

DATA WAREHOUSE PDF

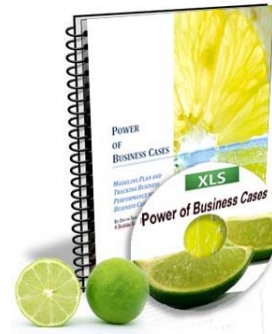
By Kresimir Futivic and Bozidar Kralj



Intention of this document is to enlarge knowledge about information generation, influences on quality, to highlight potential problems and solutions, and to present at one significant side of the company processes that definitely affect qualities and values of decisions. So, the Data Warehouse PDF is made for brief introduction into complexity of information management.



If you want to jump directly into innovative solutions visit our products





LEGAL DISCLAIMER: Refer to the Company terms of Service and other legal documents, for Company limitations on liability and other important information. Company Terms of Service can be found at www.business-intelligence-secrets.com website. The information presented is for illustrative and informational purposes only. Use these techniques, advices and strategies at your own risk. Making decisions based on any information presented should be done with knowledge that you could experience significant losses, or make no money at all. You agree that our company is not responsible for the success or failure of your business decisions relating to any information presented.



What is Data Warehouse?

Data Warehouse Definition

The Data Warehouse is a database of unique data structure with advanced data archiving capabilities that allows relatively quick and easy performance of complex queries over large amounts of data.

A classical, production information system is primarily built for gathering data inputs and their processing. It allows the company to be operational and run smoothly, and that means mostly data entry.

On the other hand, the data storage system in its structure allows fast and easy retrieval of large amounts of data. This makes it suitable for the construction of the so-called systems for business decision support (DSS – Decision Support System).

In addition, what is Data Warehouse?

The data stored daily in the production system must ultimately serve the administrative structure of the company. Administrative structure of the company should be able to extract useful information from large amounts of data, which she will serve for evaluation of the results achieved, planning and decision-making. For this purpose it is necessary to ensure a quick and easy access to data stored in complex structures of production systems. Data Warehouse provides just such a mode that is faster and easier access to information, to view and analyze large amounts of data, in which time measures the reach in seconds or minutes.

In building the data warehouse, implementation have to meet specific problems that do not appear in the construction of production (transaction-oriented) information systems. Most problems are related to the construction of a system for extracting data that is a periodically automated transfer of data from the source to the destination of the production system data warehouse. Some of the problems encountered in the construction of warehouses are:

- Integration of diverse data from multiple sources (multiple production systems) implemented on different platforms;
- Rapid detection of the changes in the source system;
- Iterative nature of model-building data warehouses and hence the iterative nature of building software for extraction.



Problems related to the construction of a data model are quite well-described in the literature and is not a problem too. On the other hand, problems related to extraction of data represent the biggest challenge, making the process of building Extraction System takes between 70 and 90 percent of the total time required to build a warehouse. When we add the problems that arise due to the iterative nature of building models and data extraction systems, we get a system which is very difficult to accurately determine the limits of construction. This is one of the reasons why the project of building data warehouses, as it turned out in practice, is largely subject to failure.

Benefits of Data Warehouse

In companies there is an increasing need for presentation of larger and larger amounts of quality information contained in production systems. For now in many companies, a part or major parts of information from production systems are exported in the data warehouse – Datamarts. There is a constant need for expanding existing and constructing new Datamarts that would include production systems that are not yet covered by existing Datamarts. The need for new Datamarts is recognized by management and employees at all levels of the organizational structure. Therefore, the article tries to explain how and when company can get involved in the building of a new segment of Data warehouses – Datamarts, and what are the dangers and benefits of data warehouse construction.

Decision support systems are fast becoming the practice in most companies. The fact that should be kept in mind is that investment in a large data warehouse is often greater than in any of the applications that the company currently uses. Initially it might seem not so big of an investment, but one must reckon that once built, the warehouse still does not work by itself and requires the attention of one or more persons depending on company size and number of users that use it. Adding the working hours of several people involved in cleaning the data identified as "garbage" and optimization of response time to inquiries and similar activities for which a warehouse is several times more expensive. Without these processes, data warehouse benefits would be significantly decreased. On the other hand, engaged and responsible persons for the data warehouse must be a person from the company who knows how to operate production systems – data sources for the DWH. Depending on the number of such sources and the time and comprehensiveness of data in the warehouse, such employees are often very valuable to the company.

Buying Data Warehouse or Building It With Own Resources?

