Analysis of the need of Data Warehouse Creation

Snezana Savoska[1],
[1] Faculty of Administration and Management of Information systems, University "St.Kliment Ohridski" Bitola, Partizanska bb, 7000 Bitola, Macedonia

Abstract. Data Warehouse is a special tehnological environment that integrates data from internal transactional company databases, external sources of data in company's repozitorium which is convinient for the data analysis, periodical and ad-hoc reports and decision making. For the strategic decission, making purpose and solving unstructured problems, we use the agregated data from the data warehouse and data marts. Data warehouse becomes the central point of the comapny's demands for information and visual data reprezentation. The concept of Data Warehousing is especially provided with new technology of relational databases with parelel processing.

Keywords: Data Warehouse, Data marts, decision making.

1 Introduction

Data Warehouse is a special technological environment that integrates data in the form convenient for analysis, periodical and ad-hoc reports. Also it makes it easy to process the reporting of managers, analytical staff and decision making process. For the strategic decision making purpose and solving unstructured problems, usually the transactional information systems aren't being used in the company's every day operating systems, but we need to use the aggregated data for the given period of time, collected in the Data Warehouses. The expansion of concept of Data Warehousing is especially provided with the technology of relational databases parallel processing model.

In fact, the Data Warehouse is an adapted reproduction of data of the transactional information systems, especially structured for queries, analysis and reporting. The transactional information data and Data Warehouse data are different entities. The data in the Data Warehouse is updated periodically. Data Warehouses are huge repositories that grow up in enormous range and have a specific structure. Although they come from transactional information systems, are transformed in a specific format suitable for reporting and data visualization. The data is taken from different sources, sometimes from different servers from different operating systems. Data can be loaded in the data warehouse using specific procedures and some transformations in a specific data structure.

They are convenient for periodical, chronological and ad-hoc reports and they are user-friendly for managers and analytical staff for the companies. For this purpose,

we often create additional aggregated tables which are capable of giving prompted information and acceptable data visualization.

The Data Warehouses become the central repositories of all organizational needs for information used for new relationship research, trends and hidden values. They are a common focus point of all organizational members; they enhance business knowledge, very important for the current knowledge era.

The saving data format is different from this transactional information data format. The Data Warehouse is not usually normalized; it can include relational databases, multidimensional databases, flat files, hierarchical databases and object databases. Big part of Data Warehouses is used for post-decision monitoring of effects of decision making called operation analysis. Also, the market pressure, the software vendors and industrial experts make a very strong influence for Data Warehouse and modern data mining tools affirmation (Figure 1). This affirmation is in the direction of development of the company and sectors data warehouses in all forms, using specific data mining tools and enabling data visualization in the decision activities [5].

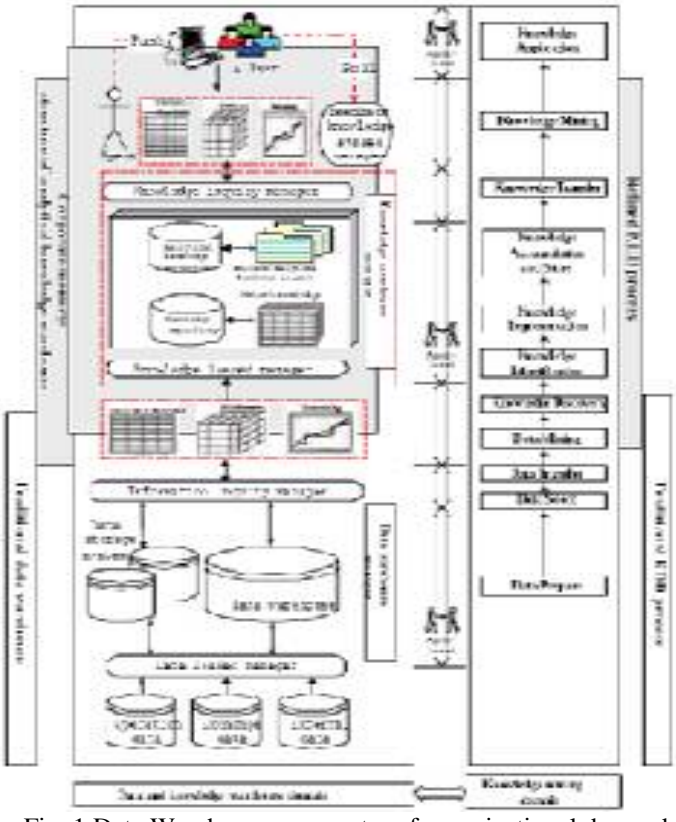

Fig. 1 Data Warehouse as a center of organizational demand [5]

2 The role of Data Warehouse in the company

The Data Warehouse is the target of the data acquisition as a source of a company's data delivery. Wholly looked, Data Warehouses subject it's use, to some primary functions that are in fact its tasks

- The data warehouse directly supports the different business rules in the company and in this way, supports managers and analytical staff in the process of management and decision making. The data warehouse properties must be flexible and easily adaptive to the business rule changes. It includes adding new entities; hierarchical relations dislocate and change the static entities relationships.

