RESEARCH PAPERS
FACULTY OF MATERIALS SCIENCE AND TECHNOLOGY IN TRNAVA
SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA

2013 Number 33

EXTRACTING PROCESS AND MAPPING MANAGEMENT FOR HETEROGENNOUS SYSTEMS

Igor HAGARA, Pavol TANUŠKA, Soňa DUCHOVIČOVÁ

Abstract

A lot of papers describe three common methods of data selection from primary systems. This paper defines how to select the correct method or combinations of methods for minimizing the impact of production system and common operation. Before using any method, it is necessary to know the primary system and its databases structures for the optimal use of the actual data structure setup and the best design for ETL process. Databases structures are usually categorized into groups, which characterize their quality. The classification helps to find the ideal method for each group and thus design a solution of ETL process with the minimal impact on the data warehouse and production system.

Key words

data warehouse, ETL process, categorisation

Introduction

When the company has already decided about the integration of the data warehouse into their structures, there is a long period of hard work which does not stop even after the implementation of the data warehouse. Creating a data warehouse as a process of creating a logical and physical model is a separate chapter, which is documented in a number of different theories. In this article, we would rather deal with the way of how to establish a data warehouse to a company and integrate it into the existing systems. Actually, important and necessary are sensitivity and knowledge of the internal structure and even the status of different production systems as the source for a data warehouse.

Integration of Data Warehouse into production system

Terms ETL (Extract - transform - load), respectively ETT (extract-transform - transport) process are widely used for integrating the data warehouse into the production systems.

Ing. Igor Hagara, PhD., doc. Ing. Pavol Tanuška, PhD., Ing. Soňa Duchovičová – Institute of Applied Informatics, Automation and Mathematics, Faculty of Materials Science and Technology in Trnava, Slovak University of Technology in Bratislava, Hajdóczyho 1, 917 01 Trnava, Slovak Republic
e-mail: igor.hagara@stuba.sk, pavol.tanuska@stuba.sk, sona.duchovicova@stuba.sk

Pump provides us with the correct and if possible quick loading of data into the warehouse. This process is much more extensive and complicated than it might appear at first sight, and ultimately causes most of the problems, especially during the design phase. Its role and benefits are undeniable.

Data Pump is designed to:

- Extract data
- Transform data
- Data Transport

Management of the data pump creation

Function of the data pump is clearly defined, but the goal of this article is to minimize influence to production systems.

For the complete implementation of all required functions and the loading of data into warehouse without affecting the operation in the company, we come out from the defined state. Therefore, in production, we have to identify the window in the main operation, either overnight or at early morning hours.

Also to be taken into account is, that time is provided not only for filling the data warehouse, but also for running the operation. From this requirement, it is quite clear that the complete recovery of data in a data warehouse is impossible due to the lack of time window for a single filling. Data warehouse has to provide actual data for a reasonable price during its operation, and, at this moment, we are unable to provide it.

On the other hand, the total recovery method could be appropriate, considering the fact, that the primary system requires no changes in the database structure and, in this direction, the data pump would work as a copier.

In addition, a very interesting solution is when we use everyday backups which are made in every company. They should work in parallel environment, and after completing the remaining resources, using the same parallelism, we would achieve all the required data in time. The given operation would be more expensive by buying the necessary licensing, hardware and governance of other systems and so on. We did not reject this opportunity because we know that the multinational corporations providing solutions in the area of data warehouse obstacles do not stop and use the possibilities of accurate copies of the production environment, because the total cost of establishing and operating a data warehouse is covered by the buyer. The Figure 1 describes a usual process of corporations using a standard model. This model is fast an easily applicable to every customer and system, but it is not efficient in comparison with performance versus price. In this process, there is a method which provides the easiest way of extracting data from each source. Usually, every next step after extracting is impacted by this technique of extraction and it can be mostly seen in load and transfer. Process is not optimized for ETL tool, but for source system, which increases the budget for maintenance and makes improvement more complicated.