The strategic value and advantage of data warehouse

The company in its daily operating-production systems collects large amounts of data. This information may be of different shapes, e.g. data – a person Y paid on 23 March 200 USD subscription service, or Z – sales in continental region has decreased for 11 percent compared with the same period of last year. Data provide a picture of what is happening in a particular segment of business enterprises. After the data are collected, the next important step is transformation to information. This information (later knowledge) allows a company to bring a number of important



decisions for future business based on available information. For example – Service XX is terminated because it created a loss of 12 percent per month. Based on sales data, it is obvious that it is outdated and unattractive to users.

Data Warehouse is a tool and a system of knowledge for people in the company responsible for making decisions. This means that the sooner a company begins to make decisions based on available data stored in the warehouse, the sooner will the DWH project prove its value to the company. Therefore, the out-of-the-box solution might sound like a good idea, but it is, unfortunately, not possible, i.e. it is possible in a narrow sense, but for now let's call the finished product a warehouse.

Reasons for success or failure of data warehouse projects

Studies show that about 50 to 60 percent of benefits of a data warehouse are not as expected. To go a little bit further, data warehouse projects fail to fulfill planned targets. About 70 percent of these failures by warehouses were due to constructed by their own forces. Why is this so, and what are the reasons that a company has its own forces not able to build a data warehouse? Here are reasons why the benefits of data warehouse are not as expected:

There is no insured sponsor in the managerial structures. Top management must be interested in building a data warehouse if the analysis proves the validity of its construction, and must support it with all the resources. Support of top management is reflected in the sponsorship of the project. BI is primarily intended for people who make decisions in the company and their assistants. If they do not want it to use for their services, the project simply collapses. On the other hand, any such project is quite expensive, and without a solid management, structure determination sooner or later will be suspended due to high costs and a large commitment of human resources.

In our company there is no sufficient number of people who can devote 100 percent to building data warehouses. It should be remembered that the construction of a typical warehouse takes one to two years and normally mobilizes 10 to 30 people. Talk to many human resources and it does not bother them. Operational problems through such a long period are usually impossible. If it is possible then these people definitely were not valuable for existing production systems and were in excess from the beginning.

Engagements only own power within the organization without consultants and subcontractors. The idea is that IT specialists within companies know better data and processes in a company of professionals out of business, but it is a fact that the external IT companies unburdened by the problems, and due to the nature of their jobs have a higher and broader experience, especially if you are engaged in a data warehouse for some time. For example, an external IT specialist on the basis of these requirements often can recognize future problems and what needs to be done for their elimination before the real need arises. As previously stated, the construction with own forces succeeded in merely 30 percent of cases and is therefore a real danger of the failure of building a data warehouse if not engaging specialists outside the company.

Resistance is possible within the company in introducing new technologies – data warehouses in the regular work. Resistance can manifest itself from the leadership structure to the lowest levels.



Employees and management who will most use applications based on the technology of data warehousing must be open to new technologies.

Roles of Data Warehousing

To better illustrate different scenarios of using data warehouse, simple model of data warehouse is used of an insurance company. Defined are four dimensions: insured person, time, services, point of sale. Measure is fee billed. In this example, a simple datamart (DM) replaces a very large number of static reports (Figure 1). Datamart will simplify business for persons engaged in regular (eg monthly) reporting at any level.

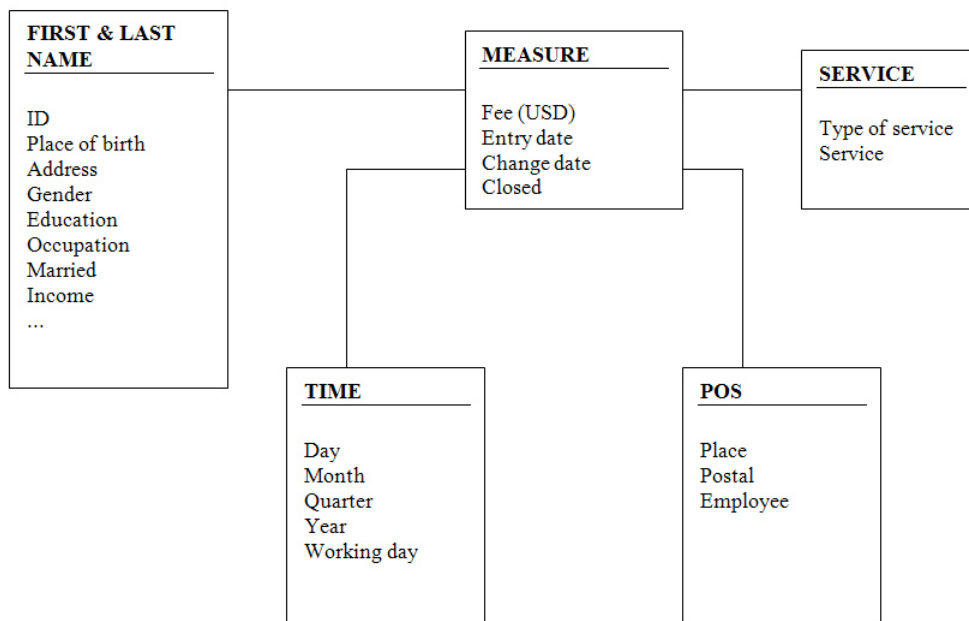


Figure 1. An example of a simple Datamart.