- Data Warehouse is a collection of rules for integral, objective and subjective strategic information which must be managed in all phases, starting from data acquisition process. The necessary characteristics that the data warehouse must own imply data modeling and usage of design techniques that support subjective orientation. It also provides data integration flexibility through additional data sources in all data live cycle.

The data warehouse is a historical store of strategic information with historical data relations with different entities. This property enables data modeling and easy usage of supporting design techniques for nonvolatile and time consistence. It is the data source from which the data is exported in the sector's Data marts. The sector's Data marts can be used for data exploration, data mining, and manager's reports or as a base for gaining multidimensional cubes for analytical data processing called OLAP cubes. This implied using a model of unbiased data which will be filtered and prepared to satisfy specific aims and demands of data marts. Some models for data aggregating support are being used and also aggregated tables are created because of the minimizing the response time and increasing the end users satisfaction.

Data warehouse is a source of stable and nonvolatile data, from the aspect of data processing and data changing. The model for this type of data must be different from the transactional data model [5], because when we load data in the warehouse, they become historical data and they can't be updated or deleted. They must be browsed or selected.

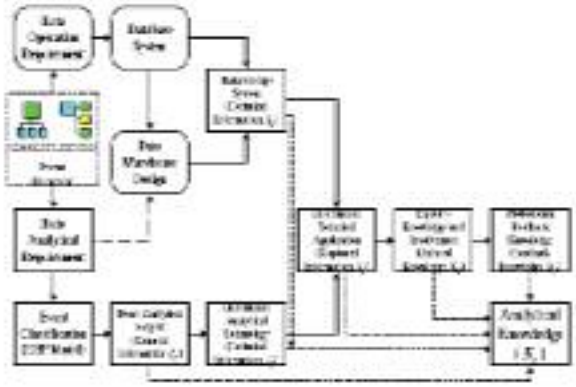

Fig.2 Relationship of Data Warehouse with other company's IT systems [5]

3 Data warehousing and Data marts

The concept of data warehousing is a new one. For this reason, we need to standardize the definitions and the core issue, because in the literature we can meet different ways and definitions related with this issue. For this purpose, we usually use the company's data dictionary. Some issues are focused on data and other on people, software, tools and the business products. W.H.Inmon created the clear definition of the concept of data warehouse in the direction of measured attributes which clarify our understanding: "Data Warehouse is a collection of integrated, subject-oriented databases, designed to support decision support systems, which data is in volatile and relevant at the same time" (1992 Inmon).
Inmon's data warehouse definition make two implicit assignment:
(1) Data Warehouse is a physical place separated from the operating data
(2) Data Warehouse contains aggregated data and atomic analytical data for management, separated from transactional processing systems
Demands of separated data environment for data warehouse are an essential element of the concept. In many cases, the operating processing system is inconvenient for many business processes such as decision making and data analysis. This is obvious especially in the process of data visualization where data preparation time is required. In this case, data must be constant, integrated and time assigned.
First of all, we must explore some associated elements: How data will be stored in the company's data warehouse, how to make data marts and the data warehouse's metadata. The sources of data for the data warehouse - the transactional databases, are subject oriented, volatile and customer and product oriented, or focused on the operator' demands. Data transformation and integration in the data warehouse's counterpart or transformations of data events instances are not suitable for loading in the data warehouse. Transactional data has to be integrated with other unrelated transactional data, sets of internal or external company's data. There must be some procedures for data distillation and integration for the given flow of "loading" data in the organizational data warehouse. These procedures are created as scripts or triggers and run automatically in specified time, creating an updated data warehouse – refreshing data from data sources valuable for data warehouse.

