System approach to planning and implementation of enterprice data warehouse

Viktor Gopeyenko [a] , Maksim Levchenko
Department of Natural Sciences and Computer Technologies
Information Systems Management Institute, Ludzas 91, LV-1019, Riga, Latvia

Received 1 February 2011, accepted 23 February 2011

Abstract. Nowadays the importance of information exchange, accumulation and processing in modern business is increasing. It is hard to imagine a big company not having tens, hundreds or even thousands of terabytes full of accumulated data. Server systems of large multinational corporations demand significant costs to maintain and apply huge amounts of the accumulated data. The question is: how are these amounts of data formed, are they really necessary, and how can they be disposed? Therefore, in the long run company's management is to face these questions because the storage of information requires certain expenses, which increase depending on the growing amount of data.

Citations: Viktor Gopeyenko, Maksim Levchenko. System approach to planning and implementation of enterprice data warehouse – *Innovative Infotechnologies for Science, Business and Education,* ISSN 2029-1035 – 1(10)2011 – Pp. 10-12.

Keywords: Online transaction processing; OLTP; data warehouse; DWH; business intelligence; BI; data mining; systems analyst; decision support system; DSS; online analytical processing; OLAP; relational online analytical processing; ROLAP; multidimensional online analytical processing; MOLAP; structured query language; SQL; Oracle BI; OBI; Microsoft SQL Server.
Short title: System approach.

Introduction

In order to answer these questions it is essential to understand the nature of information and its content. Everything is simple when concerning emails, working papers and files: if they are relevant and valuable, they are stored, otherwise they are deleted. In turn, *online transaction processing* (OLTP) systems contain data with completed transactions and directories necessary for the formation of these transactions. From the point of view of OLTP system the concept of transaction includes any information, generated by any system used by an enterprise for its daily activities. As an example we can consider the situation with the sale of wares in a shop. The system captures the information of each item of goods that/which pass through checkout counter. Thus, a huge amount of transactions is recorded in OLTP data storage, which accumulates information from all the supermarkets of the chain. Having clarified the source of the formation of data, it is necessary to consider its content. As a rule, initially all sales contains the minimum of information which is useful for an analysts. For example, such minimum information could be: the date and the place of purchase, the name of the product, the amount of the sale and tax, the number of wares sold.

Having minimum information, it is worth considering how beneficial its systematization and processing can be. A quite serious analytical system can be designed on the base of the aforementioned example. This analytical system could reflect the dynamics of sales concerning the items of wares, compare the current amounts of sales with the amounts of sales of the previous periods, calculate the average purchase amount, the amount of tax paid, and other useful information. In practice, the OLTP systems of big enterprises store much more useful information, which enables monitoring warehouse stocks, logistics, perform financial analysis and accounting, develop marketing activities, etc [1,2].

1. Methods

Let's consider information as data after its representation in comprehended, ordered and focused on *business intelligence* (BI) instruments form.

As a rule, in the majority of organisations, there are several different groups of users that need various representation of the information. *Report readers* prefer to receive answers to the questions in a form of the same report that is well familiar to them. Their approaches to the analysis and to the information requirements are usually the same. Usually, individual help is required to this group of users to in order to start working. The other group of users which could be de-

[a]Corresponding author, email: *viktors.gopejenko@isma.lv*

fined as *information browsers* needs a slightly longer training to develop the means to work with data, however having the corresponding means they are able to prepare simple reports independently. Generally, they do not require powerful tools for the construction of the reports.

Advanced users prefer to build their own queries and reports. The possibilities of aggregated values search are important to them, as in this case they may quickly prepare reports without going deep into the details of the data model. Power users constantly put forward new demands for the data storage. They make considerable impact on all aspects of *data warehouse* (DWH) activity and aspire to have the information concerning data units. This group can be distinguished by the most intense usage of metadata repository and is the most interested in data mining.

One of the errors made by the developers of many DWH projects is an attempt to select the only software as a standard for all users. Usually it is better to select several instruments for different DWH users. As a result, the price of these programs will be significantly lower than time costs of system analysts.

The means of data access might also be concerned to various groups. It might be the general purpose means to manipulate multidimensional data arrays or more complex *decision support system* (DSS). Data mining means are used to find regularities in the data. Data visualization means represent data in a way, that helps to reveal these regularities.

Applications that use data from DWH and represent information to end-users are concerned with the means of *online analytical processing* (OLAP) group [3]. The main reason of OLAP usage for the query processing is the working speed. OLAP creates a DWH relation snapshot and restructures it into a multidimensional model for queries. The declared time for query processing in OLAP is approximately 0.1% from similar queries in the relational database. This means that use data from the relational databases are classified as *relational online analytical processing* (ROLAP) instruments. ROLAP is an alternative to the *multidimensional online analytical processing* MOLAP technology. While both ROLAP and MOLAP analytic tools are designed to allow the analysis of data through the use of a multidimensional data model, ROLAP differs significantly since it does not require the pre-computation and storage of information. Instead, ROLAP tools access the data in a relational database and generate *structured query language* (SQL) queries to calculate information at an appropriate level when the end-user requests it. The access means that use multidimensional databases are concerned with *multidimensional online analytical processing* (MOLAP) group. MOLAP is an alternative to the ROLAP technology. While both MOLAP and ROLAP analytic tools are designed to allow data analysis through the use of multidimensional data model, MOLAP differs significantly as it requires the pre-computation and storage of information in the cube.

To create an analytical system an idea is needed as an integral part of a final product. The correct selection of BI instrument is the other integral part of the successful development of the project. In the modern world various analytical solutions are available from multiple vendors, the most recognized among them are Oracle, Microsoft, SAP, SAS, IBM [4]. Globally they can be divided into two groups.