DM replaces also reports that are typically performed only once and never again use the (for example who are insured in place A and have bought XXX service at all retail outlets in place B, or for example which service makes the most money per quarter) Often for various reasons, which are needed within the company, reports that appear to be generated very rarely or only once. These are in fact all statements that do not belong to a set of reports that are standard and serve as leading indicators of business structures in the company, but serve for example to market research or development services, for example at some point company decides to cancel certain insurers, which often damage the cars to raise monthly insurance. From this point of view data warehouse model can be determined with the help of tools for making statements closer to the person who must report. User very quickly and efficiently create reports, which frees up considerable human resources in the IT department. Data Warehouse has several parallel roles (in text below) but mostly during construction of DWH only one role is being favored. Decision about favorite role is based upon recognized need for improvement of the existing information systems companies.



1. Controlling role of DWH

During the process of extracting, data can be controlled and filtered (data cleansing). Data that failed to pass filters are candidates to be the data - garbage, errors. Take for example the simplest filter that checks is entering date of payment of any party in relation with the current year. In most production systems during the year 2000 for various reasons year 1900 was entered.

Filters can be much more complex, some of them may be subject to the business process (non-existent customer can not pay) and by statistics. For example, it is likely an error if one day we have over 120,000 USD. at a particular point of sale where Y is normally charged 10.000 USD dailly. Record of 110.500 USD for application is acceptable, but employees can enter errors and this the case of manual error. Therefore it makes sense to build a warehouse alert system that tells us that sale of the location Y is outside of the statistical framework.

Figure 2 presents schematic process of extraction of data through the filter. These filters are installed in the source code extraction.

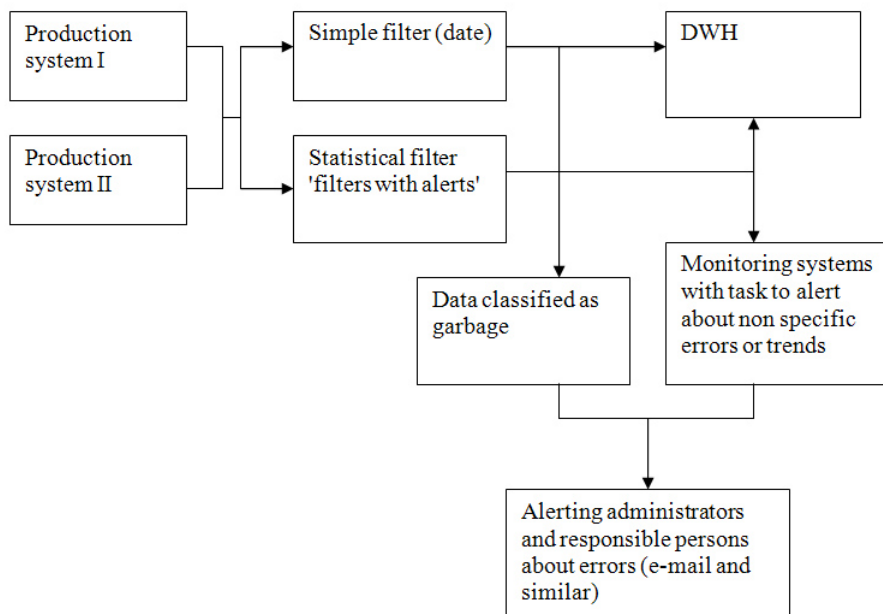


Figure 2 The process of extraction of data through the filter.

2. Strategic Role

Consider for example that the data warehouse insurance data has back up of fifteen years. The amount of data can contribute to defining its knowledge of enterprise business processes with datamining tools. It is possible to discover simple things such as which services over the past n years were the most profitable in the long term and which generated losses. Some services are analyzed through a number of years to see whether the investment in the service was worthwhile. On the other hand, sometimes users can find out more complex data that can not be assumed before analysis, like for example which streets are highly at risk of burglary and which are safe. With this information insurance fee will be settled for clients. On the other side this example shows how DWH role falls again under the control role, because it can identify workers who often make mistakes when entering data (if the DWH contains dimension for each employee entered data).



