROCESSING FOR DW

The final quality of output – from government establishments or business organizations or research laboratories or financial institutions – depend on the selection and processing of raw materials - here data from multi sources. Low quality data in terms of missing, dirty, inaccurate and inconsistent data will result in low quality and unreliable DW which will cause chaos and catastrophe in decision making forecasting.

Data refinement is required to avoid such situations and the process of refinement especially when data are entered from multi-discipline and multi-sources involve the following techniques:

- Data cleaning
- Data integration/transformation and
- Develop DW model.

IV. DISCUSSION

A. Cleaning and Transformation

Generally all databases/Data Warehouses suffer from noisy, incomplete, inaccurate and inconsistent data collection and entry specifically from heterogeneous data sources [6]. Therefore this part of refining data has attracted many research activities resulting in extensive studies and subsequent reports of varying degrees. Most of the research activities are based on specific case studies and have commented on work flow models and tools developed by individual scholars [7]. This suggests that even though there are many tools (like "Data Scrubbing", "Data Auditing") available for data cleaning and data transformation, before loading into DW, it is necessary to improvise the available tools to suit individual management system to achieve real time decision making and accurate forecasting [1].

It has also been reported that there is a need for the appropriate design of data based schema and cleaning rules (like queries) during DW design when improvements may also be suggested to the constraints in the scheme followed [2]. This also opens doors for innovations in the tools used for enhancing reliability and repeatability suggesting that innovation step is not a onetime process but a continuous one. This is especially true as problems and constraints encountered in static source system are aggravated in the case multi source and multi disciplined system. Therefore the research scholars have reported the following points to be considered: - [5]

- Data analysis
- Define Transformation and mapping Rules
- Verify corrections and effectiveness
- Adopt multiple interactions of the above three points to prevent some errors that may become apparent during Transformation and Loading of data into a Data Warehouse.
- Include provision in the schema for achieving conflict detection by back flow of data
- Use appropriate language for cleaning and transformation processes in system operation.

B. Model for DW-ETL Processes

A good number of reviews are available on modelling based on conceptual designs for ETL Modelling. Regular and periodical Auditing are performed to ensure the accuracy and reliability and also to take care of changes taking place at sources as the expectations of corporate executives often keeps changing. Three main approaches have been suggested to ensure this step of Auditing: [7]

- Use queries to represent mapping between source(s) and target(s) (required by corporate Systems)
- Based on this it is possible to introduce a framework and propose a model for ETL-DW Processes
- With the experience of 1 and 2 the available model(s) to arrive at a refined working ETL Model.

C. DW Model and OLAP Technology

The corporate managers always look for data from DW tools to organize, understand and use the data to make effective strategic decisions in the competitive and fast evolving global market. It is a very expensive process to build an effective and efficient DW which should be designed, if needed, on subject oriented basis with integrated, time variant and non-volatile collection of data in support of management decision making processes. This also helps to evolve a better supporting system to business management [9].

D. Real Time or Active Data Warehouse

Real time or active DW is the need for business strategy to meet the global demands in the competitive market. This is also due to data management for development of analytics and business intelligence. In real time DW, instead of periodic loading as in traditional ETL approaches, loading is done continuously. Besides the available information in this field, further research activities are recommended as the nature of data delivery from different data sources are becoming more and more complicated and must simultaneously enter the DW in real time DW [1].

E. Efficient Business Decision Making

Having studied all DW processes, developments and requirements, it is recommended in the reviews that four components are required for efficient business decision making:

- Proper scheduling in ETL work flow management
- Identify the resources for data collection
- Locate priority controllers
- Decide concurrency controllers.

And the most important part of all four components is that all of them should work in parallel to each other to achieve efficiency and performance in any ETL work flow.

V. FORECAST

Forecasting from data analysis is an important feature in predicting future trends in business environment. An effective forecasting is based on predictive accuracy, speed, reliability and repeatability. Many methods have been reported and comparisons are available for different classifications and predictions in [9]. It has been found that accurate use of any method must consider training time, robustness and interpretability. For further guidance, it is recommended to refer relevant text books and research papers to learn more about machine learning, statistics and model recognition perspectives and also for detailed discussion on selection measures.

VI. CONCLUSION

A great progress has been made and is also continuously being activated in the field of Data Warehousing. It is necessary to develop new DW methods, systems and application of tools (either general or tailor made models) with numerous enhancements and reorganizations to meet the ever enlarging business decision making requirements. This leads to more and more research activities with the ongoing systems.

ACKNOWLEDGEMENT

We would like to thank Mr. Venkataraman, for helping us in doing this survey on Datawarehouse which is also widely used in the area of business Intelligence.

References

1. Dr. Kamal Kakish and Terasa A Kraft, "ETL Evaluation For Real Time DW," Proceedings of the conference on Information Systems Applied Research, New Orleans, Louisiana, USA. ISSN: 2167-1508, VSN 2214.
2. Paulraj M & Sivaprakasam P, "Functional Behaviour Pattern For Data Mark Based on Attribute Relativity," International Journal of Computer Sciences Issues, Vol.Issue 4, No.1, July 2012.
3. Saptarisi Goswami, Samir Ghosh, and Amlan Chakravarti, "Outlier Detection Technique for SQL and ETL Tuning," International Journal of Computer Applications (0975-8887), Vol. 23, No.8, June 2011.
4. MakinSaifur Rahman Malik. Azra Shamin and Sajid Ullah Khan Gorsi, "Revised Framework For ETL Workflow Management For Efficient Business Decision," International Journal of Computer Theory and Engineering, Vol. 5, No. 3, June 2013.
5. Erhard Rahm and Hong Hi Do, "Data Cleaning: Problems and Current Approaches," University of Leipzig, Germany. (This paper was prepared at Microsoft Research, Redmond, WA)
6. Payal Pahwa Taneja and Gorims Thakar, "UCLEAN: A Requirement Based Object Oriented ETL framework," Guru Govind Indraprasta University, Delhi. International Journal of Computer Science and Engineering Survey. Vol. 2, No. 4, 2011.
7. Shekar H, Ali El. Sappagh. King Saudi University, Saudi Arabia.Abdeltwab M. Ahmed Hendawi and Ali Ahmed El Bastawissi. Faculty of Computers and Information, Cairo University, Cairo, Egypt. "A Proposed Model for DW ETL Processes," Journal of Saudi King University, Computer and Information Sciences, (2011)23, 91-104.
8. Paul Raj M and Sivaprakash P. "Functional Layer Interfaced Data Mart Architecture (FLIDMA)," Department of Computer Science, Vinayaka Missions University, Salem. Department of Computer Science, Sri Vasani College, Erode. T.N. India
9. IJCSI, International Journal of Computer Science Issues, Vol. 9, Issue 4, No. 1, July 2012.
10. Jia Wei & Micheline kamber, "Data mining-Concepts & Techniques" Publisher: Morgan Kaufmann.

