Modeling and Optimization of Extraction-Transformation-Loading (ETL) processes in Data Warehouse: An Overview

Nitin Anand

Research Scholar

AIACT&R

New Delhi

Manoj Kumar

Associate Professor

AIACT&R

New Delhi

Abstract:

ETL processes are responsible for the extraction of data from several sources, their cleansing, their customization and transformation, and finally, their loading into a data warehouse. In this paper we are reviewing the optimization of its execution plan. We review on the optimization of the sequence of the ETL operations involved in the overall process.

.***Keywords- Data Mart , Data Quality (DQ), Data Staging Area, Data warehouse, ETL , Metadata ,OLTP***

1. INTRODUCTION

The term data warehouse, in a high-level description, stands for a collection of technologies that aims at enabling the knowledge worker (executive, manager, analyst, etc.) to make better and faster decisions by exploiting integrated information of the different information systems of his/her organization. 'A data warehouse is a subject-oriented, integrated, time-variant, non-volatile collection of data used to support the strategic decision-making process for the enterprise [1]. It is the central point of data integration for business intelligence and is the source of data for the data marts, 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.

During the ETL process, data is extracted from an OLTP databases, transformed to match the data warehouse schema, and loaded into the data warehouse database [2] . Many data warehouses also incorporate data from non-OLTP systems, such as text files, legacy systems, and spreadsheets. ETL is often a complex combination of process and technology that consumes a significant portion of the data warehouse development efforts and requires the skills of business analysts, database designers, and application developers. The ETL process is not a one-time event. As data sources change the data warehouse will periodically updated. Also, as business changes the DW system needs to change – in order to maintain its value as a tool for decision makers, as a result of that the ETL also changes and evolves. The ETL processes must be designed for ease of modification. A solid, well-designed, and documented ETL system is necessary for the success of a data warehouse project.

2. PHASES OF ETL

An ETL system consists of three consecutive functional steps: extraction, transformation, and loading:

2.1 Extraction

The ETL Extraction step is responsible for extracting data from the source systems. Each data source has its distinct set of characteristics that need to be managed in order to effectively extract data for the ETL process [3]. The process needs to effectively integrate systems that have different platforms, such as different database management systems, different

operating systems, and different communications protocols.

2.2. Transformation

The second step in any ETL scenario is data transformation. The transformation step tends to make some cleaning and con-forming on the incoming data to gain accurate data which is correct, complete, consistent, and unambiguous. [5]

This process includes data cleaning, transformation, and integration. It defines the granularity of fact tables, the dimension tables, DW schema (star or snowflake), derived facts, slowly changing fact tables and dimension tables . All transformation rules and the resulting schemas are described in the metadata repository.

2. 3. Loading

Loading data to the target multidimensional structure is the final ETL step. In this step, extracted and transformed data is written into the dimensional structures actually accessed by the end users and applications [6]

3. FUNCTIONALITY OF ETL TOOLS

ETL tools represent an important part of data warehousing, as they represent the mean in which data actually gets loaded into the warehouse. To give a general idea of the functionality of these tools we mention their most prominent tasks, which include:

(a) the identification of relevant information at the source side,

(b) the extraction of this information,

(c) the transportation of this information to the DSA,

(d) the transformation, (i.e., customization and integration) of the information coming from multiple sources into a common format,

(e) the cleaning of the resulting data set, on the basis of database and business rules, and

(f) the propagation and loading of the data to the data warehouse and the refreshment of data marts.

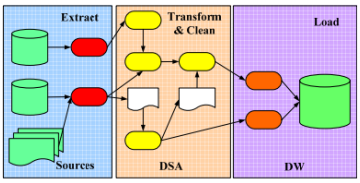

Figure 2. The environment of Extraction-Transformation-Loading process

In Figure 2, we describe the general framework for ETL processes. In the left side, we can observe the original data stores (Sources) that are involved in the overall process. Typically, data sources are relational databases and files. The data from these sources are extracted by specialized routines or tools, which provide either complete snapshots or differentials of the data sources. Then, these data are propagated to the data staging area (DSA) where they are transformed and cleaned before being loaded into the data warehouse. Intermediate results, again in the form of (mostly) files or relational tables are part of the data staging area. The data warehouse (DW) is depicted in the right part of Figure 2 and comprises the target data stores, i.e., fact tables for the storage of information and dimension tables with the description and the multidimensional, roll-up hierarchies of the stored facts. The loading of the central warehouse is performed from the loading activities depicted in the right side before the data warehouse data store.