3. Hardware accelerator

Reporting and data input on the same particular server at a same time slow down operative work of server. Therefore, it is necessary to purchase a new server that is used for reporting only, to be specific for reporting processes only. Fact is that if the server on which the data are entered only will not have any more complaints from users who import data if reporting function is removed from server. Usual complaints will not be any more on slow work that sometimes lead to serious delays in the work. Tasks of reporting are removed on the server which is usually extension of data warehouse.

Data Warehouse as software structure even if it is placed on only one server speeds up the process of reporting due to functional customization for needs of reporting. Because of this feature data warehousing is functioning as hardware accelerator simply due to fact that existing structure of production systems are released of demanding reporting process. Before the introduction of data warehousing statistical reporting usually slowed performance of production systems.

The relationship of market types and technologies of data warehousing

Common idea that the amount of stored data such as measured in Tera bytes (TB) for Data warehousing is necessary to forget. Also if Information system has 4 TB of data for 1.5 million customers in some way be justifies being proud on the company information system, but also recalls for caution since it could also mean hundreds of generated reports and none of them actually were not read. Value of generated reports is not only in number of pages. Information about using 4 TB of data itself is probably useless, ie, it could only mean a large amount of numbers that are in the warehouse from which it has little value if they are of low quality. The primary question that business leaders need to know is what the data warehouse has to serve. For what type of questions data warehouse must provide answers.

Data Warehouse may be a way to display data from production systems that are already adapted to the analysis of data entry – Reporting function.

Another often neglected aspect is the fact that the data warehouse as a system serves for data storage. Of course data warehouse is usually both, but the question that has a primary role. In production system data is deleted periodically, and if it is not then it slow system. If production system is not slow dispat large data quantities this could mean purchased hardware was at the time of instalation overpaid and unnecessary. It is certain that it does not have to keep your payment receipt from permanently disconnected subscribers for decades. On the other hand, these data can be useful in statistical terms, and then it comes to other essential functions of a data warehouse to the prediction of customer behavior based on past behavior. The production system can satisfy most of reporting functionalities (they are more complicated to do than if they are not generated in data warehouse technology) but without historical data. Such reports from production systems can not never supply historical context for reporting. The fact is also that the data storage is somewhat neglected in recent years in the IT industry since the development of explosive speed that has brought it mobilized a lot of IT experts to develop this segment. In following list is presented an overview of usually stored data – frequently deposited information.



- Detailed sales figures for the financial and tax purposes.
- Data on trends of use of services and purchase of goods that have strategic value.
- Information required by various state offices (eg statistical office, etc.).
- Medical records generally must be kept very long time.
- Documentation of toxic substances, etc. which also must be kept forever.

Reporting in the classical sense and prediction of customer behavior (such objective analysis generally) justify the investment in data warehouse. Prediction of the behavior of client that production system can not efficiently provide.

Predicting behavior of customers will depend on the categorization and on two important parameters from market: the interaction between consumer and product selection coincidence. Looking at these two parameters can be analyzed 4 combinations or 4 types of markets that are really environments for business companies:

Linear market (interaction of consumers and product selection coincidence are low). An example of waste enterprise in which most customers do not care too much for the service (and indeed do not think of company if it performs as expected (if such waste is collected regularly). On the other hand the number of users is also subject to low fluctuation. Fluctuation, customers churn is almost neglectable because population is relatively constant in analyzed area. Predicting customer behavior is very accurate and possible, because last year's behavior can tell a lot about the behavior of consumers in current year.

Statistical market (small interaction of consumers and HIGH accidental selection of products). Sales of household goods is a good example of this. Although some consumers know who will probably buy detergent at the same time are not so sure if they need windshield cleaner. Time of the year still affect whether customer will be better to sell air conditioning or snow chains for cars. Correlation of different sizes in data stock is definitely a useful and new information about consumer behavior. Especially if we have a well developed attributes of the buyer (which is usually not the case because most sellers may know that the buyer is a man of 42 years but whether the household has 3 or 7 members, it has been something else which definitely defines its purchase)

Chaotic markets (HIGH interaction of consumers and small product selection a coincidence) A good example is the high fashion market where the taste is formed on the basis of interaction between consumers and what was at one point "in". Data warehouse in such conditions do not provide any new information.

Real-time market (HIGH-consumer interactions and randomness LARGE selection of products) stock financial markets is a good example of real-time market. By itself it is clear that, unfortunately, data warehouse will provide answers to questions where to invest in the past years, although dwh has great historic data and data about customers and everything. The market is still complicated as



chaotic market, the number of parameters that influence the purchase are large, crucial parameters are often secret.

