Medical Informatics Europe '99
P. Kokol et al. (Eds.)
IOS Press, 1999

Outpatient Health Care Statistics Data Warehouse - Implementation

Dejan Zilli
Vizija d.o.o., Ljubljanska 42, Celje, Slovenia

Abstract. Data warehouse implementation is assumed to be a very knowledge-demanding, expensive and long-lasting process. As such it requires senior management sponsorship, involvement of experts, a big budget and probably years of development time. Presented Outpatient Health Care Statistics Data Warehouse implementation research provides ample evidence against the infallibility of the above statements. New, inexpensive, but powerful technology, which provides outstanding platform for On-Line Analytical Processing (OLAP), has emerged recently. Presumably, it will be the basis for the estimated future growth of data warehouse market, both in the medical and in other business fields. Methods and tools for building, maintaining and exploiting data warehouses are also briefly discussed in the paper.

Keywords:
Data warehouse, outpatient health care statistics, On-Line Analytical Processing (OLAP), Decision Support Services (DSS).

1. Introduction

In our country there are several outpatient health care information systems that are used to store medical data in health care centers, private practices and other health care institutions. The Institute of Public Health of the Republic of Slovenia has defined several requirements for collecting outpatient health care statistics data on the state level. Application for statistical data distribution, made according to these requirements, served as a basis for the presented data warehouse solution.

The outpatient health care statistics collects the data about diseases and conditions as well as those about attendance and referrals. The research was limited to diseases and conditions, but can be easily used for the whole outpatient health care statistics eventually. The logical design of Outpatient Health Care Statistics Data Warehouse is discussed in the paper by Natek S. [1]. This paper introduces some methods and tools for building, maintaining and exploiting the data warehouse. It also presents some practical experience gathered during the implementation of Outpatient Health Care Statistics Data Warehouse.

The technology and techniques for building a data warehouse are not new to the world of database experts. They are thoroughly discussed by Inmon [2], Kimball [3] and by lots of others [6, 7]. It is not the intention of this paper to discover any new method or technique for building the data warehouse. The purpose of this research is to show how to effectively employ new inexpensive technology to do the work that took a lot of money and time in the past, with a very small budget and in a short period of time.

To achieve this goal the following hypotheses were proposed:

1) The *PC environment* is mature enough to serve as the basis for our data warehouse solution,
2) *Microsoft SQL Server 7.0* [4] with its *Decision Support Services* [5] can provide a powerful and reliable platform for implementation of almost any OLAP system.

One of the most difficult decisions in designing Outpatient Health Care Statistics Data Warehouse was related to the data granularity [2]. The decision to omit the lowest level of detail from the data warehouse [1] had a major influence upon all further implementation activities. The decision about data sources, data extraction and transformation, the design and population of data cubes, and the exploitation of Outpatient Health Care Statistics Data Warehouse are all the subject of this article.

2. Data Sources

The decision about a higher detail data level was largely influenced by the state level standard for distribution of outpatient health care statistics data. By not having adopted this standard, the data source for Outpatient Health Care Statistics Data Warehouse would be the transaction data from various medical information systems. Such a decision would make the data warehouse implementation much more complex and time-consuming. However, benefit of a lower detail data level would be time dimension (as opposed to half a year period dimension in summarized statistical data). The amount of used disk space would increase considerably, and data sparsity [3] would be much higher.

The state level standard regulates all obligatory statistical data transfer. The statistical data has to be submitted in a fixed format textual file. The data in those transfer files could serve as the data source for the statistical data warehouse. The problem of that solution is the fact that data in transfer files could be of low integrity, therefore complex data cleansing procedures should be developed to check the data integrity. Another deficiency is the lack of dimensional data (diagnoses, reporting units, specialties...) in transfer files. For these reasons the data from statistical data distribution application was used as the data source instead.

The data from statistical data distribution application are in Paradox tables and therefore ODBC readable [8]. Another benefit of this decision is that data from those tables are already checked for consistency by the statistical data distribution application. Every record has a flag stating whether it is consistent. The consistency check verifies the existence of any of the twenty possible kinds of mistakes. The statistical data distribution application also stores dimensional data needed for the data warehouse. For all stated reasons the selected data source is the most suitable for our data warehouse solution.

