HETEROGENEOUS DATA TRANSFORMING INTO DATA WAREHOUSES AND THEIR USE IN THE MANAGEMENT OF PROCESSES

Pavol TANUŠKA, Igor HAGARA

Authors: ***Assoc. Prof. Pavol Tanuška, PhD., MSc. Igor Hagara***
Workplace: ***Institute of Applied Informatics, Automation and Mathematics, Faculty of Materials Science and Technology, Slovak University of Technology***
Address: ***Hajdóczyho Street 1, 917 24 Trnava, Slovak Republic***
Phone: ***+421 (33) 5447 736 kl. 320***
E-mail: ***pavol.tanuska@stuba.sk, igor.hagara@stuba.sk***

Abstract

The main issue of this paper is data pump and its design. Data pump is input part of large-scale data warehouse and it is the first narrow point of it. For continual optimal business is necessary owned good design of data pump. Usually it is disvalue its responsibility, but on the other hand, if we have bad data in warehouse it means, that we will have bad report for strategic decisions. Also we cannot forget the rate which is data provided by pump. It is possible to happen that restoration of data warehouse will have been claimed disproportionate amount of time.

Key words

data warehouse, data pump, responsibly of data pump, design of data pump

Introduction

The practice is increasingly faced with the concept of data warehouses. It is mainly due to the optimization of processes and management. Still a large number of firms manage large-scale information systems and quantum data. These data are produced daily, whether in manufacturing or services.

It is still a few people in senior positions in companies that are not aware of the possibility of processing the data and their subsequent use for strategic decisions. It is often satisfied with the already existing and operating processes and systems, while frequent these were not and cannot provide sufficient information about production. Just only to realize that with increasing amounts of data is necessary to stored data somewhere outside, because theirs begin to make slower production systems. But we can handle them in the further development

of the company. It is also clear that a larger amount of data also means more of the various sessions between them and thus more dependencies.

In this process of accruing data is beneficial select which are the envy of the entity are less important and which essential. Since this process can then be reflected to other challenges the company and determine where their objectives to optimize processes. Such an approach reflected in the market more competitive.

Data Warehouse

A data warehouse is defined by Bill Inmon in 1990 as: "A warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process" [3]. On the web you can find a number of precise definitions and explanations of the data warehouse, and understand what is intended. In fact, this is a special type of large-scale relational database that is exclusively designed for analytical queries. If we want provide relevant and accurate information at the right time by data warehouse is extremely important to fill up it with the right values and the correct method. In practice I have often met, when the company had introduced a data warehouse but the filling was not commensurate with the services which provide the data warehouse.

For independent business as usual of the data warehouse is more requirements than just correct it right design. Its comprehensive schema you can zoom in on the next figure.