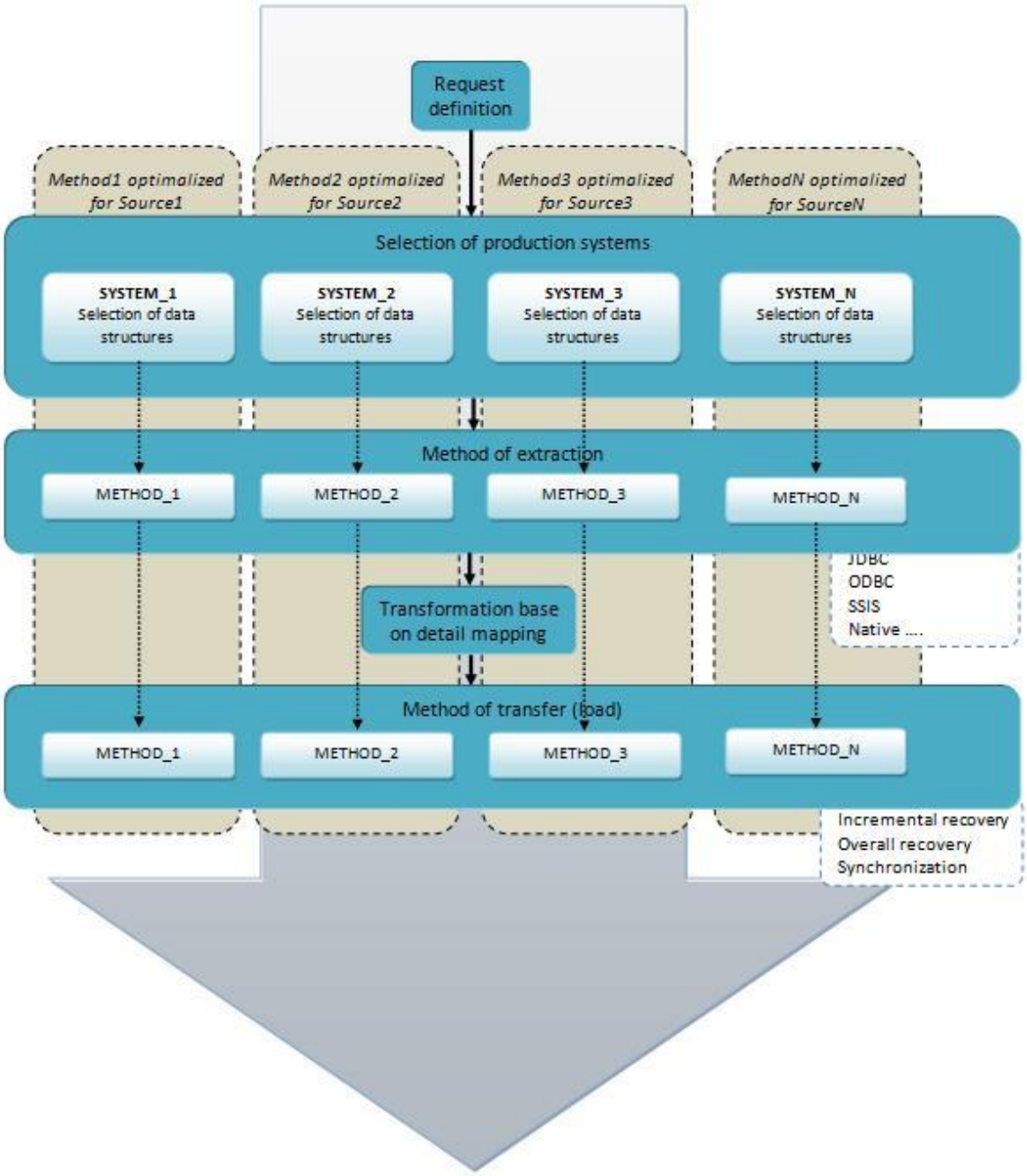

Fig. 1 *Selection process for method of extracting used by multinational corporations*

We ourselves selected a way how to effectively use the space at night, respectively in other empty hours and opportunities of data pump and its components. If we decide to intervene into the production systems, it could have considerable consequences such as the slowdown during normal operations, but also during night operations (5). In this regard, it is also necessary to have on mind a longer-term growth of data and also the length of the processing of runs. Having considered all the facts, the combination of both ways, i.e. the overall recovery and incremental recovery, seems to be the most reasonable way. Thanks to its variability, the data pump, will choose which part will be managed and in which way. To fully use the variability, it is necessary to define the process of data selection (4).