The trend that CRM and Internet imposes is the quality of communications between manufacturer and a buyer. Internet is increasing interaction of consumers (see for example the reader reviews of books on www.amazon.com or hundreds of sites that deal with high end electronics <http://www.hifiworld.co.uk/>). This trend contributes to the growing interaction between consumers (growing and chaotic real-time market) which is one of the factors why today's data warehouse terms are less effective in their role as predictive of behavior in the future. Terabytes of data will remain just terabyte of data for the named markets.

The advantage and benefits of data warehouse over other forms of information reporting system is in:

- Merging diverse data from multiple sources (multiple production systems) implemented on different platforms.
- Rapid detection of the changes in the source system.
- Iterative character of the model building data warehouses and hence the iterative nature of building software for extraction.
- Detection of errors in a production system
- Lasting data storage (typically 5-10 years) in relation to production systems (typically 1-2 years)

Data gathering

Data gathering is an important feature of the data warehouse. The information system of companies in many cases, is comprised of multiple subsystems, physically separated and built on different platforms. Such non-integrated information system is a major problem for a system of reporting within the company. The problems of timely collection of all necessary data, inconsistencies among the reports obtained from various sources covering the same area of business within the company, are reporting poor. Data Warehouse performs integration of all existing data sources and makes them accessible in one place. That process of collecting and integrating data from all available sources, the most difficult task in building a data warehouse.

Each component of the production information system is a potential source of data for data warehouse. Only the data warehouse does not allow a direct, manual entry of data into it. Manual entry of individual records in a data warehouse is not permitted nor necessary because the data already entered the production information system company (this is a basic purpose of the production system). On the other hand, input data into data warehouse is done automatically, periodically and in large quantities. For example, it may be decided that the end of each working day to perform data collection and aggregation of data from any available source, production systems and to transfer them in a data warehouse. This job will perform software system that must be built and run in fixed time intervals. The time interval between transmission of data (data warehouse refresh period) may be one day, week or month for example, depending on how up to date data is needed. In the time between two refreshes, the base data warehouse is quiet, ie the base is not for any input data but only read from the database of data warehouse. Of course, here we face the fact that the warehouse data always have old data from yesterday or last week or month. This may seem



like a disadvantage, however, the purpose of data warehousing is such that it does not require the condition as it was in this moment. Data Warehouse uses the administrative structure of the company that sets the types of questions: “How much I earned in the last month when it comes to foreign business partners of the realization of what is charged in the same period last year,” or “What is the most problematic categories of users in terms of returning the loan and how much the average delay in the case of married, male with more than two children? “. Daily data warehouse refresh period is quite sufficient for your first question when the monthly period more than good for the other issue that takes into account the historical data that can sezi up to ten years. One day delay can make a significant difference. Further, as between two refresh does not perform any input into the database storage of data, statements made during this time certainly would be consistent, which may not be the case with statements from the production system in which the data fluctuate due to the continuous input. Of course, if we ask the question: “What is the state party’s bank account?” then we will use the data warehouse for production information system that shows the way this is just a minute or seconds.

Prerequisites for building the system for data transmission

The detailed elaboration of the process data can be set off only after the following conditions are met:

- 1st Defined as (initial) requirements of users in terms of necessary data.
- 2nd Are available to people well enough to know the structure and content of the source system.
- 3rd The source system is not capable of the construction or modification of a logical structure.

The first requirement means that there is at least a sketch data model, ie a list of measures and dimensions that the user wants the data warehouse database. Without fulfilling this requirement is, obviously, can not go into the design phase retrieval.

The second requirement is extremely important. Because of the complexity of the sources, it is very difficult to find contact information, to determine the exact algorithm to obtain information without a good connoisseur of the source system. The importance of this requirement has proved to be of practical project to build a data warehouse of the Zagreb Fair, when the source familiar with the system were not always available team to build warehouses. Because it is often ignored some important facts related to the complex structure and content of source data, a result was inaccurate retrieval algorithm. This situation is repeated several times in a project to build a warehouse, resulting in constantly changing, but mostly written in code, and thus necessarily delay the project.

The reasons for setting up the third terms are obvious. If the structure of the source system is not stable, the algorithm retrieval may be subject to frequent change. Time limits can be incorporated, which leads to customer dissatisfaction, etc.

Data warehouse in operative daily work

Working with data warehouse can be viewed as two separate parts. One is the automated process of daily data and the other interactive work users with applications where the data source data warehouse.



