

A Survey “Off the Record” - Using Alternative Data Models to increase Data Density in Data Warehouse Environments

Victor Gonzalez-Castro

Lachlan MacKinnon

School of Mathematical and Computer Sciences, Heriot-Watt University,
Edinburgh UK, EH144AS.
<mailto:{victor.lachlan}@macs.hw.ac.uk>

Abstract. It is proposed to use alternative Data Models to increase data density in Data Warehouse environments by abandoning the traditional record storage structure.

Introduction

In Data Warehouse environments data sparsity is a common issue that has been addressed in past years using different techniques but remains an unresolved problem. The techniques developed to date to tackle this problem have been based on the relational model. At the same time other area of research has been focusing on alternative data models that abandon the traditional record storage/manipulation structure. We are investigating the use of alternative data models to increase the internal data density in Data Warehouses with the idea of decreasing the data sparsity and there by decreasing the final size of the database.

State of the art

We intend to compare several alternative Data Models:

- Relational Model. Uses the traditional record storage/manipulation structure. It is the base model against which the other models will be compared.
- The Triple store [1] [2]. This model does not use a record structure; instead it uses a structure called *the name store* to keep all the names, and *triplets* to construct the processing structure.
- The Binary Model. [4]. It considers that all tables in the system are *binary tables*.
- The Associative Model [3]. It comprises two types of data structures *Items* and *Links*. It differs from binary and Triple store in one fundamental way; Associations themselves may be either the source or the target of other associations.
- The Transrelational Model TM [5]. It *keeps the Relational model itself but abandon the record storage structure*.

Proposal

Data sparsity is originated from the aim of answering all the possible user queries from the information stored in a Data Warehouse. It is the result of the Cartesian product of all the dimensions and all the aggregation levels. In Figure 1 we show a Time dimension and aggregation levels demonstrating how nulls are generated. Figure 2. Is a graphical representation of a Cartesian product of two dimensions and shows how Data sparsity is originated and its relation with data density.

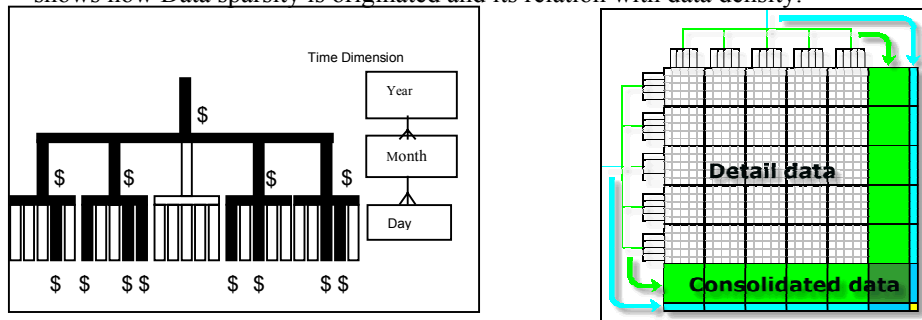


Fig. 1. A three level dimension and Nulls.After [6]. **Fig. 2.** Data sparsity and density, from [6]

The aim of our research is to carry out an impartial survey and comparison of whatever or not the use of alternative data models can improve the Data density in Data warehouse environments and at the same time observe the effect that such data density increase has on the Data sparsity.

We intent to use a Data base implementation of each data model. Relational Model – A commercial RDBMS. Triple Store Model – TriStarp. Binary Model – Monet DB. Associative Model – Sentences DB. Transrelational – since there is currently no instantiation of the Transrelational model available, We will build an implementation of the essential algorithms.

We will use the TPC-H data set with the Data models/products mentioned above, to run a set of benchmark metrics to determine relative performance, and then consider relative data density and sparsity.

References

1. G C H Sharman and N Winterbottom, The Universal Triple Machine: a Reduced Instruction Set Repository Manager. Proceedings of BNCOD 6, pp 189-214, 1988.
2. TriStarp Web Site: <http://www.dcs.bbk.ac.uk/~tripsarp>. Updated November, 2000.
3. The Associative Model of Data. Simon Williams,2002, Second Edition, Lazy Software Ltd. ISBN: 1-903453-01-1 www.lazysoft.com
4. MonetDB. ©1994-2004 by CWI. <http://monetdb.cwi.nl>
5. Date, C.J. An introduction to Database Systems. Appendix A. The Transrelational Model , Eighth Edition. Addison Wesley. 2004. USA. ISBN: 0-321-18956-6.
6. Pendse Nigel. Database explosion. <http://www.olapreport.com> Updated Aug, 2003.