Identify primary sources and tables

The first step is to identify primary sources for data warehouse. List of the sources is based on customer requirements. Designer has to make a list of sources and reduce the effect of one perspective. By analysing the systems, the designer gets a high level analysis of architecture. This process helps better categorize the primary systems into the main systems, the systems with a support role or the systems with redundant data and the systems which are important for overall reconciliation. We also cannot forget the various other sources, except for conventional production systems, there are different external inputs mostly defined as manual inputs. Those might be different Excel spreadsheets or other databases, possibly code lists that were hard coded in the code due to the poor primary system architecture. Structure of these resources is usually the most difficult to collect and define. On the other hand, they can bring colossal value to the company. The manual inputs are then treated individually, while we are trying to unify them from regarding their structure. This can help us later when defining interfaces and the process of transformation. Consequently, it is necessary to determine how the data is transferred from the individual primary sources into the basic layer of data warehouse. In this case, it is easier to define a single path for all sources, what is then easier for management of data pump and mainly for maintenance. It is not necessary to look for a perfect way of selection of each source, because during the life of data warehouse the technology will evolve and come with a better solution for a source system. What is more, this method of finding the ideal solution of data transfer could increase the time for implementation and make the maintenance more complicated. It is more correct to choose one stable and verified solution and if possible, use it for all sources (2). Important in transferring data into the data warehouse is to set up independent process for obtaining data from the source system and independent import the data into a data warehouse. Their independence is important for two reasons. The first one is not to affect normal operation. The other reason is an independent import of data from the source systems into the data warehouse. This independence provides us with flexibility in processing the data warehouse, which does not affect the primary operations of the systems. With the correct time management, it does not affect the users of data warehouse either.

It is necessary to realise that the processing of the data warehouse is usually computationally more demanding as for the capacity, rather than the operation of primary systems, and independence is therefore desirable.

Next step is to identify the individual tables in the production databases necessary for data warehouse.

The production systems are often extremely overloaded with historical data, tables supporting the application run, various auxiliary calculation tables and other redundant data warehouse. It is necessary to exclude these tables from the data transfer pump, thus saving some time and also storage space. This removal of redundant information provides us with higher speed, when the record to data storage takes less time during the copying of data. After initial cleaning the data structure lists, selected tables are divided into several groups, namely those that have a low number of records, tables with higher number of not very frequently changing records, tables with a higher number of very often changing records and tables with high number of records.

Aggregation process and mapping management

The preceding paragraphs described the principles that allow us to correctly define resources and loading the data readings which are closely related. These are just parts of the ETL process. It continues with a detailed mapping of the entity (i.e., aggregating source systems into the data warehouse), so that data is ready for reporting. The described process is independent from the chosen method of building the data warehouse, using either a top-down method (Inmon method) or bottom-up method (Kimball method).

Entities are defined by reporting requirements of customer and represent the subjects of interested subjects from the customer's point of view. Each entity has its own attributes (values), which represent this entity. Every single attribute is mapped to the source systems while mapping all the values from the source system.

Mapping in this sense is defined as the implementation of transformation rule in metadata in an ETL tool programming language.

Mapping ends up with the process of setting up historization, in order to minimize the volume of data with actual value and therefore optimize the performance of data warehouses. Historical mapping takes place to clearly define the rules which are set up.

When describing the aggregate mapping, I purposely did not mentioned any schemes, facts, metrics or dimensions, because the same method will be used whether we decide to go through integration layer, or we do mappings directly into stars schemes; mapping can be also used between the integration layer and stars schemes. The difference between the above-mentioned layers is in the depth of aggregation of the given entities, and therefore also in the complex pre-calculations of data.

