The Design and the Implementation of an HEALTH CARE STATISTICS DATA WAREHOUSE

Dr. Sreèko Natek, assistant professor, Nova Vizija, srecko@vizija.si

ABSTRACT

Health Care Statistics on a state level is a central point where all relevant statistic data are collected from various sources from all over the country. Various and complex requirements for processing and reporting data makes Health Care Statistics on a state level a perfect example for efficient implementing of Data warehouse technology. The research investigates logical design and implementation of data warehouse with a special attention on a different data modeling technique in various phases. The research clearly shows that a requirement for processing and reporting statistical data determines the basic design decision and thus the basic scope and semantic value of final data warehouse.

Keywords: Data Warehousing, Health Care Information System, Decision Support Solutions, On-line Analytical Processing (OLAP)

INTRODUCTION

Outpatient Health Care Statistics on a state level depends on data coming from health care centers, physicians and institutions on a primary health care level in Slovenia. To ensure statistical integrity of data from different health care software, several requirements were defined by from Institute of Public Health of the Republic of Slovenia, including the data structure for reporting Disease and Conditions as well as Attendance and Referrals for Outpatient Health Care. To obtain cost effective solution for Outpatient Health Care Statistics on a state level, the following technology was employed: floppy disks for data transfer, Paradox for data store and Delphi application is used to merge data from different providers and for basic statistical and reporting purposes. The project obtained some excellent results. To achieve a perfect system, the major objective of our research was to establish whether the employment of the latest data warehouse and OLAP technology design techniques would create the better solution.

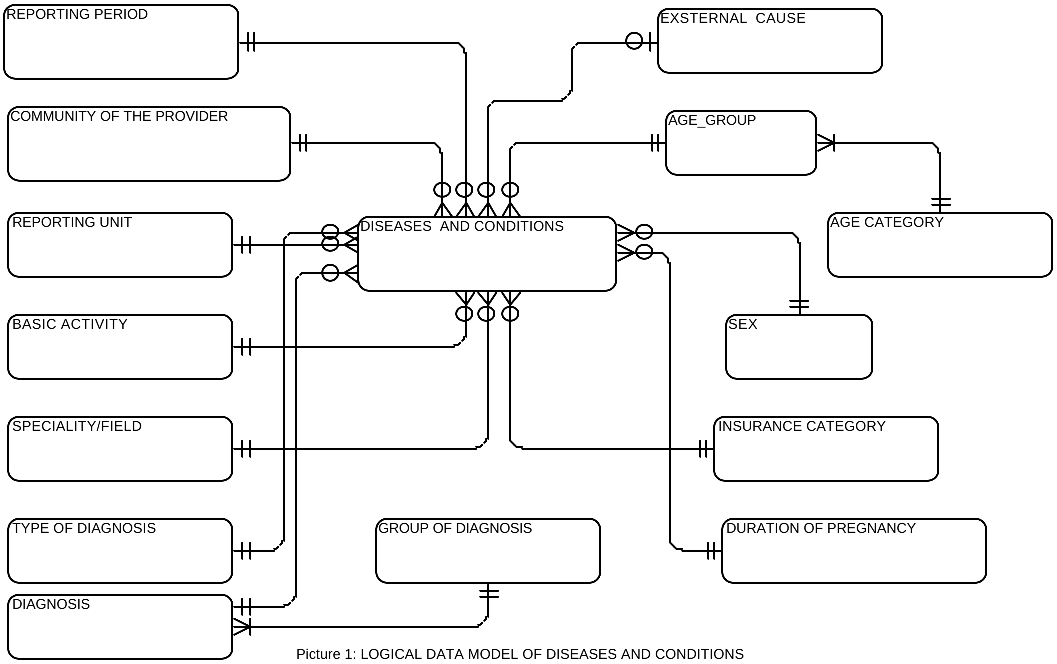
Knowing the existing solution, and having in mind the existing data warehouse technology and the design technique, we made the initial hypotheses:

1. It is possible to make a better solution without changing the data transfer structure. Lowest level of details would further improve the solutions, but cost more.
2. Microsoft SQL Server with included data warehouse OLAP – Analytical Services is the suitable technology for our solution.