Although the central aggregation concept, there is some kind of specific aggregated data which can be presented in same business organizational departments but is tangential to other departments. Alternative concept is adopting of scalable, less expensive version of data warehouse, called data mart. The data mart concept is sometimes called mini-mart and is posed to minimize the expenses and maximize the usefulness for organizational business departments. From the company's perspective, data mart can be an isolated island of information, accessible for top managers and the owner - organizational department which use this data mart. But, it isn't accessible for the other company staff. Because of these reasons, first it is convenient, to create data warehouse for the entire company with a complete view of the company and then organizational department's data marts. Another possibility is first to create the organizational department's data marts and then to integrate them, they create central integrated company data warehouse, depending of the manager's demands. This is more targeted and less expensive method of gaining DW from the existing data marts

The metadata is an important data warehouse characteristic; they are the information for data in the data warehouse. Each system must save the data attributes and the data transformation algorithm for the data warehouse, from where data is created, place where they are located, how are they transformed and how can be accessed. In this way, we obtain the possibility to crate a personal warehouse to meet specific customer needs. In fact, data warehouse is subjectively oriented, data integrated, time dependant and in volatile [1].

Creation of applicative data design of the data warehouse is always done with the defined variable names, the screen's variables and other issues are placed in the data dictionary of the data warehouse. All changes in the names and variables lead to inconsistency of applications with data dictionary and they must be automatically updated and changed in the organizational manual, but this is allowed until the moment when the data is loading in the data warehouse. In that moment, the discussion of which convention will be accepted and which names will be used, is finished. All the accepted conventions are saved in the metadata dictionary.

Another result of the data integration is retrieved of the common unit measures for all of the synonyms for the data elements in the data warehouse tables. Common unit measures integrate different measuring attributes and also create global acceptable units for all of the final data in data warehouse. For these reasons, there must be an organizational consensus for all of the data measures and all of the data units must be converted in these units when they are loading in the data warehouse. This agreement is available with creating a standard that will be accepted by the organizational staff and managers and will be used via all data warehouse's users. Sometimes the unit's unification must be obstruct and limited with a wide range of the company member's demands and demands for different data query with various unit measures. In this case, it is necessary to convert the data in a non-standardized unit measures and permitting the flexible staff's conversation setup. In the data visualization process, the retrieved of unified measures play an important role because of the fact that the unified measures of data give a valid visual data representation, and opposite, the visualization process is meaningless. Without a doubt, the data must be stored in data warehouses in an integral, global acceptable way defined with business rules. Figure 3 shows the organizational data flow for data warehouse.

Time period included in the data warehouses contains data from many years, opposite of the transactional information systems which store the data in short period of time, maximum - one year. Each primary key contains an explicit or implicit time elements (day, month, quartile, and year). In the presentation, time unit must be included with the primary key. When the transactional data is transformed, some implicit data element must be added and loaded at the end of the period and this time period must be exact and constant. Exception is made just when the data is incorrect or poorly transformed. Inhibit changing and updating data in data warehouse have sense because of its in volatile nature. These operations are allowed in the transactional databases, but in data warehouse, only the operations of loading and data browsing are supported.

Another difference between transactional databases and data warehouses is the normalized form and minimized redundancy of transactional data, opposite of demoralized form of the data warehouse. Data in the transactional databases are usually in the third normal form, with eliminated, derived data elements and

aggregated data. Designers of Data warehouses don't care about redundancy because redundancy is recommended and it provides a rapid system response for reports, query and data visualization[1].

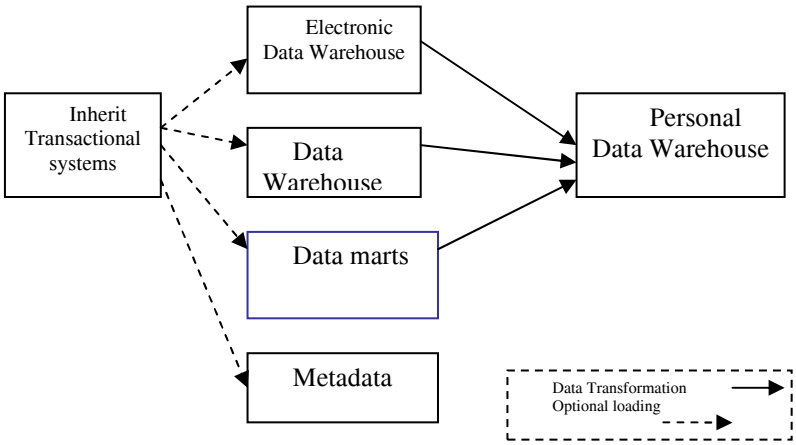

Fig. 3 Organizational data flow for Data Warehouse [1]

4 The Data Warehouse's Architecture

The data warehouse's architecture must support the organizational working data structure, communications and data processing. Figure 4 shows the relationships between data the warehouse's elements. The source level is presented by transactional database systems and external databases. The aim of the data warehouse is to release the data, locked in transactional information systems and to mix it with those from external databases. Additional aim of the data warehouse is to cause minimum influence to the system's operations to the data for analysis and processes monitoring. Web and internet technology allow easy and more economic access to data and incorporate the data in company data warehouse.

