International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014
ISSN 2091-2730

Analysis of ETL Process in Data Warehouse

N.Nataraj[1], Dr.R.V.Nataraj[2]

[1]PG Scholar,[2]Professor, Department of Information Technology, Bannari Amman Institute of Technology, Sathyamangalam,
nataraj.se13@bitsathy.ac.in

Abstract— ETL is responsible for the extraction of data, their cleaning, conforming and loading into the target. ETL is a Critical layer in DW setting. It is widely recognized that building ETL processes is expensive regarding time, money and effort. In this, firstly we review commercial ETL tools and prototypes coming from academic world. After that we review designing works in ETL field and modelling ETL maintenance issues. We review works in connection with optimization and incremental ETL, then finally challenges and research opportunities around ETL processes.

Keywords— ETL, Data warehouse, ETL Modelling, ETL Maintenance

INTRODUCTION

Enterprises as organizations invest in DW projects in order to enhance their activity and for measuring their performance. It aims to improve decision process by supplying unique access to several sources. In this we have two types. The two famous types are databases and flat files. Finally let note that sources are autonomous or semi autonomous. It is the integration layer in DW environment [1]. ETL tools pull data from several sources (databases tables, flat files, ERP, internet, and so on), apply complex transformation to them. ETL is a critical component in DW environment. Indeed, it is widely recognized that building ETL processes, during DW project, are expensive regarding time and money.A. Data Warehouse layersSources: They encompass all types of data sources. They are data provider. The two famous types are databases and flat files. Finally let note that sources are autonomous or semi autonomous.

ETL: It is the integration layer in DW environment. ETL tools pull data from several sources apply complex transformation to them. Finally in the end, data are loaded into the target which is data warehouse store in DW environment.

Data Warehouse: is a central repository to save data produced by ETL layer. DW is a DB includes fact tables and dimension tables. Together these tables are combined in a specific schema that may be star schema or snowflake schema.

Reporting and Analysis: Collected data are served to end-users in several formats. For example data is formatted into reports, histograms.

Extraction:The problem data from a set of sources which may be local or distant. Logically, data sources come from operational applications, but there is an option to use external data sources for enrichment. External data source means data coming from external entities. Thus during extraction step, ETL tries to access available sources, pull out the relevant data, and reformat such data in a specified format.

Transformation:This step is the most laborious one where ETL adds value. This step is associated with two words: clean and conform. In one hand, cleaning data aims to fix erroneous data and to deliver clean data for end users (decisions makers). Dealing with missing data, rejecting bad data are examples of data cleaning operations. In other hand, conforming data aims to make data correct, in compatibility with other master data.Checking business rules, checking keys and lookup of referential data are example of conforming operations.

Loading: This step conversely to previous step, has the problem of storing data to a set of targets. During this step,

ETL loads data into targets which are fact tables and dimension in DW context.

Commercial ETL Tools We have two types of ETL tools. On the one hand, there is subfamily of payable ETL Data Stage and Informatica [7][8] . On the other hand, the second subfamily of commercial ETL comes with no charge [9].

Informatica: Informatica is broadly used ETL tool for extracting the source data and loading it into the target after applying the Required Transformation .ETL developers map the extracted data from source systems and load it to target systems after applying the required transformations [9].

Data Stage: Its basic element for data manipulation is called "stage." Thus, for this tool an ETL process is a combination of "stages." Thus we speak about transformation stages and stages for extracting and loading data (called connectors since release which are interconnected via links.

SSIS:SSIS imposes two levels of tasks combination. The first level is called "Flow Control" and the second level controlled by the first, is called "Data flow." Indeed, the first level is dedicated to prepare the execution environment (deletion, control, moving files, etc....) and supplies tasks for this purpose. The second level (data flow) which is a particular task of the first level performs classical ETL mission. The Data-Flow task offers various tasks for data extraction, transformation and loading.

SIRIUS:It develops an approach metadata oriented that allows the modelling and execution of ETL processes [13]. It is based on SIRIUS Meta model component that represents metadata describing the necessary operators or features for implementing ETL processes. In other words, SIRIUS provides functions to describe the sources, targets description and the mapping between these two parts [12].

ARKTOS:ARKTOS is another framework that focuses on the modelling and execution of ETL processes. Indeed, ARKTOS provides primitives to capture ETL tasks frequently used. More exactly, to describe a certain ETL process, this framework offers three ways that are GUI and two languages XADL (XML variant) and SADL (SQL like language).

DWPP:DWPP is a set of modules designed to solve the typical problems that occur in any ETL project. DWPP is not a tool but it is a platform. Exactly, it is C functions library shared under UNIX operating system for the implementation of ETL processes. Consequently, DWPP provides a set of useful features for data manipulation.

