Holger Günzel, Wolfgang Lehner, Stein Eriksen, Jon Folkedal

Modeling of census data in a multidimensional environment

Erstveröffentlichung in / First published in:

Advances in Databases and Information Systems. Poznan, 7.-10. September 1998. Springer. S. 363–368. ISBN 978-3-540-68309-4.

DOI: https://doi.org/10.1007/BFb0057748

Final edited form was published in "Advances in Databases and Information Systems. Poznan 1998", S. 363–368,
ISBN 978-3-540-68309-4
https://doi.org/10.1007/BFb0057748

Modeling of Census Data in a Multidimensional Environment

Holger Günzel[1], Wolfgang Lehner[1], Stein Eriksen[2], Jon Folkedal[2]

[1]Department of Database Systems, University of Erlangen-Nuremberg,
Martensstr. 3, D-91058 Erlangen, Germany
{guenzel, lehner}@informatik.uni-erlangen.de
[2]Statistics Norway, KOSTRA Development Team,
Postbox 8131 Dep, N-0033 Oslo, Norway
{ser, jfo@ssb.no}

Abstract. The general aim of the KOSTRA project, initiated by Statistics Norway, is to set up a data reporting chain from the norwegian municipalities to a central database at Statistics Norway. In this paper, we present an innovative data model for supporting a data analysis process consisting of two sequential data production phases using two conceptional database schemes. A first data schema must provide a sound basis for an efficient analysis reflecting a multidimensional view on data. Another schema must cover all structural information, which is essential for supporting the generation of electronic forms as well as for performing consistency checks of the gathered in-formation. The resulting modeling approach provides a seamless solution for both proposed challenges. Based on the relational model, both schemes are powerful to cover the heterogeneity of the data source, handle complex structural information, and to provide a versioning mechanism for long term analysis.
Keywords: Data analysis, metadata, multidimensional model, census data

1 Introduction

In this paper we report the results and experience of an international cooperation between the data warehouse research group of the University of Erlangen-Nuremberg (Germany) and the KOSTRA development team of Statistics Norway (SN). The acronym KOSTRA (Kommune STat RApportering; [6]) stands for a project setting up a new statistical database, which is suitable for gathering, accumulation and analysis of census data of institutions of all norwegian municipalities.

1.1 The KOSTRA Reporting Chain

The currently existing reporting chain consists of a manual flow of data from widespread municipalities to SN. Data according to a specific topic are gathered by filling out forms, transported in a more or less heterogeneous way like sheets of papers, disks or modem to SN and fed into a database. The scenario causes problems in handling the immense number of reports, data inconsistencies, and missing data. In general, this errorprune process incorporating a lot of manual corrections should be replaced by an automated solution.
The KOSTRA project (figure 1) intends to simplify this reporting chain and increases the correctness of data by replacing paper sheets by an electronic one. As a

Final edited form was published in "Advances in Databases and Information Systems. Poznan 1998", S. 363–368,
ISBN 978-3-540-68309-4
https://doi.org/10.1007/BFb0057748

consequence, the reporting will be standardized and automated from the point of data gathering, storage, and analysis. In a first step, data from several electronic sheets are gathered into a local database at each municipality. In a second step, data are encrypted and transmitted to SN. A Common Reception Service (CRS) at SN receives and stores the reported data for enabling an extraction process for analysis purposes. In a last step, the collected and cleansed data from the central database are analyzed at several places like internally within SN for public statistics or at different municipalities for their internal analysis providing a loop back analysis. Due to the analysis process, SN already uses an existing solution called Regional Database ([3]).

1.2 The Common Reception Service

As illustrated in figure 1, our design approach of the CRS consists of the Central Reception Server and two different kinds of structural information database. The Central Reception Server collects and physically stores the incoming data. Generally spoken, this server corresponds to a "Data Warehouse" ([4]). Structural information additionally covers data about the structure and the process correctly filling out an electronic sheet. Thus, we divide structural information into a dynamic part, used for the interaction with the user, and a static part, remaining stable during one gathering period. Metadata about transmission or the identification of different sources must be seen in a more general context and is beyond the scope of this paper ([8]).