The metadata level contains data about data. When we insure universal access to data, some forms of directories with data and some data repository must be managed. Metadata includes a directory where data is stored with their names and mining, rules of summarization and data cleansing. In this directory some facts are provided, how data was prepared and how was stored in the data warehouse, from where data sources are taken and how they are saved.

The process management level is focused on planning tasks for creating and managing the data warehouse over data directories. For these reasons, this level leads us to create updates and managing procedures for data warehousing. Tasks as periodical downloading data from identified data sources, data aggregation and

[1] Inmon in 1992 show this and he was faced with the first impression was that there is minimal redundancy between transactional databases which are the sources of data and the data warehouse (around 1%). He enumerates couple of factors for this

downloading external data, as the metadata updates are allowed in this data warehouse architecture level.

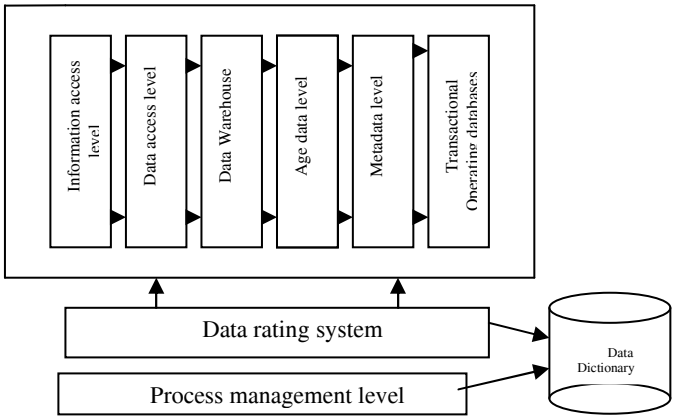

Fig. 4 Components of Data Warehouse Architecture [2]

Data access level transfers all of the information across the network. It is called middleware and includes network protocols and a routine of searching. Applicative sending messages can be used for some isolated applications. This level is also used to collect transactions or messages and deliver them to some locations at the same time. This level can be imagining as data warehouse message transport system.

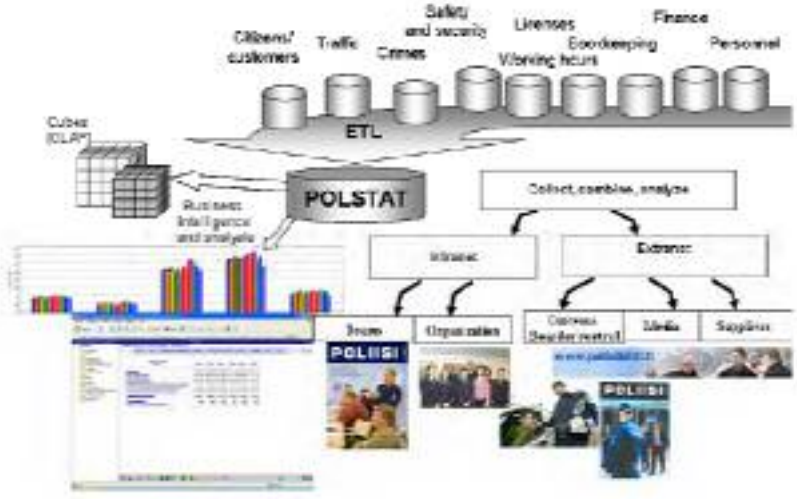

Fig.5 All DW levels must be adapted to business rules [5]

Physical data warehouse level is the place of the actual useful data locations in the data warehouse. The final component of the data warehouse architecture is the level of data saving which includes processes as selection, editing, summarizing,

combining and data loading. Also, they include information about internal transactional company databases and external databases. The physical data level often includes complex programming. But, the data warehouse vendors reduce the process's complicity, making the tools easy to use. The demands of this level are analysis of data quality, filtering, model identifying and structuring data in the resent operational databases (Figure 5).