4. THE LIFECYCLE OF DATA WAREHOUSE AND ITS ETL PROCESS

The lifecycle of a data warehouse begins with an initial Reverse Engineering and Requirements Collection phase where the data sources are analyzed in order to comprehend their structure and contents. At the same time, any requirements from the part of the users (normally a few power users) are also collected. The deliverable of this stage is a conceptual model for the data stores and the processes involved. In a second stage, namely the Logical Design of the warehouse, the logical schema for the warehouse and the processes is constructed. Third, the logical design of the schema and processes is optimized and refined to the choice of specific physical structures in the warehouse (e.g., indexes) and environment-specific execution parameters for

the operational processes. We call this stage Tuning and its deliverable, the physical model of the environment [11]. In a fourth stage, Software Construction, the software is constructed, tested, evaluated and a first version of the warehouse is deployed. This process is guided through specific software metrics. Then, the cycle starts again, since data sources, user requirements and the data warehouse state are under continuous evolution. An extra feature that comes into the scene after the deployment of the warehouse is the Administration task, which also needs specific metrics for the maintenance and monitoring of the data warehouse. [12]

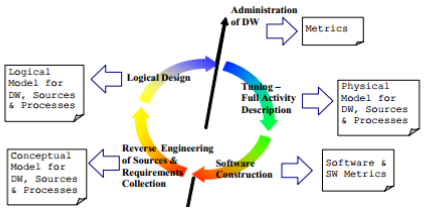

Figure 3. The lifecycle of a Data Warehouse and its ETL processes [14]

4. Related Work

The MetaData Coalition (MDC), is an industrial, non-profitable consortium with aim to provide a standard definition for enterprise metadata shared between databases, CASE tools and similar applications. The Open Information Model (OIM) [4] is a proposal (ledby Microsoft) for the core metadata types found in the operational and data warehousing environment of enterprises. MDC uses UML both as a modeling language and as the basis for its core model. OIM is divided in sub-models, or packages, which extend UML in order to address different areas of information management. The Database and Warehousing Model is composed from the Database Schema Elements package, the Data Transformations Elements package, the OLAP Schema Elements package and the Record Oriented Legacy Databases package. The Database Schema Elements package contains three other packages: a Schema Elements package (covering the classes modeling tables, views, queries, indexes, etc.), a Catalog and Connections package (covering physical properties of a database and the administration of database connections) and a Data Types package, standardizing a core set of database data types.

The Data Transformations Elements package covers basic transformations for relational-to-relational translations. The package is not dealing with data warehouse process modeling, i.e., it does not cover data propagation, cleaning rules, or the querying process), but covers in detail the sequence of steps, the functions and mappings employed and the execution traces of data transformations in a data warehouse. The ETL process, in data warehouse, is a hot point of research because of its importance and cost in data warehouse project building and maintenance. The method of systematic review to identify, extract and analyze the main proposals on modeling conceptual ETL processes for DWs [13] Generating ETL processes for incremental loading [13]

5. Different perspectives for an ETL workflow

We follow a multi-perspective approach that enables to separate these parameters and study them in a principled approach. We are mainly interested in the design and administration parts of the lifecycle of the overall ETL process, and we depict them at the upper and lower part of Fig. 2, respectively. At the top of Fig. 4, we are mainly concerned with the static design artifacts for a workflow environment. We will follow a traditional approach and group the design artifacts into physical, with each category comprising its own perspective. We depict the logical perspective on the left-hand side of Fig. 4, and the physical perspective on the right-hand side. At the logical perspective, we classify the design artifacts that give an abstract description of the workflow environment. First, the designer is responsible for defining an execution plan for the scenario. The definition of an execution plan can be seen from various perspectives. The execution sequence involves the specification of which activity runs first, second, and so on, which activities run in parallel, or when a semaphore is defined so that several activities are synchronized at a rendezvous point. ETL activities normally run in batch, so the designer needs to specify an execution schedule, i.e., the time points or events that trigger the execution of the

scenario as a whole. Finally, due to system crashes, it is imperative that there exists a recovery plan, specifying the sequence of steps to be taken in the case of failure for a certain activity (e.g., retry to execute the activity, or undo any intermediate results produced so far). On the right-hand side of Fig. 4, we can also see the physical perspective, involving the registration of the actual entities that exist in the real world. We will reuse the terminology of [8] for the physical perspective. The resource layer comprises the definition of roles (human or software) that are responsible for executing the activities of the workflow. The operational layer, at the same time, comprises the software modules that implement the design