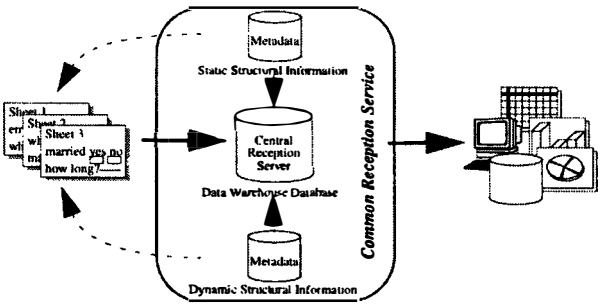

Fig. 1. The KOSTRA reporting chain

1.3 Difficulties in Modeling the Scenario

An automated reporting of census data causes several general problems and requirements. The data sources, the gathered data, and the structural information about the data cause an immense heterogeneity. To consider only the data perspective, there are forms with more than 100 questions. Two further requirements improving the current service concern the reporting process. On the one hand, every form or component may change during the reporting time, i.e. versioning approaches have to be considered. Moreover, SN requires a mechanism for generating and delivering forms to the different municipalities without an explicit intervention. On the other hand, monitoring of the reporting chain and checking against inconsistencies of the incoming data is extremly important. Altogether our innovative approach for an adequate data model contains exceedingly flexibility. A

Final edited form was published in "Advances in Databases and Information Systems. Poznan 1998", S. 363–368,
ISBN 978-3-540-68309-4
https://doi.org/10.1007/BFb0057748

straightforward and traditional modeling approach would focus either on the data gathering and storage or on the perspective of an efficient analysis.
The structure of the paper is as follows: In section two we deal with the structural information of forms according to their usage and mapping to the relational model. The third section defines the requirements of an efficient analysis and proposes a conceptual database schema. The paper concludes with a summary and an outlook.

2 Metadata Schema Design for Structural Information

The structure of an electronic sheet or form represents the fundamental basis for gathering and analyzing of census data. This structural information controls the sheet generation process, supervises the fill out process, and determines the necessary consistency checks. As we will see in a consequence, it intensively affects the CRS.

2.1 Physical Structure of a Form

Census data can be divided in several topics, where each topic is related to a set of questions and represented by a specific sheet. Therefore the set of questions and the structure of the form are associated to a topic. Although the forms appear to be different and heterogeneous, a structural analysis of the existing forms shows that they all are based on the same skeleton. Each form is composed of an identification block consisting of information like the social number, the address, and one or more information blocks carrying data provided by the sender. In turn, each information block consists of a set of single questions. This combination of questions and information blocks belonging to the same topic is static during a reporting period.
In most cases the structure of electronic forms consists of more than one information block and more questions depending on each other. Therefore, a design model for a 'dynamic' structure is required. For example, an information block may be skipped, if a start-up question of one block links another block. Therefore, the structure at the instance level is highly dynamic caused by different ways of answering questions.

2.2 Static Structural Information

The knowledge of a form structure needs to be centrally stored for enabling a consistent data gathering process. During filling out a form, static structural information only influences the layout of the form and is independent of the data itself. Forms need to be designed very flexibly, because every form may be modified every reporting period resulting in a versioning of the single components of a form. Therefore the different structures of the sheets are stored within a static part of an information repository. The term 'static' emphasizes the fact that this kind of structural information keeps stable at least throughout one reporting period. To implement the physical structure, our proposal includes a metadata repository having an own conceptional schema and reflecting the relevant structure of the forms.
For the implementation of the scheme based on a traditional relational datamodel, we propose several tables for the electronic sheet itself, the information blocks and the questions as the basic items of the structural requirements. To fulfill the versioning, several time clauses denote the validity periods of the single components of the sheet.