The research was divided into two parts. The first part investigates the differences in designing data models for logical, data transfer and data warehouse purposes. The differences, designing decisions, possible results are discussed and the first hypothesis was confirmed. The second part covers the implementation phases of designed data warehouse with special attention on data transformations and physical design. The second hypothesis was also confirmed.

LOGICAL DATA MODEL

The logical data model for Outpatient Health Care Statistics for Diseases and Conditions reflects the real world medicine entities at a primary level. Picture 1 shows a complex logical data model without technology limitations. The model simply explains that Outpatient Health Care Statistics for Diseases and Conditions need data (instances of diseases and conditions) assigned to the reporting period (twice a year: the first half of the year and then the whole year), community of the provider (territory dimension), reporting unit from National data base, basic activity (e.g. general medicine etc.), specialty / field (pediatric specialty), type of diagnosis (draft diagnosis and final diagnosis), diagnosis classification (ICD–10 code, also for injuries at work), external cause (also ICD-10 code), age (group and category), sex, insurance category (e.g. workers, etc.) and duration of pregnancy (valid only for some ICD-10 groups).



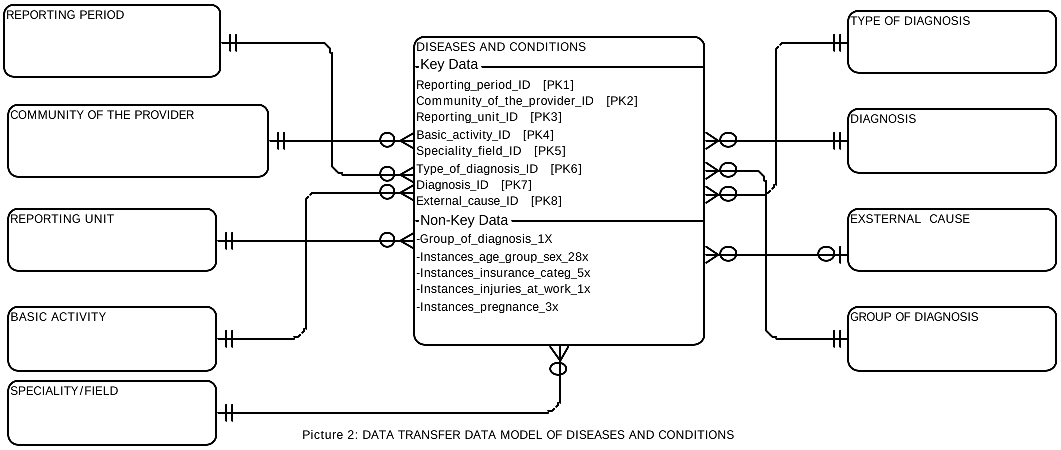
Picture 1: LOGICAL DATA MODEL OF DISEASES AND CONDITIONS

DATA TRANSFER MODEL

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 (7). 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.

Data transfer data model in Picture 2, illustrates the existing standard data structure. The designer of this model made important decisions, necessary to achieve cost effective solutions without employing data warehouse technology. The data structure follows technical limitations of floppy disk data transfer and tries to minimize the volume of data in data transfer and later data store phases. Eventually electronic transfer will replace the floppy data system. The designers did not follow any international data standard, like HL-7. The most important differences in Picture 2, compared with Picture 1 are:

- The primary key is not assigned to all referential entities of the logical model, making the level of details higher than in the logical model. The designer decided that the age / sex, insurance category, injuries at work and pregnancy duration do not interact to such a degree that the lowest level of detail is necessary for statistical purposes.
- The age group in conjunction with sex, insurance category, injuries at work and pregnancy duration instances became attributes, depended only on reporting period, community, reporting unit, basic activity, specialty / field, type of diagnosis, diagnosis and external cause.
- The group of diagnosis is implemented as an attribute, which is redundant for easier analysis.



Picture 2: DATA TRANSFER DATA MODEL OF DISEASES AND CONDITIONS

DATA WAREHOUSE MODELS

Logical data warehouse design follows several design principles (6). The most important design criterion is the usage of data warehouse. This semantic principle is mainly concentrated on the question of **how detailed the data warehouse should be?** The answer determine the amount of data expected and the data warehouse development effort and operational costs.