1. *Off-the shelf* solutions are installed on clients servers and adjusted to their needs.
2. *Development kits* provide a development environment only and make it possible to create custom solutions.

Oracle Business Intelligence Enterprise Edition (Oracle BI-EE) can serve as an example of a off-the-shelf solution, which is necessary to integrate in the existing software environment and to adjust to the requirements of each client. Oracle BI is a business intelligence platform which delivers a full range of analytic and reporting capabilities. Designed for scalability, reliability, and performance, Oracle BI-EE delivers contextual, relevant and actionable insight to everyone in an organization, resulting in improved decision-making, better-informed actions, and efficient business processes.

2. Realization

Microsoft SQL Server 2008 including SQL *Server Analysis Services* (SSAS), SQL *Server Integration Services* (SSIS), SQL *Server Reporting Services* (SSRS) can be considered as an example of the development instrument, which is a quite powerful tool to create customized solutions. It is a powerful and reliable data management system that delivers a rich set of features, data protection, and performance for embedded application clients, light Web applications, and local data stores. It is designed for easy deployment and rapid prototyping.

The data storage creation is a long-term and quite an expensive enterprise. When buying the data storage, analytical system or any other project, the customer wants to acquire not a set of algorithms and ready procedures, but a product that allows minimizing its expenses by optimizing working processes and the quantity of workplaces. Correctly adjusted system is able to substitute a large number of workers, freeing human and financial resources. In such a way we get to the first and, probably, the most important stage in the creation of any system.

The first stage of any system creation is its projecting. The customer will not pay anyone for unnecessary options that do not satisfy his requirements. The system must be carefully planned and organized, thereby responding to the requests of the customer. At this stage a plan is created which is a logical structure of the system. This plan is coordinated with the customer and sooner or later after a number of long and painful negotiations is approved.

The next stage is defining the necessary configuration of the hardware and software. This stage is also important since

without the skilfully chosen hardware and software base the created system will not be stable. As a consequence the periodic unforeseen faults in the system are possible. As a result the customer will bear losses which might be addressed to the developer of the system. Usually the contracts accompanying such projects clearly specify such nuances.

Then when everything is approved, we have the action plan; it is time to realize it. At first it is necessary to choose the team of developers, select the specialists with appropriate skills. In practice, the necessary number of specialists with certain (sometimes, specific) skills is not always available. Large companies with a large number of employees in different branches organize multinational groups. Smaller companies which, as a rule, have smaller capabilities either search for corresponding personal on the labour-market or join with other developers. As large companies have enough financial resources for their project realization they, usually, cooperate with other large companies.

When the team of developers is created and the responsibilities between them are distributed, it is possible to start the creation of the project. If the customer already possesses any data set, a part of array which is enough to develop software and to perform the test operations is taken from this data set. Then, as a rule, the end product is created as modules. This is a long-term process that can take a long period of time depending on the desired results. Usually, the average duration of the project is defined by the time period of one or two years. A large project may take two years or even more.

Of course, after the creation of the system it should be tested. This is an important step in its creation as a poorly working project is useless. During testing, the system is checked according to many parameters. The main testing parameters are the correct operations of mechanisms which form the system as well as the tolerance to stress loads. Paradigma *correctness of operations* means that all mechanisms (interfaces, services etc) should work without any significant faults and system overloads in any cases. *Tolerance to stress* loads is a relative concept, although it might be clearly measured. Initially the system tolerance is checked at the stress level defined by the customer (e.g. the number of transactions or connections). Then its working capabilities are checked at double load. And finally, the testing is performed at the maximum load level in order to understand when the system becomes inoperative.

The last and the final stage is the system support. Usually, after the creation of a large-scale project, the customer desires to get not only the warranty of its operability but also to get its support. Generally it is a long-term cooperation which is specified for several years ahead.

Special attention should be devoted to the security question during the project development. The best approach to such approach is formulated as follows: *The confidential information is the information that is not defined as non-confidential information*. The Developer company should thoroughly protect clients' data preventing any possibilities of the information leak. Usually, there is a set of clear instructions, the main idea of which is that in order to evade the problems it is better not to create them initially.

Conclusion

The choice of the software is one of the key features during the planning of the analytical system, as the functionality, licence costs, development and support costs, development perspectives depend exactly on the software choice.

The implementation of an analytical system at an enterprise is a rather expensive activity that requires involvement of significant financial and human resources. If the approach is correct, the return of investment is high and fast. A correctly adjusted system can perform a large set of different operations repeatedly, the fulfilment of which could require the work of several departments.

Due to the automation of the calculations the cost of data processing is significantly reduced. Other advantages are as follows:

i) a significant increase in calculation speed;
ii) an increase in the availability of data (the portal of reports can be seen in the Internet);
iii) an opportunity to obtain information in real time (during the report formation);
iv) an improvement of security (the data is available only to those staff members who have access).

Consequently, the presence of the data array and the lack of its usage strategy create prerequisites for the OLAP solution planning.

In the modern world, where in order to be successful in business it is necessary to be faster, wiser and more flexible than your competitors, exactly OLAP is capable to help to achieve success.

References

1. John M. Coe. The Fundamentals of Business-to-Business Sales & Marketing. – New York: McGraw-Hill, 2003.
2. Daniel Amor. The E-Business (R)evolution, 1st edition. – New York: Prentice Hall, 1999.
3. Hanson Ward A. Principles of Internet-marketing. – Cincinnati: South-Western College Publishing, 2001.
4. Eric Sperley. Enterprise Data Warehouse: Planning, Building, and Implementation, volume 1. – Upper Saddle River: Hewlett-Packard Company, 1999.