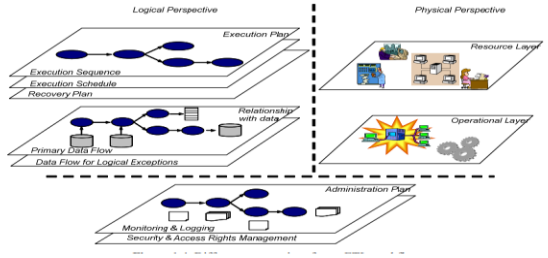

Figure 4.Different perspectives for an ETL workflow [7]

5. CONCLUSION AND FUTUREWORK

ETL processes are very important problem in the current re-search of data warehousing. In this paper, we have investigated a very important problem in the current research of data ware-housing. This problem represents a real need to find a standard conceptual model for representing in simplified way the extrac-tion, transformation, and loading (ETL) processes. In a recent study [9], the authors report that due to the diversity and heterogeneity of data sources, ETL is unlikely to become an open commodity market. This paper describes the flexibility of adding various types of information sources, which ultimately helps in storing the data into the Data Staging Area. Since quality plays an important role in developing software products, we have presented functional requirements along with non-functional requirement i.e., security requirements. This approach is better compared to existing systems. In [10], the authors report on their data warehouse population system.

The architecture of the system is discussed in the paper, with particular interest (a) in a 'shared data area' which is an in-memory area for data transformations, with a specialized area for rapid access to lookup tables and (b) the pipelining of the ETL processes.

The future work may include to dealing with other non-functional requirements like reliability, performance etc. In this paper, we have focused on the data-centric part of logical design of the ETL scenario of a data warehouse.

6. REFERENCES

[1]. W.H. Inmon. Building the Data Warehouse. 2nd edition. John Wiley & Sons, Inc., New York, 1996.

[2].Alex Berson, and Stephen J. Smith, Data Warehousing, Data Mining, and OLAP (New York:. mcgraw-Hill, 1997

[3] D. Theodoratos, S. Ligoudistianos, T.Sellis. View selection for designing the global data warehouse. Data & Knowledge Engineering, 39(3), pp. 219-240, 2001.

[4] MetaData Coalition. Open Information Model, version 1.0. 1999. Available at: http://www.MDCinfo.com

[5] E. Rahm, H. Hai Do. Data Cleaning: Problems and Current Approaches. Bulletin of the Technical Committee on Data Engineering, Vol. 23, No. 4, 2000.

[6] Oracle Corporation. Oracle9i™ SQL Reference. Release 9.2, pp.17.77-17.80, 2002

[7] Won Kim et al (2002)- "A Taxonomy of Dirty Data " Kluwer Academic Publishers 2002

[8] W.M.P. van der Aalst, A.H.M. ter Hofstede, B. Kiepus-zewski, A.P. Barros. Workflow Patterns, BETA Working Paper Series, WP 47, Eindhoven University of Technology, Eindhoven, 2000, available at the Workflow Patterns website, at tmithttp://www.tm.tue.nl/research/patterns/documentation.htm

[9] Giga Information Group. Market Overview Update:ETL. Technical Report RPA-032002-00021, March 2002

[10] J. Adzic, V. Fiore, Data Warehouse Population Platform,in: Proceedings of the Fifth International Workshop on the Design and Management of Data Warehouses (DMDW"03), Berlin, Germany, September 2003

[11] J. Trujillo, S. Luján-Mora. A UML Based Approach for Modeling ETL Processes in Data Warehouses. In the Proceedings of the 22nd International Conference on Conceptual

Modeling (ER'03), LNCS 2813,pp. 307–320, Chicago, Illinois, USA, 2003.

[12] A.Simitsis .Mapping Conceptual to Logical Models for ETL Processes DOLAP 05, ACM, (2002) 67-76

[13] Jo¨rg, Thomas, Deßloch, Stefan, 2008. Towards generating ETL processes for incremental loading. In: ACM.roceedings of the 2008 International Symposium on Database Engineering and Applications.

[14] Simitsis, Alkis, Vassiliadis, Panos, 2008. A method for the mapping of conceptual designs to logical blueprints for ETL processes. Decision Support Systems, Data Warehousing and OLAP 45 (1),22–40