Final edited form was published in "Advances in Databases and Information Systems. Poznan 1998", S. 363–368,
ISBN 978-3-540-68309-4
https://doi.org/10.1007/BFb0057748

2.3 Dynamic Structural Information

From a processing point of view, we demand that metadata monitors and guides each step of answering a sheet because some questions or other parts like information blocks depend on the answers of other components. To come up with a solution for this requirement, the semantic structure of the sheets need a dynamic part of an adequate representation within a metadata repository. In this way, a supervision of answering the question of a form and providing online as well as offline consistency checks are possible. However, the content of the dynamic part is influenced by the recorded information and directly effects the fill out process: On the one hand the dynamic structural information checks the plausibility of the values through rules or threshold values. On the other hand it monitors the correct input procedure of the sheet. If somebody answers that he is male, then all questions about pregnancy automatically should be skipped or should not be possible to answer.
Our proposal of modeling dynamic structural information is based on the ECA concept ([2]), considering the answer of a question as an event which is checked with a condition and followed by an action, if the condition evaluates to true. Since semantic checks are necessary before entering a sheet component and after answering a question, we require enter as well as exit conditions and actions. The enter condition always considers previous answers, whereas the exit condition always uses the answers of the current component. This proposal of ECA chains needs an intensive connection of user interaction and system-based semantic checks. It should mentioned here that the ECA principle may be also used for global off-line consistency at SN. Since these dynamic structures exist on all granularities of a sheet, relational tables are required for questions, information blocks and sheets.

3 Multidimensional Schema Design for Analysis Information

The promising profit of the discussed reporting chain emerges from the possibility of analyzing the gathered data. This yields in the requirements for an adequate conceptional schema to close the gap between integrating and storing the census data in a database system and efficient analyzing at the user level. This last requirement corresponds to what is generally known under 'OLAP' ([1]). Within the design phase, the multidimensional data model (cube) turned out to be an adequate model for the kind of sophisticated data analysis. In the multidimensional way of thinking, the data cube consists of several dimensions covering the structural information and cells storing the numeric (census) data. For a seamless implementation of this model, we fall back on existing relational technology. Our proposal uses a relational database engine with a relational data model generally known as star-schema ([4]). Within our star-schema approach, a fact table holds all census data. Dimension tables, organized around the fact table like a star, represent all structural information.

3.1 Multidimensionality

Typical analysis queries may be classified into time series analysis, sender analysis or a topic oriented analysis. Every analyst wants an uniform view on the data. In general not the individual object but a global view on a set of data is desired.

Final edited form was published in "Advances in Databases and Information Systems. Poznan 1998", S. 363–368,
ISBN 978-3-540-68309-4
https://doi.org/10.1007/BFb0057748

Therefore we need a data model which covers analysis perspectives and is tightly coupled with the application. In the following we propose a reasonable and simple multidimensional model to avoid sparsity and achieve a clear structure. Time and sender reflect two of four dimensions, since for each sender a form is filled out exactly once a reporting period. The other two dimensions follow the structure of a question, which can be divided into a header ('answer dimension') and a stub ('objective dimension'). All dimensions together determine the single facts, i.e. an answer of a single question. Figure 2 illustrates the multidimensional modeling idea and the connection between the time, sender and the 'question' array.

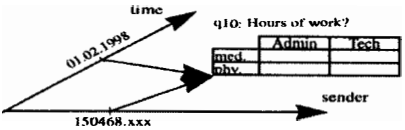

Fig. 2. The multidimensional approach

3.2 The Data Cube Model

Our proposal uses the dimensional modeling or star schema approach at the relational level. The central fact table within a star schema contains a composite primary key of the participating dimensions and the single numeric facts. The CRS fact table requires a composite key with the sender identification (SID), a reporting time interval and an ID of the objective and answer dimension. The fact attribute covers exactly one answer belonging to exactly one sender, in a time interval for one objective-answer relation.

```
FactTable (SID, ReportingInterval, ObjectiveID, AnswerID, Fact)
```

Since different questions have answers of different type, problems arise with that heterogeneous fact types of the fact attribute. The trouble is solved with our 'hyperfact' approach. A virtual hyperfact table consists of a union of several type specific fact tables with only one homogeneous fact type. The fact attribute of the 'hyperfact' table in turn holds the table names of the real fact tables (subordination, [7]). Therefore, an access of a specific fact requires two steps. In a first step the name of the real fact table is determined. In a second step, the answer is looked up type specific in the real fact table.