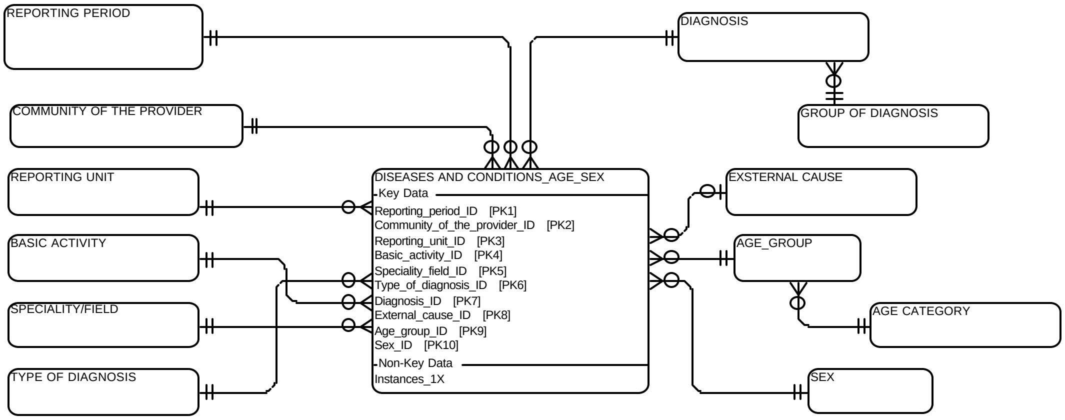
In this example, the designer decided that the lowest level of possible details (Picture 1) is not needed. Therefore the data transfer data model uses some previously mentioned limitations, producing higher granularity. Without these limitations, a designer should accept the logical data model on the lowest detailed level (transaction level) as the data warehouse model. 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 (6) would be much higher. In this case all current answers, and of course all detailed ones impossible with the current solution, would be possible from such data warehouse. A different data transfer model from the existing one should be accordingly designed. Another data transfer and data warehouse technology should be employed on a larger amount of data and, at a higher cost. This research tries to overcome some of the problems, caused by the breakdown of data warehouse into four schemes, by designing and implementing a virtual cube, which is described in detail in the second part of the research.

On the assumption that no improvement in the requirements are needed and that the existing data transfer model is the optimum solution, the following data warehouse and OLAP design should be appropriate. Because of special design principles of data warehouse where star and snowflake structure (6) are the optimum data models, four snowflake data models were designed:

- Four models are needed for optimum OLAP possibilities (3). Common data warehouse can not be defined from existing data transfer model, as the detailed level was lost by purpose.
- All four models have similar fact tables, where primary key is compound from several standard dimensions: the reporting period, community of the provider, reporting unit, basic activity, specialty / field, type of diagnosis, diagnosis code, group of diagnosis and the external cause (mandatory only in the picture 6: injuries at work, in all other cases external cause is present only for a particular diagnosis code which represents injuries at work).
- All fact tables consist of numeric attributes, continuously valued and additive across time.
- Some shared dimensions have a further level of hierarchy: the reporting period (calendar hierarchy), reporting unit (organizational hierarchy) and diagnosis (group of diagnosis).
- All four models are basically star schemes, except the group of diagnoses connected with the diagnosis and the age group connected with the age category – snowflake schema.
- All four models have only one attribute – "instances" according to the primary key and share common dimensions.
- Beside common – shared dimensions, models employ special dimensions: in picture 3 the age group and sex, in picture 4 the insurance category and in picture 5 the pregnancy.
- Nearly all of the described dimensions are basically slowly changing over the time. The research incorporates the simplest method of solving the problems, overwriting old values with losing the ability to track the history. If a more precise solution is requested, the data warehouse model can be promoted to such capabilities, following the known techniques (6).