Daily data recovery

Data Warehouse has a certain amount of time in which the data is refreshed. Typically, the data warehouse is updated once a day, in order to relieve hardware resources to be carried out mostly at night and not disturbing the normal operation. Refreshing the data is completely automated and requires no action by the people. In case of any mistakes in the production system was in the process of refreshing the relevant people (developers and builders warehouse administrators), it is automatically noticed.

Working with tools for viewing data warehouse

Tools for interactive viewing of data warehouses (which has already been done, therefore they should be differentiated from the program for building a data warehouse) can be already finished products such as Oracle Discoverer or user applications such as dairy dates in Oracle Application Express. These programs are adapted to work with data warehouses and are intended as decision support for administration, etc. They differ from Oracle Discoverer, which are generally adapted to the company for koj statements were made and what they need. The results of these programs get basically all the summary level and do not deal with individual records (for example, are typical of such reports by region rather than statements by customers “from first to last, such reports as the user must wait for the (quality) production system and not from the data warehouse). Very generalized, saying the production system should have the better answer to the analytical data warehouse to the question of synthetic character.

Interaction with Office package (Word, Excel)

Because the distribution of MSOffice package here will make a brief reference to these programs. Data warehouse and applications that are based on it are the result of the efforts that all information relevant to the functioning of businesses unite in one, for (of speaking) of each access point. If the user data on its sale of holding in Excel it is perhaps easy, but they can not be integrated with other information such as sector (ie, least not without the intervention of each of these users manually re-enter, copy + paste određeneih parts which decides integrated, ie Popćeno action dependent on customer service this spring). So the summary report for a company always at the request must work again, which is the objective circumstances (the presence of all persons who are required to do so, updating of data, etc.) means that the report came much later (perhaps too late), but if the data on one place (in the data warehouse) came to their synthesis is concerned computer. Excel as the application can be helpful in the extreme view of some data and prepare them for eventual printing and presentation rather than as a tool to interactively look at these data because it was not intended.



Maintenance of Data Warehouse

Relatively little attention is paid in the literature on the maintenance of data warehouses. This business segment is not only irrelevant in the life cycle of any IT system but the Warehouse perhaps more pronounced than in other applications. Article will present some of the problems in the 'life' of data warehouses. Dealing with problems depends primarily on the environment in which the warehouse is in general about and the possibility to actually make anything.

Knowledge Takeover

When submitting the product for maintenance it is necessary to have in mind the complexity of data warehouses. Although most procedures and codes are essentially the same thing, it is important and necessary **to understand the issues and learn about the production systems** from which data are retrieved for the data warehouse. Person responsible for maintenance will have to learn a lot about all the systems where data are retrieved. One thing should be kept on mind, these production systems are living systems and are subject to change for various reasons, and that they also have their own lifecycle of changes, smaller or bigger. Any changes to the source system can cause error getting data from the source. It should be considered that the standard data warehouse is full of different applications and that the person who maintains the data warehouse should be at least generally familiar with business and process logic of these production systems and is able to evaluate how changes in the source applications, cause changes in the way of the reach and quality of data for data warehousing. A person who deals with maintenance of the warehouse will be very exposed to changes, much more than someone who deals with the maintenance of typical transactional applications. Monitoring possible changes in source systems and responding to them is more complicated than responding to business issues that come into ordinary transactional applications for several reasons:

- Lack of familiarity with the work of transactional sources
- Not being informed about changes
- The complexity of the code change
- Too little time or opportunity to test

For these reasons it is recommended that the DWH developers and administrators working on data warehouse maintenance to be present at all meetings concerning changes in the sources for the data warehouse.

Deleting Data

Next important issue of data warehouse maintenance is proper periodical data delete exercise. At a certain moment, DWH will certainly need to delete part of the information. Prudent is not deleted, but consider the possibilities of alternative methods of storing such data. One possibility is in keeping all data in storage system, but the greater the volume of data is - worse is performance, response to inquiries. Although, there are ways to circumvent the problem. Another option is to pull old data from tables and storage in case of possible subsequent calculations. Comparisons of actual archival data is often difficult to make a quality manner, because the business concept has changed, completed the restructuring of income, expenses, etc. If there are no quality transformation tables, comparison will be possible only on the basis of estimations. The process of restoring the old data is often time consuming and restoration of the old data may require several days with a number of employees engaged to create the desired report. Certain databases allow partition table on n logical unit. Some of these units can be virtually excluded in order to shorten response time to inquiries, if necessary parts included. It is therefore not necessary to load table with data and to use parts of the table, which is of course much more elegant solution.

The third way is to eject parts from a warehouse of information in aggregated form. It is now the most common way



to delete data. The decision to retain the aggregated information is based upon response of the most business users of data and in collaboration with the people who maintain the data warehouse.