3. Data Extraction and Transformation

The data from On-Line Transaction Processing (OLTP) systems are almost never in the form suitable for the data warehouse. The same goes for the outpatient health care statistics data. The data transfer data model has several differences from the data warehouse model [1]. Therefore the data extraction and transformation is not trivial. This process usually requires complex programming and the use of data import/export and transformation tools. In our case most of the data cleansing and integration was already performed by the distribution application. Data Transformation Services (DTS), a new facility in Microsoft SQL Server 7.0 [9], was used for the data transformation.

DTS provides a graphical environment for describing data transformation process packages. DTS package stores data source description, data flow description, execution sequence, and transformation description. The transformation is performed by SQL sentences, which are supplemented by VB scripting language and the usage of external transformation tools. DTS package execution scheduling enables total data transformation automation. Four DTS packages were made to extract and transform the fact tables data. The usage of SQL sentences was sufficient for the realization of outpatient health care statistics data transformation.

Several problems were brought to attention during extraction of dimensional data:

1. Overlapping periods in reporting period dimension; there are two standard overlapping reporting periods (first half of the year, whole year). One way of solving this problem would be by calculating the fact data for the second half of the year (the difference between the whole year and the first half of the year). Fortunately, the statistics distribution application also supports the user-defined period, which was set to the second half of the year for the purpose of the data warehouse implementation.
2. Various special values representing the 'NULL' value; all those values had to be translated to 'NULL' value.
3. Dimensions inherent in the data transfer data structure (sex, age group, insurance category, and duration of pregnancy). The data for those dimensions had to be entered into dimensional tables. Modification of those dimensional data would be necessary only in the case of data structure changes.
4. Data integration problem; as the result of the data transfer data model, some of the dimensions could not be integrated in the same data cube. Several data cubes were proposed during logical design [1]. The integration of data cubes in the virtual data cube removes all dimensions, which are not present in all data cubes.

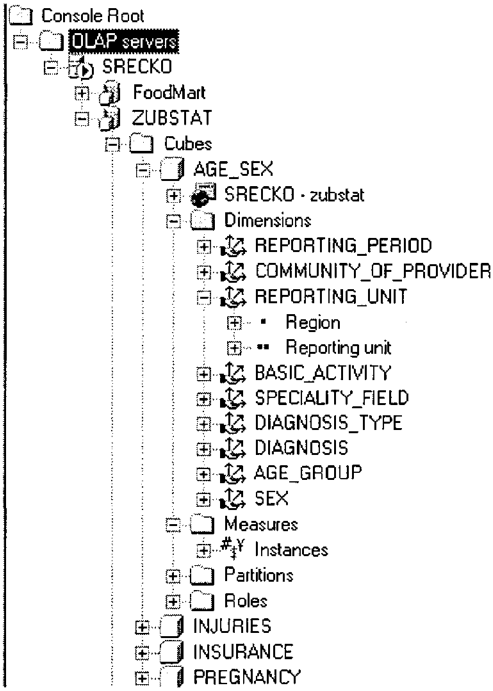
Visual database diagram [9] simplifies the task of changing the data structure. It provides a simple graphical environment for definition of tables, attributes, primary keys, and foreign keys in the SQL Server database.

4. Design and Population of Data Cubes

There are several important decisions that have to be made while designing data cubes [5]:

1. *Storage choice*: DSS offers the possibility to choose between three different storage models: Relational OLAP (ROLAP), Multidimensional OLAP (MOLAP), and Hybrid OLAP (HOLAP). MOLAP model is the most suitable for Outpatient Health Care Statistics Data Warehouse. The data are stored in multidimensional database, which ensures fast response times.
2. *Dimension hierarchy*: some dimensions are hierarchical by their nature. The decision, how to store the data of hierarchical dimensions, was already made during the logical design phase. In the *star schema* the data of all dimension levels are stored in a single table. This solution ensures fast response times for the price of redundancy. *Partial snowflake schema* was adopted in the case of Outpatient Health Care Statistics Data Warehouse, which means that the data of hierarchical dimensions are stored in multiple tables.
3. *Level of preaggregation*: DSS provides the possibility to set the extent of aggregation stored by OLAP server. This can be set with performance gain percent or with maximum storage space. DSS uses heuristics to select and store the most useful aggregations. Full performance gain was set for the sake of this research to achieve the best possible response times.