II. MODELLING AND DESIGN OF ETL

ETL are areas with high added value labelled costly and risky. In addition, software engineering requires that any project is doomed to switch to maintenance mode. For these reasons, it is essential to overcome the ETL modelling phase with elegance in order to produce simple models and understandable. This method is spread over four steps: 1. Identification of sources 2. Distinction between candidates' sources and active sources. 3. Attributes mapping. 4. Annotation of diagram (conceptual model) with execution constraints. A. Meta-data models based on ETL Design the designer needs to: 1. Analyse the structure and sources. 2. Describe mapping rules between sources and targets. The based on meta-model, provides a graphical notation to meet this necessitate [13].

III. ETL PROCESS MAINTENANCE:

When changes happen, analyzing the impact of change is mandatory to avoid errors and mitigate the risk of breaking existent treatments. As a consequence, without a helpful tool and an effective approach for change management, the cost of maintenance task will be high. Particularly for ETL processes, previously judged expensive and costly [14], [15]. Using ETL terminology, above previous research efforts focus on the target unlike the proposal of which focuses onchanges in the sources. In these proposal dealing with change management in ETL are interesting and offer a solution to detect changes impact on ETL processes. However change incorporation is not addressed.

IV. RESEARCH OPPORTUNITIES

1. Many conceptual models enrich the ETL design field. However no proposal becomes a standard neither widely accepted by research community like multi-dimensional modeling in data warehouse area.
2. Mapping rules are an important delivery in ETL design.
3. Big data technologies arrive with exciting research opportunities. Particularly, performance issue seems solvable withthis novelty.
4. Tests are fundamentals aspects of software engineering. In spite of this importance, and regarding ETL, they are neglected.
5. Meta data and unstructured data [2].

V. CONCLUSION

ETL is identified with two tags: complexity and cost. Due its importance, this paper focused on ETL, the backstage of DW, and presents the research efforts and opportunities in connection with these processes. It is widely familiar that building ETL processes is expensive concerning time, money and effort. It consumes up to 70% of resources. Therefore, in current survey, firstly we give a review on open source and commercial ETL tools, along with some ETL prototypes coming from academic world. Namely, SIRIUS, ARKTOS. Before conclusion, we have given an picture of performance issue along review of some works dealing with this issue, particularly, ETL optimization and incremental ETL. Finally, this surveys ends with presentation of main challenges and research opportunities around ETL processes.

REFERENCES:

[1] W. Inmon D. Strauss and G.Neushloss, “DW 2.0 The Architecture for the next generation of data warehousing”, Morgan Kaufman, 2007.

[2] A. Simitisis, P. Vassiliadis, S.Skiadopoulos and T.Sellis, “DataWarehouse Refreshment”, Data Warehouses and OLAP: Concepts, Architectures and Solutions, IRM Press, 2007, pp 111-134.

[3] R. Kimball and J. Caserta. “The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data”, Wiley Publishing, Inc, 2004.

[4] A. Kabiri, F. Wadjinny and D. Chiadmi, "Towards a Framework for Conceptual Modelling of ETL Processes ", Proceedings of The first international conference on Innovative Computing Technology (INCT 2011), Communications in Computer and Information Science Volume 241, pp 146-160.

[5] P.Vassiliadis and A.Simitsis, “EXTRACTION, TRANSFORMATION, AND LOADING“, http://www.cs.uoi.gr/~pvassil/publications/2009_DB_encyclopedia/Extract-Transform-Load.pdf

[6] J. Adzic, V. Fiore and L. Sisto, “Extraction, Transformation, and Loading Processes”, Data Warehouses and OLAP: Concepts, Architectures and Solutions, IRM Press, 2007, pp 88-110.

[7] W. Eckerson and C. White, “Evaluating ETL and Data Integration Platforms”, TDWI REPORT SERIES, 101communications LLC, 2003.

[8] IBM InfoSphereDataStage, http://www-01.ibm.com/software/data/infosphere/datastage/

[9] Informatica, http://www.informatica.com

[10] C. Thomsen and T,B Pedersen, "A Survey of Open Source Tools for Business Intelligence", DB Tech Reports September 2008. Homepage: www.cs.aau.dk/DBTR.

[11] C. Thomsen and T,B Ped- ersen, "A Survey of Open Source Tools for Business Intelligence".International Journal of Data Warehousing and Mining.Volume 5, Issue 3, 2009.

[12] Talend Open Studio, www.talend.com

[13] A.VAVOURAS,"A Metadata-Driven Approach for DataWarehouse Refreshment", Phd Thesis, DER UNIVERSITÄT ZÜRICH,ZÜRICH, 2002.

[14] J.F. Roddick et al, "Evolution and Change in Data Management - Issues and Directions", SIGMOD Record 29, Vol. 29, 2000, pp 21-25