Another model to planning and development of the data warehouse architecture is a system that consists from six subsystems:

1. Operating data from transactional information systems which provide company's data for loading data in the data warehouse over application for extraction, data propagation, conditioning and data standardizing.
2. Data migration and collecting in data warehouse. In this case some form of synchronic and asynchrony conversion of data can be used[2]. Also, we can mention operations such as data refreshing, updating and propagation. In this phase, the processes of data restructuring as renormalization, adding new files, entities and keys, data combining and aggregating are also allow data to become information. There is data from internal and external sources, prediction data, assessments and simulations. The quality of this system is depends of the transport mechanism in the data warehouse such middleware.
3. Data warehouse designing and administration understands great knowledge for DW metadata, the relationships between them and their sources (internal or external). Also the designer must have information about the data relationships, their granularity, and period of updating, backup period. As the data warehouse becomes bigger, the more expensive it is to maintain it. Although the higher dimension of the data warehouse, the simplification of data is not recommended because of the essentiality of analytic data analysis and decision making. Today's data warehouse administration system includes usage of visual HCI interface which simplify data administration, data cleansing, data transferring and data loading (For example Oracle Data Warehouse builder).
4. The middleware is a system software for machine learning with purpose to collect and use knowledge of the abstraction level and it depends on the number and structure of the users, their connection via LAN or WAN in client-server environment[3].
5. Decision support and analysis applications demand creating database models, databases and application software which provide their connection and usage from the end users. The methodology includes different data access, mathematics and statistic models, visualization, heuristic models or using knowledge databases (or library) and case studies. Sometimes it is very hard to connect the solution complicity with the simplicity of the end-user solutions. For these reasons, such systems include data mining software, data exploration and sometimes data models in knowledge databases. Although all problems complicity, interfaces mustn't be overcrowded and with intensive color. The colors are welcomed for end users but sometimes, fear the user. This means that we can use reasonable screen visualization and understandable user interface.

---

[2] Detail in N.Balaban,J.Ristic, Sistemi za podrzavanje odlucianja, Saraevo, 1998

[3] This concept implied physical separation of applications and databases

6. Presentation interface is the most important system of data warehouse usage because it provides communication with users and courage or discourages them. Depending of its intention, the interfaces can be classified as:
a) Simple information interface in form of tables or visual presentation
b) Interactive systems with queries and drill-down possibility
c) Simulated system performing "what-if" analysis
d) Functionality systems that support corporative function providing company functions
e) Automated expert systems which help solving specific expert problems
Few of them may be combined in an unique interface. We can use interfaces such as: command prompt, menu-interface query language, graphics interface, groupware and multimedia and hypertext interface. The interface has to be easy for use and providing easy data access and data understanding for end-users.

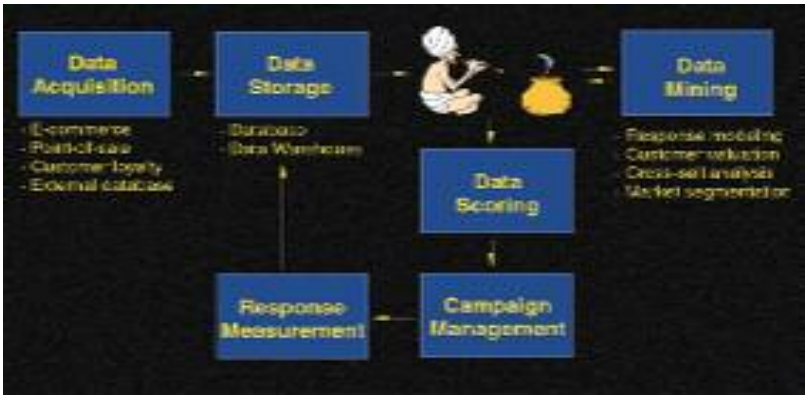

Fig. 6 Relationship between Data Warehouse and Data mining processes [4]

5 Conclusions

Data Warehousing is a preferable concept for all manager's and analytical demands because of the growing needs for data analysis and demanding information for decision support. For this reason, the data warehouse becomes focused on all organizational and informational needs. But, to build and maintain an organizational data warehouse it is not an easy process. There are some conditions to be satisfied, such as the company consensus for building warehouse, support from top managers and proper organizational behavior. Also, there must be support from experienced supporters for technical and software processes in the creating of the data warehouse. The most important thing is defining the business rules, granulation, metadata, data sources and model of data transformation. For these reasons, the data warehouse must be planned and created very carefully. We often think about data warehouse as a part

of data mining processes in the company. The representation of the relationship of the Data Warehouse with data mining is shown on Figure 6.

References

1. Marakas D., “Modern Data Warehouse, Mining and Visualization”, Chapter 2 and 3, (2005),
2. Balaban N, Ristic J., “Sistemi za podrzavanje odlucivanja”, Saraevo, 1998
3. Inmon, W, H, “Building the Data Warehouse”, Third edition, WCP, 2007
4. Turban E., “Decision Support Systems and Intelligent Systems”, 2008
5. Wang J., Data Warehousing and mining, Second edition, USA, 2008