Creating Reports in Data Warehouse Maintenance Mode

The fact is that some users, because they are more agile than others, may be able to make themselves some reports. Even though, these users can become so skilful in knowing reports and business logic with the reporting tool that they can eventually be better than the staff responsible for maintenance of Data warehouse system. On the other hand it is expected that after the introduction of DWH and by going in operation IT will no longer use the reports and they will no longer have to write reports. But, there will be always reports that users will not know to create or to do them without errors.

It is also often the case, that the appetite after the introduction of DWH systems significantly increases and that the initial process of creating reports and publishing of the results to customers or other business segment. Because users including management can get the wrong impression that it is sometimes very easy to create a report by the maintainers just as it is simple to activate it. It is of course wrong image, right picture would be for proper interpretation of DWH efforts = invested great DWH effort by staff enable ease of use of data warehouses for customers. Greater IT effort means better simplicity of reports for users. In reality efforts are just reciprocal.

Unnecessary Information

It is expected over a time that DWH becomes a bag for any information or data that is found. It is required to be brought into the warehouse 'and such as you find' data sources for a sense of completeness of the staff that maintains DWH systems. DWH system thus becomes a bag in which to throw all sorts of information, many of which will be never used, and which will be further burden for the system. Unnecessary data can be split into those where the warehouse is not the place for them and to those who are ineligible for entry into the model. As an illustration, certain users requested information from the newspapers should be found in the data warehouse...

Endless Possibilities and Requirements for 'Tuning' DWH System

There are users who are sensitive to the work of DWH system just as well as production users who can not tolerate such as more than half a minute unavailability of the production system. The reason for poor response of DWH may be certain ad hoc queries. DWH system is more vulnerable to performance slowdown than a production system that is often resistant to such things, secured and such queries rarely appear on production due to nature of work. One "poor" ad hoc query can block DWH system to the extent that for few hours is much slower response. Normal DWH users who regularly use this system will be unhappy with performance decrease.

It's often impossible to predict from designers point of view all combinations what users can do to DWH performance, and it can be seen only through practice. Active education of users is therefore very important to avoid customer dissatisfaction and in favor to IT staff to which this dissatisfaction



are reflected.

Aggregation of Data

Many DWH systems imply the need for the construction of aggregated structures that contain less data and thus have a better response to ad hoc inquiries. These structures can be technically implemented in various ways such as tables in the database, aggregated in OLAP systems, or something else. Common to all these is that data warehouse maintenance experts must spend time for building them in useful way on regular basis and without many errors ('up to date'). The growth of aggregated reports can complicate maintenance of DWH. To avoid too complicated maintenance it is required to somewhere draw a line and to say that certain statements are no longer worthwhile to build, and it is sometimes necessary to wait a longer time to respond in a query to avoid customer dissatisfaction.

Reports from the DWH system Vs. Reports from the Production System

Erroneous is the idea that the OLAP tools can completely replace the reports of the production. According to the experience, aggregated reports should be made with the data from the DWH system (statements produced from global experience) and reports that provide a single image should be provided from the production systems.

Examples of reports / views	Source	Reason
Members depth	Production systems	Relatively fast response
Total depth of members group, Comparison of member accounts for group A	DWH	Customization for aggregation

Data Corrections

Once noticed the error should be corrected if the ratio of information importance and resources availability is positive. Filters and data-cleansing process will observe errors from the production system. Errors should be corrected sooner or later in a production system, and communication about error correction is important whether they can be corrected. Rectification can be done in the software procedures which data are retrieved after they are observed. Typical errors can be made such as wrong date (eg, 4.3.0002). DWH systems are "read – only 'systems, in comparison to production systems. DWH can with little effort create a specific table of recorded errors, table as a writable location and description of the observed error, and this is important information for cleaning data in production system. With such table of error information production system can accept and act on the basis of them.

Selecting Tools

DWH tools often overlap in their functionalities. Sometimes, company is investing considerable money in OLAP and multidimensional database. This investment might look irrational if the relational model was satisfactory with common reporting tools for most tasks. But, company should be aware that with the relational database, old star schemes and together and with various reporting tools, often come to the insoluble problems, even though they created barely 5% of all reports. For this reason relational databases are replaced and investment is made into multidimensional databases and OLAP. The importance of unsolvable problems is a measure and trigger for the introduction of DWH system.



The fact is, however, that investments in complex OLAP and other reporting tools needs to be somehow justified and that is not always easy since this few percent of reports sometimes can be made with extra expert work and without DWH tools.