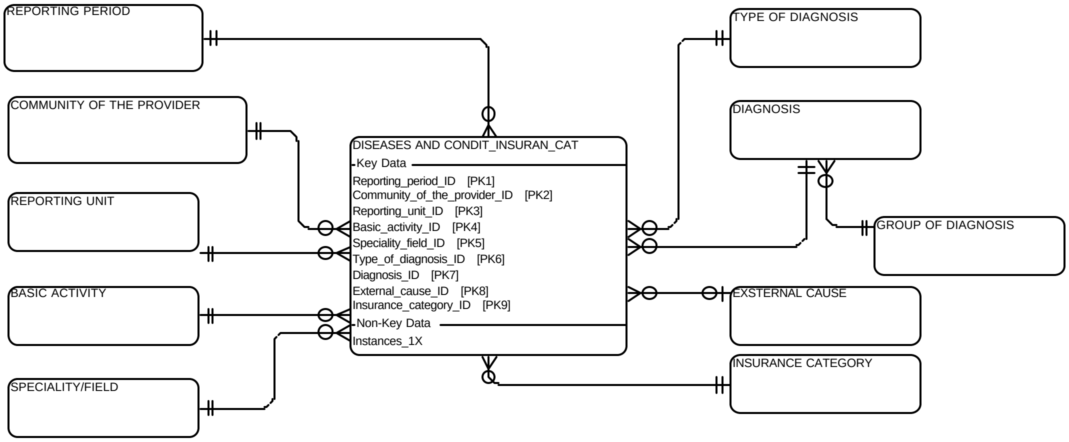
The design of research data warehouse involved the following steps: choosing the Outpatient Health Care Statistics as a potential research problem, choosing the granularity – level of detail (as defined in data transfer structure), choosing the dimensions and finally choosing the measured facts, which are the final goals of data warehouse design.

Picture 3 describes a data warehouse model in which instances are depended on common – shared dimensions together with age groups and sex dimensions. Age dimensions have a further level of hierarchy – the age category.



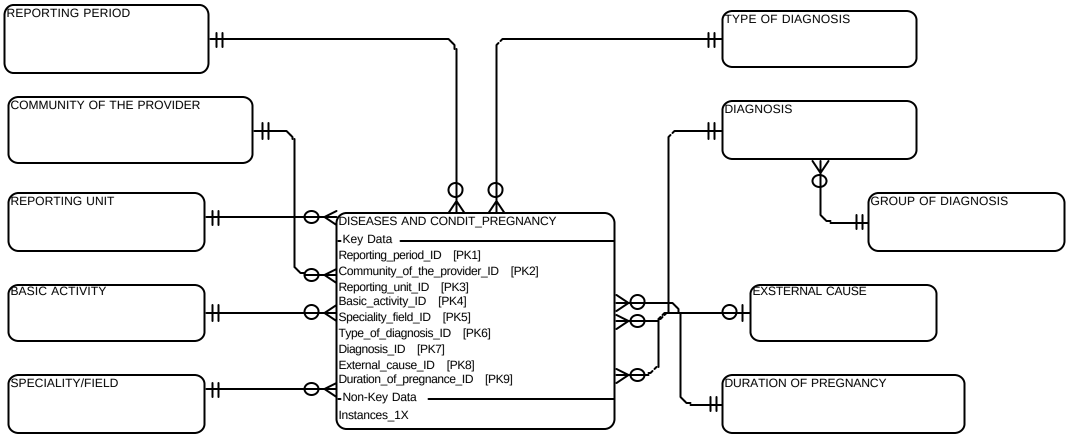
Picture3: DATA WAREHOUSE MODEL OF DISEASES AND CONDITIONS BY AGE AND SEX

Picture 4 describes data warehouse model where instances depend on common – shared dimensions, together with the insurance category.