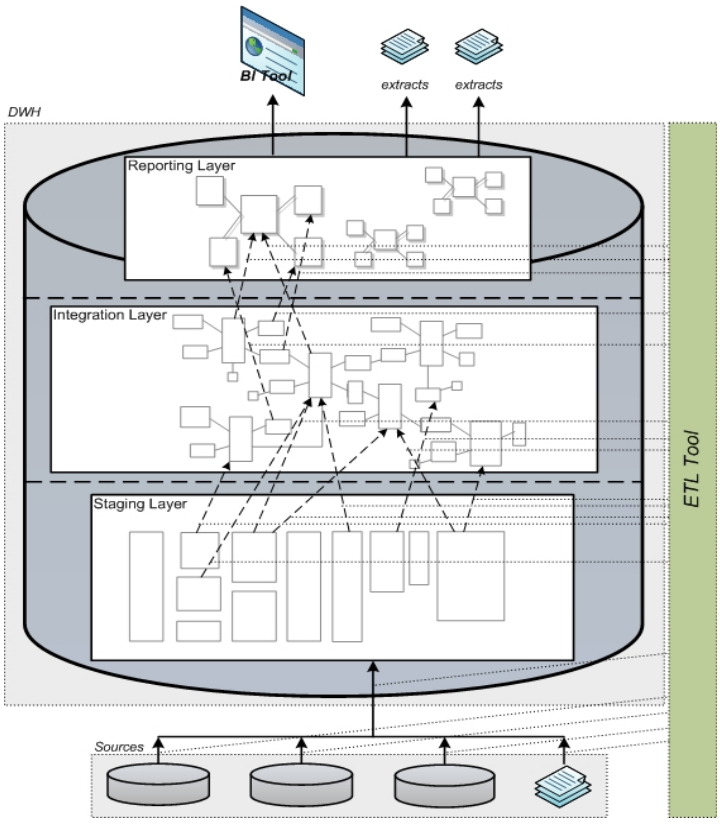

Fig. 4 *ETL process from extraction to higher layer mapping*

The mapping process depends on various circumstances, unlike the previous process, which should provide us with the data independent from the primary systems. A process map of loading and processing the individual entities allows a flexible reaction to data processing in case of data error or other unexpected situations.

The ETL process does not end when data is loaded into the data warehouse, but passes across the entire spectrum of the data warehouse and ends up with history mapping of the most aggregate entity. To facilitate the operation and readability of such an extensive process, various ETL tools are used.

Conclusion

All data structures of source system should be categorized into groups. After categorization, it is necessary to think about what various primary systems provide right now, and what the limits or benefits of the current solutions are. The process proposed in this paper provides a unified solution of how to be more efficient in categorization and extraction of data from sources. It helps remove a complicated design of data extraction, and thus focus on the specific features of the solution, such as mapping data in the data warehouse. Proposal of process provides a solution support independent from the system or database platform, data structure and data model of sources. What is more, this proposal supports uniformity and simplicity of solution and it can be used also for the maintenance and improvement of ETL process.

ETL process does not finish even after loading the data into the data warehouses, but it is necessary to focus on the correct management of the integration of data into entities. Here, it is important to know the rules of filling the attributes of the aggregated entities of the primary systems. The dependencies created by the integration of multiple systems ETL tools are used to manage and administer these rules.

Since we provide solution not for a few months, years or decades, it is necessary to carefully analyse the situation, especially in general areas such as export, loading data, handling ETL processes and dependencies. This does not lessen the importance of mapping the data into an aggregated entity, because we need to ensure the correct load of the data warehouse even in aggregated layers.

References:

1. AKASH, [online]. [cit. 15-09-2012]. *Methods of incremental loading in data warehousing*. DWBI Concepts (2012), Internet source: <http://www.dwbiconcepts.com/etl/27-basic-etl-concepts/109-methods-of-incremental-loading-in-data-warehouse.html>
2. GREENFIELD, L. [online]. [cit. 20-10-2012]. *Maintenance issues of data warehousing systems*. LGI Systems Incorporated, Chicago (USA) (2012). Internet source: <http://www.dwinfocenter.org/maintain.html>
3. KOTOPOULIS, A. [online]. [cit. 21-11-2012]. *Best practices for real-time data warehousing*. Oracle Corporation, Redwood Shores (USA) (2012). Internet source: <http://www.oracle.com/technetwork/middleware/data-integrator/overview/best-practices-for-realtime-data-wa-132882.pdf>
4. Microsoft Corporation. [online]. [cit. 20-10-2012]. *Extracting data from operation systems.* Internet source: <http://msdn.microsoft.com/en-us/library/aa906009%28v=sql.80%29.aspx>

5. Execution – MiH. [online]. [cit. 19-10-2012]. *Data warehouse ETL extraction*. Internet source: http://www.executionmih.com/data-warehouse/etl-extraction-design.php

Reviewers:

doc. Ing. German Michaľčonok, CSc. – Faculty of Materials Science and Technology in Trnava, Slovak University of Technology in Bratislava

visiting Prof. Ing. Miroslav Božik, PhD. – JAVYS, a.s., Bratislava