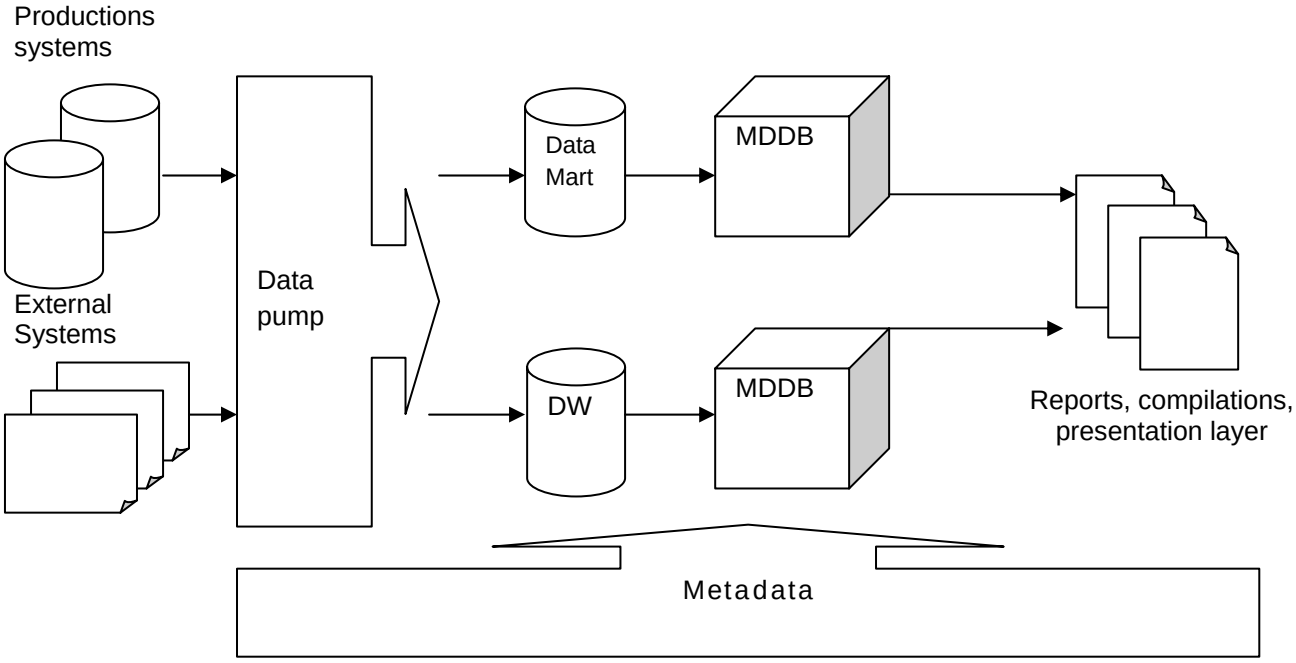

Fig. 1. *Data Warehouse*

The final product, which is interesting for the customer is the output, i.e. reports, compilations and presentation layer. Before it we found application layer. Since the data warehouse is typically modeled as multidimensional database structure there is also a representation of presenting multidimensional cube. This is realized via OLAP server. Application layer is to be built on a data layer presented data warehouses or data marketplace. They are mainly used in large companies, where we require finer breakdown of the data

warehouse, for example regional, provincial directorates. The first part of the data warehouse is show as a data pump, which provides data transmission. Metadata are information about the system data and data model.

Data pump

As can be seen an important part of the data warehouse is a data pump, which I would like to focus. Also called as the ETL (extract - transform - load), respectively. ETT process (extract - transform - transport). Pump can deliver a correct and, if possible, the speedy saturate with data of the warehouse. This process is much more extensive and complicated than it seems, and ultimately leads to some issues most especially in the drafting. Data pump represents a weak point and when we passed the terminology in the field of logistics can be rightly considered as the narrow space. It is for this reason that the customer sees at first glance his office and takes only a kind of pump data and the need to reduce its functionality. What is to the detriment of the case, this procedure and opinion, you also find by a database of experts and make its role easily like: download and transport data. In doing so, its role and contribution are not.
The data pump is responsible for:

1. Extract data - it's the ability to connect to the production systems and to select relevant data from them. Production systems could be built on different platforms, and also generate a different data format. This is often seen in particular in production systems supplied by different suppliers. Moreover, transaction systems to generate the amount of redundant data and is therefore essential to choose the correct method of choice. This is crucial in downloading the data warehouses.
   For the extraction of data are often used native approaches and utilities due to the diversity and the optimization of the process. But it is also possible to use other interfaces such as JDBC, ODBC, SQL, etc. gates. It should be taken into account when and what type of use with respect to performance.
2. Transforming the data - this role is continuously builds on the previous one. The potential of the selected data to verify their validity and clean them from any mistakes that we could introduce into the data warehouse. This action is like a feedback for the transactions systems whether such systems generate errors. Of course, then there is a process of integration and transformation. Data from different sources are consolidated in a consistent forms and structures, or within one destandardized whole, which is used as the starting input for the data warehouse. This process appears to be calm and just theoretical in good simple definition. But in practice it is about a separate access to each source. Often we have to talk of poor systems, which the company developed as ad hoc programs. Above this selection then we can perform a few calculations and adjustments. This is particularly the assignment of their own identification numbers, non-homogeneous data updated by the internal code data pump, or even the application of certain aggregate functions and set the index for optimization.
3. Transporting of data - the data pump feature, which is for the customer about the most visible and most tangible. It is the actual data transfer between the systems. Among the transport of data include the initial migration of data and then meet. Alternatively, we can not forget even for the update.
   There appears to be very important to achieving the correct method:
   - Total renovation
   - Incremental restoration
   - Synchronization.

Design of data pump

Each of these methods has its advantages and disadvantages and in practice I have met already several variations. Total restoration appears to be the simplest solution. However, it is not appropriate for extremely large-scale production systems. Such a method is too long and too many times to be addressing the data warehouse is not updated on a daily basis. Incremental method and synchronization of the reason seems to be acceptable. In finally result they are much more difficult not only to implement but also for maintenance. It is very important to choose the appropriate method by which the processes are set in the company. In the event that we require reports on a monthly respectively weekly basis is inefficient to use an incremental method for the synchronization. In this regard, we would unnecessarily overpay to a service provider or they seek. For daily reports and analysis is again extremely important to have current data.