4. *Virtual data cube design:* the disintegration problem caused by the data source decision can be solved by the means of a virtual data cube. The virtual data cube provides a view over several data cubes, which have shared dimensions. Dimensions that are not common to all data cubes are left out. Measures are summarized over omitted dimensions. In the case of Outpatient Health Care Statistics Data Warehouse virtual data cube can also serve for the verification of the data transformation process (the number of instances of all age groups and sexes must correspond to the number of instances of all insurance categories).



```
Console Root
⊟ OLAP servers
    ⊟ SRECKO
        ⊞ FoodMart
        ⊟ ZUBSTAT
            ⊟ Cubes
                ⊟ AGE_SEX
                    ⊞ SRECKO - zubstat
                    ⊟ Dimensions
                        ⊞ REPORTING_PERIOD
                        ⊞ COMMUNITY_OF_PROVIDER
                        ⊟ REPORTING_UNIT
                            ⊞ • Region
                            ⊞ •• Reporting unit
                        ⊞ BASIC_ACTIVITY
                        ⊞ SPECIALITY_FIELD
                        ⊞ DIAGNOSIS_TYPE
                        ⊞ DIAGNOSIS
                        ⊞ AGE_GROUP
                        ⊞ SEX
                    ⊟ Measures
                        ⊞ Instances
                    ⊞ Partitions
                    ⊞ Roles
                ⊞ INJURIES
                ⊞ INSURANCE
                ⊞ PREGNANCY
```

Picture1: Structure of outpatient health care statistics data cubes, dimensions and measures

After data cubes have been designed it is time to populate them. In DSS this is called processing of data cube. The process is fully automated. It can also be scheduled for the off hours period.

5. Exploitation of Outpatient Health Care Statistics Data Warehouse

Every data warehouse solution, simple or complex, is built to support a number of decision support functions that can not be implemented in the OLTP system. Most frequently used function of OLAP systems is *drill-down analysis* [2]. It gives the possibility to overview summarized data and then drill-down in the direction of interest. The summarized medical data of Outpatient Health Care Statistics Data Warehouse could be used for reporting, analysis of known facts, and discovery of unknown facts (Data Mining). However, it is not the scope of this article to examine all possibilities of Outpatient Health Care Statistics Data Warehouse exploitation.

There are already some client tools available on the market that support Microsoft DSS Server. They use interface called Pivot Table Services (PTS), which provides easy access to the whole DSS functionality. Such open client architecture ensures that many other software providers will follow. Arguably the most important OLAP client tool will become Office 2000 with its spreadsheet application Excel. Relative high availability of those tools should bring data warehousing and OLAP systems closer to the average user.

6. Conclusion

The research shows that broadly available PC based data warehousing technology is becoming mature enough to support the decision-making process. SQL Server 7.0 database functionality and included Decision Support Services provide excellent environment for building and maintaining a data warehouse. It is the increasing availability of client tools that will make the expansion of OLAP technology possible in the near future.

The Outpatient Health Care Statistics Data Warehouse implementation study provides several tips for building a data warehouse with limited resources. The decision about the data source and consequential data warehouse granularity had the greatest influence on the choice of implementation techniques. With the help of this research, the experts from the Ministry of Health have finally been given the possibility to analyze the outpatient health care statistics data. The transition to utilization will bring new challenges and could be the subject of another research.

References

[1] Natek S., Data Warehouse for Outpatient Health Care Statistics - Logical Design, MIE 99, 1999.

[2] Inmon W.H., Building the Data Warehouse, 2.ed., John Wiley & Sons, Inc., USA, 1996.

[3] Kimbal R., The Data Warehouse Toolkit, John Wiley & Sons, Inc., USA, 1996.

[4] Microsoft Corporation, http://www.microsoft.com/sql : Microsoft SQL Server 7.0

[5] Microsoft Corporation, Microsoft SQL Server 7.0 Decision Support Services: Lowering the Cost of Business Intelligence, USA, 1998.

[6] Singh H., Data Warehousing Concepts, Technologies, Implementations, and Management, Prentice Hall, USA, 1998.

[7] Anahory S., Murray D., Data Warehousing in the Real World, Addison Wesley, USA, 1997.

[8] Lazar D., Microsoft Strategy for Universal Data Access, USA, 1997.

[9] Microsoft Corporation, The Microsoft Data Warehousing Strategy: A platform for improved decision-making through easier data access and analysis, USA, 1998.