3.3 The Dimensional Tables

The dimensional tables cover all the structural information identified by the primary key attributes of the fact table. According to the four dimensions mentioned earlier, we require an answer, objective, time and source dimensional table to provide additional information and to enable OLAP analysis processes. The answer/ objective tables hold the structures for the header/ stub of a question. Since the time dimension orients at the gregorian calendar built in every database system, we do not require a specific time dimension table. The dimension table for reflecting the different senders differs to the answer or objective table. Problems arise, because a data source may be a person or a department, but the submitter, who is responsible for delivering these facts, may be always an institution. Therefore, there is a need for

Final edited form was published in "Advances in Databases and Information Systems. Poznan 1998", S. 363–368,
ISBN 978-3-540-68309-4
https://doi.org/10.1007/BFb0057748

an explicit description of the source type and the submit type. Furthermore, additional attributes for the validity and attributes covering data for a personal identification like the social number or the department identifier.

3.4 Conclusion of the Modeling Approach

Using the proposed multidimensional data schema, the fact table grows fast, implying that no new techniques at the database-level are required, but for commonly used methods for improvements of access performance like bitwise indices or partitioning. The dimensional tables compared to other modeling techniques are very small. Altogether, the effort keeps low, because the dimensions and therefore the overhead is limited. New questions or different forms lead to a slightly larger cube, but do not result in more dimensions or relational tables. The heterogeneity of the electronic sheets are seamlessly covered through the hyperfact-approach.

4 Conclusion and Open Issues

Our modeling approach of the CRS for the new KOSTRA reporting chain is based on the structural information of the existing forms. In that way, the data model for the Central Reception Server is designed to handle incoming data and provide efficient analysis access to outgoing information. The fundamental basis is the multidimensional view on data. Instead of modeling each single answer and question within a straightforward approach, we divided the structural information from analysis information to achieve independency and performance. This implies, that the model is flexible enough to deal with extensions like new electronic sheets or data, versions and new data types. From a theoretical point of view, we handled versioning problems and immense heterogeneity within a multidimensional context. From the practical point of view the implementation is currently under development at SN. Our approach seems specific for this scenario, but could be simply modified and adopted to similar problems.

References

1. Codd, E.F.; Codd, S.B.; Salley, C.T.: Providing OLAP (On-line Analytical Processing) to User Analysts: An IT Mandate, White Paper, Arbor Software Corporation, 1993
2. Dayal U.; Hsu, M.; Ladin, R.: Organizing Long-Running Activities with Triggers and Transactions, SIGMOD Conference 1990, pp. 204 – 214
3. Eriksen, S.: Data Warehousing in Statistics Norway - The RD application, Statistics Norway, internal report, Oslo, 1997
4. Kimball, R.: The Data Warehouse Toolkit, John Wiley & Sons, Inc., New York, 1996
5. Smith, J.M.; Smith, D.C.P.: Database Abstractions: Aggregation and Generalization, ACM Transactions on Database Systems 2(1977)2, pp. 105-133
6. Titlestad, G.: KOSTRA - Model and solutions for a management for information ressources (in norwegian: KOSTRA - Modell og løsninger for informasjonsressurs-forvaltning), Statistics Norway, internal report, Oslo, 1996
7. Wedekind, H.: Database Systems I (in german: Datenbanksysteme I), Mannheim, BI Wissenschaftsverlag, 1981
8. Yang, J.J.: Overall user requirements to Metadata Management Systems, Statistics Norway, KITH, internal report, Trondheim, 1997