The following diagram illustrates only the actual ETT process.

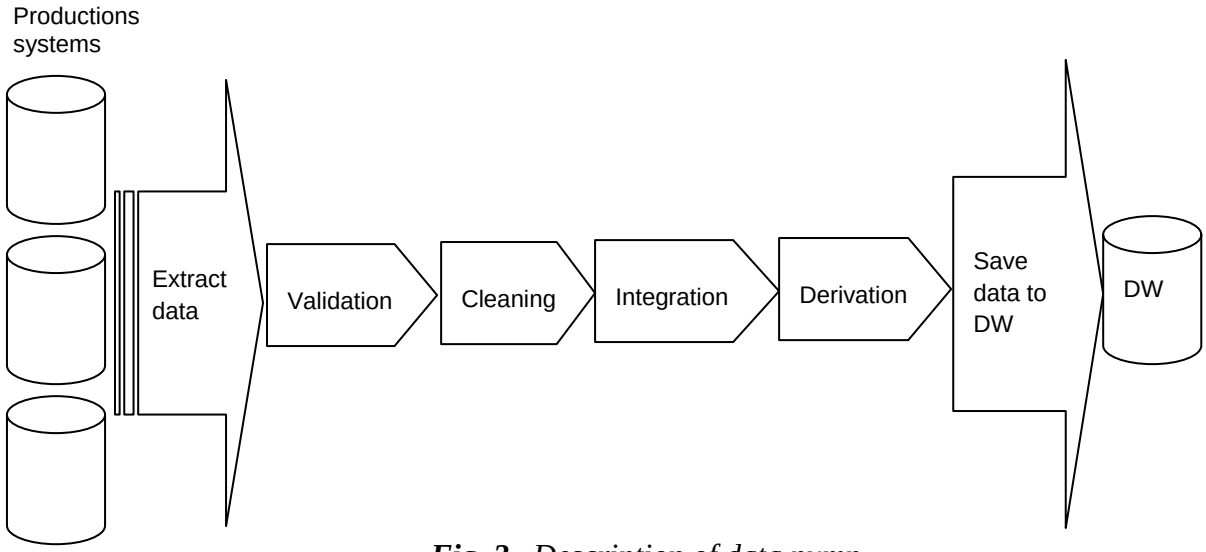

Fig. 2. *Description of data pump*

When we want build a data pump, we must also reflect the fact that we have it running in the least affected by production systems. This means a suitable choice of what data will be extracted and where the data will be extracted. As I mentioned do not always require all the data and we sort out from them the information. In rare cases, to capture such a choice of transaction systems require modification or even its extension to the new structure. Here we now present the question of what with redundant data make in primary databases that don't decelerate the production.

The second important point is the place where the already selected data will be processed and transformed. Aggregate calculations are always a burden on the hardware used. If we decide to transform the data to the transaction system has again been hampered production. Solution process for entering data into the data warehouse is the same sort of optimal mainly because of setting the server and database platforms. Server for data warehouse is an ideal setting for their needs. Indeed, it seems best reserved small area just for a data pump, which is important, however, to separate the data and application layer. Then in this constellation it is possible to pump data from the data and stored for some time in the archives. While this option seems redundant, its role is particularly useful for implementing a data warehouse. For

deferred data is to determine whether the pump is set correctly in the event of an issue it is possible to evaluate error and where it is better to analyze and make corrections. Whether it occurs on the production systems, data pump or even the warehouse.