Testing Structural Changes In the Source Systems

Very soon one of the sources in the life of DWH system will change. Testing the transfer of the modified information from production system might not be easy at all. The implication of problems are described in a few steps:

- Temporary suspension of data transfer
- Developer intervention on procedures for data retrieval
- Testing the data transfer - accuracy and quality
- Inclusion of the modified data transfer in the corresponding extraction

The third point can be sometimes very problematic, particularly if implemented data warehouses do not have testing environments. Testing DWH can be sometimes build as smaller replica of productive environment (partly with pieces of data, etc.), but usually testing requires an identical copy of the productive DWH and for such requirement there are usually no resources.

Consequences of Problems from Source Systems

If the report is made based on the data from several source applications, it is very likely that most of the report will still be correct despite the possibility that some reporting parts contain errors! If there is a dimension introduced that contains a list of sources and associated attributes, example of table can be as presented:

Source	Date of last extraction	Status on source
System A	15.4.2002	OK
System B	15.4.2002	OK
System C	16.4.2002	Sum – OK
System A	16.4.2002	Error – transfer did not pass
...

... ..

In this case, it can be said that the report was good until 15.4.

There can be a display format which immediately targets error as presented in the following table:

Source	Indicators	Amount
OK	Number of users	200,000
OK	Number of new users in month	17,350
ERR	Total traffic of new users	145,703
...		

...

where the error are expressed in bold font.

Problems in Changing the Attributes and Their Behavior



For example, if the code of products are marked for years with the prefix code 100 – xxxxx and suddenly from a particular reason there is a need to start products with the 120 - xxxxxx. Question is, shall all changes be appropriately implemented and performed in a manner that final reports will not show mix of products. Despite the possibility that this is resolved with slowly changing dimension, it is possible that some users of the tool just took the code as a criterion for the report. They will have errors in reports. All variations of errors can not be predicted when attributes are changed.

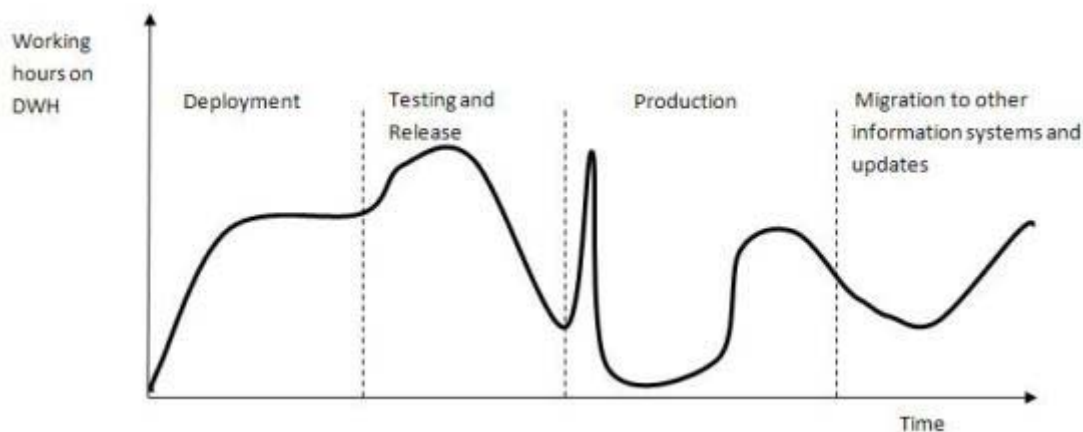
Access Rights

The fact is that employees within the company change jobs, organization units and responsibilities. New employees are hired and others are leaving. There is a fluctuation of DWH users. If for example there are hundreds of DWH users there is a great opportunity DWH system administrators will be addressed almost every day with access requirements: who must see what, who should be restricted from access or why XY user sees what should not be seen. The solution may be in a maximum granting rights to specific users – for example, managers who have the right to decide who you can watch that attributed to themselves the same rights to other users – with their subordinates.

“Dying” of certain DWH parts

Some parts of the DWH system – data marts in time are not used, and simply there is no benefit of them despite the fact that these parts were presented as great value during the development phase. A number of reasons that were previously described cause the death of some datamarts. Keeping the data updated, unfortunately, due to time and memory limits can become impossible. Then there can be raised certain negative questions – why was it done at all? Because of this, it is a good approach to record and document all project requirements and to keep it for a long time.

At the end it is easy to notice that it is easier and cheaper to build a data warehouse, than to maintain it for a longer period of time. The cost of maintenance should be included in the price calculation of costs and profitability of DWH system, otherwise the DWH system look like a bag without a bottom. Typical life looks something like the image:



From the figure it is evident that after a release and maintenance can consume resources that were not intended for maintenance purpose.



Genuine solutions brought from experience. Guess what? They work!