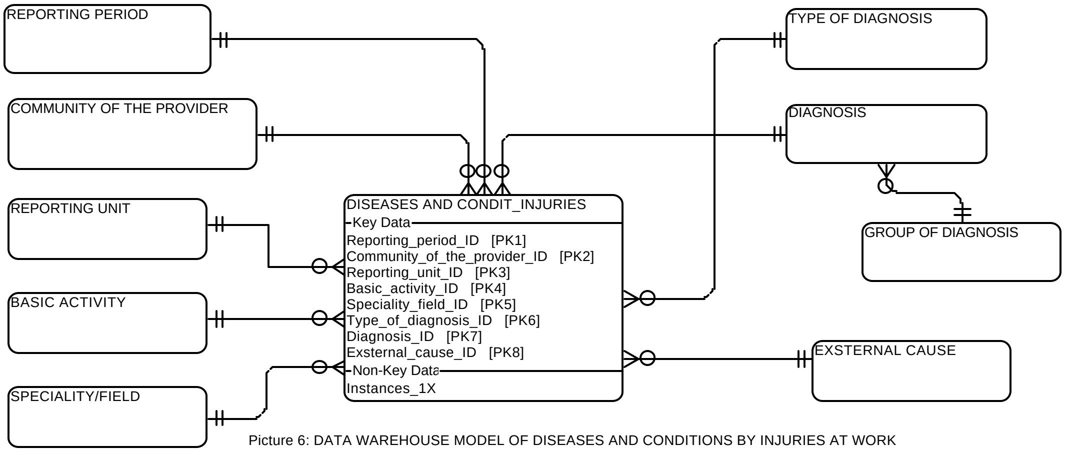
Picture 4: DATA WAREHOUSE MODEL OF DISEASES AND CONDITIONS BY INSURANCE CATEGORY

Picture 5 describes data warehouse model where instances depend on common – shared dimensions together with the duration of pregnancy. There is however the semantic question if an external cause is interrelated with a diagnosis.



Picture 5: DATA WAREHOUSE MODEL OF DISEASES AND CONDITIONS BY DURATION OF PREGNANCY

Picture 6 describes data warehouse data model where instances depend only on common – shared dimensions, but the diagnosis covers only injuries at work.



Picture 6: DATA WAREHOUSE MODEL OF DISEASES AND CONDITIONS BY INJURIES AT WORK

During extraction of dimensional data several implementation problems arise:

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 values representing the 'NULL'. All those values had to be translated to 'NULL'.
3. Dimensions inherent in the data transfer data structure (sex, age group, insurance category, and pregnancy). The data for those dimensions had to be entered into dimensional tables. Modification of dimensional data would be necessary only when data structure change.
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. The integration of data cubes in the virtual data cube removes all dimensions, which are not present in all data cubes.

There are important implementation decisions that have to be made while designing data cubes:

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. 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 virtual data cube. The virtual cube provides a view over several cubes, which have shared dimensions. Dimensions that are not common to all cubes are left out. Measures are summarized over omitted dimensions. In case of research 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).

CONCLUSION

The results of the research celarly confirm the hypotheses. Better solutions can be designed and implemented to fulfill all of the goals of a data warehouse: provide access to statistical data on national level, ensure consistent data, separate and combine all the data in a warehouse (slice and dice) so data can available for all statistical requirements to effectively support the Ministry of Health in managing Health Care National Wide Policy. The solution was successfully developed by using Microsoft SQL Server with its OLAP – Analytical Services capabilities.

REFERENCES

1. Anahory S., Murray D. (1997). Data warehouse in the real world, Addison Wesley
2. Imhof C., Loftis L., Geiger J.G. (2001). Building the Customer-Centric Enterprise: Data Warehousing Techniques for Supporting Customer Relationship Management, Wiley.
3. Inmon W.H., Welch J.D., Glassey K.L. (1997). Managing the Data Warehouse, Wiley Inc.
4. Inmon W.H. (1996). Building the Data Warehouse, 2.ed. Wiley
5. Kimball R. (1996). The Data Warehouse Toolkit, John Wiley & Sons, Inc.
6. Kimball R. (1998). The Data Warehouse Lifecycle Toolkit, Expert Methods for Designing, developing and Deploying Data Warehouses, Wiley.
7. Lazar D. (1997). Microsoft Strategy for Universal Data Access, USA.
8. Microsoft Corporation. (1998). The Microsoft Data Warehousing Strategy, USA.
9. Microsoft Corporation. (1998). Microsoft SQL Server Decision Support Services. USA.
10. Microsoft Corporation. (2000). http://www.microsoft.com/sql/beta
11. Silverston L., Inmon W.H., Graziano K. (1997). The Data Model Resource Book, Wiley.
12. Willcocks L.P., Feeny D.F. (1997). Managing IT as a Strategic Resource, McGraw-Hill.