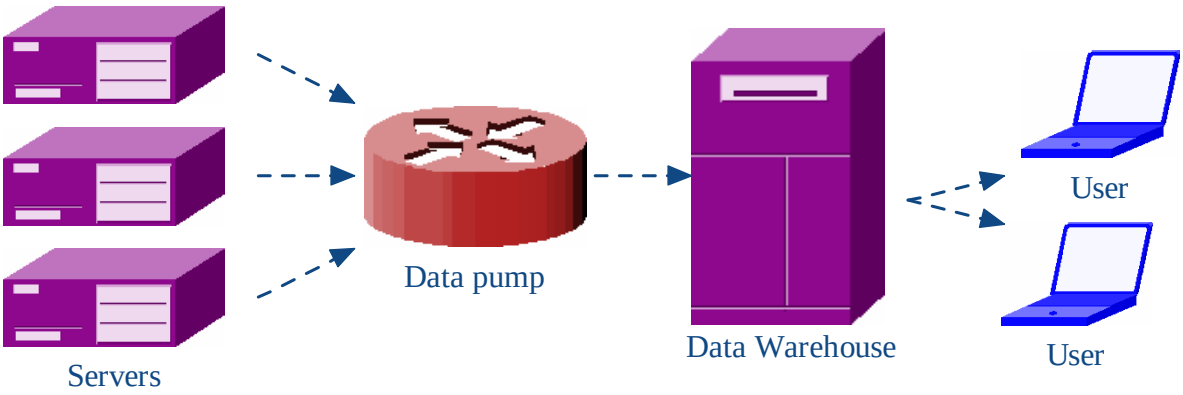

Fig. 3. *Architecture with separate data pump*

The biggest points occur particularly in poor or lack of definition of data that require the production systems. This is the classic point between customers and suppliers. However, here has reflected the strength of good definitions in the data warehouse, and bear remember that it has destined for strategic decision making. It is frequent that the implemented data pump must be spreading. In such case it has been proven to implementation modules. When a new module can be added at any time, respectively existing module can be upgraded. This solution is advantageous also from this perspective, if eliminated a production system. In this case, the right setting a scheduled task to transfer data from other systems and defective complement system after repair. Of course, that during this period will receive only partial or distorted output. On the other hand, if the production system will be dysfunctional extended period of time, in the final assessment may not necessarily be completely weaned data warehouse, respectively performed difficult maintenance. At this point, it seems equally important, the pattern of feeding the data warehouse. Not only for the crisis the question do what, if a production system falls. Basically shows that we are able to continuously expanding the production and production systems to ensure sufficient time to meet the warehouse. It is necessary only in the growth internally within one organization may be to fuse two or more companies each will use its own information technology. Of course, here comes the entire amount of solution, it is clear that in this case, there must be a central data warehouse and the resulting adjustment is either used or developed new what appears to be non-perspective. In this process it actually has closed the circle of development and maintenance of business terms.

Special functions and providers

It is thus obvious, that each data pump has to be special and manufactured precisely tailored customer itself. Not considered appropriate to use the "generalized pump". For solitary data warehouse can be used in this theory but for the unique individual systems to pump data in almost impossible. Otherwise, such a solution can affect the performance and optimization. This is because of issue that is mostly inedible by only a small utility but a comprehensive program, which has the application interface. They can be programmed in conventional languages such as C #, C + +, Delphi or ever extended web technologies like

NET, Java and so on. But solitary business end has been hidden in codes in the revised procedures and SQL. During this issue should be mentioned that the special features of the data require a pump. It is particularly detailed planning of individual tasks and then it is to check whether calls did not fall down, which is in the optimal solutions to ensure client information via email.

Given the financial requirements to design and implementation of data warehouse is the product mainly granted to large multinational corporations such as Oracle, Microsoft, Compekon, IBM etc. The experience of the whole world can translate into rapid and effective deployment data not only pump but also the entire project, which is associated with the introduction of the product. Prices in the data warehouse and any of its components, we must count on the thousands to tens of thousands of Euros, while the actual price is negotiated for specific orders. It is however a possibility to find a smaller specialized companies, for example MF Support, AR MANAGER, PVT, which brings the possibility of hiring a new data warehouse using the ASP model. Such a solution is acceptable, in particular for small and medium enterprises. This does not mean lower financial costs, it is therefore necessary to consider all offers, testimonials and guarantees.

Conclusion

When designing a data pump, it is important to convince the customer of its importance. Experience shows what efforts are not necessary to make the case that was not properly designed or was undersized. In rare cases occur even its redesign and re-development. Such treatment is obviously bad business card not only for the service provider, which usually has been constrained with unrealistic deadlines by the customer. In addition to the customer and the term breach of the exponentially outweigh the costs.

References:

[1] Dátové sklady a jejich optimalizace, [online], [cit. 29.5.2009], Available on web:<http://www.systemonline.cz/clanky/datove-sklady-a-jejich-optimalizace.htm>

[2] Charlie – dátové pumpy, [online], [cit. 29.5.2009], Available on web:<http://www.systemonline.cz/clanky/charlie-datove-pumpy.htm>

[3] A Definition of Data Warehousing, [online], [cit 29.5.2009], Avialble on web:http://www.intranetjournal.com/features/datawarehousing.html

[4] INMON, W.H. *Building the Data Warehouse.* Wiley Computer Publishing Inc., 2002. ISBN 0-471-08130-2