

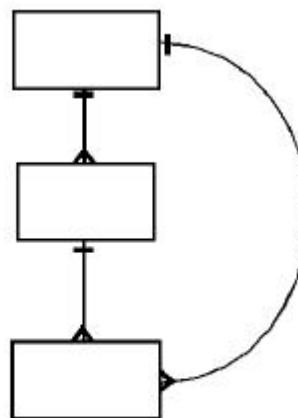
Thoughts about Data (1): Data Warehouses and Entity Models

By Mike King

(Read the left column completely before the right column - it seems easier to read in two columns.)



Normalized Model



De-normalized Model

A recent article in Datamation started with the words 'Forget Everything you know about entity relationship modeling. If your DBA is proud of normalizing all your databases into Fifth Normal Form, tell her to read this article.'. The article went on to discuss dimensional modeling and the effectiveness of a star schema for high performance response to business questions.

The message of this discussion is 'Forget Everything you know about entity relationship modeling - at your peril'. This will be motivated from a number of perspectives.

In his book 'Building the Data Warehouse', W.H. Inmon says (page 186) 'In theory, it is possible to build the architected environment without a data model. However, in practice it is never done. Trying to build an architected environment without a data model is like trying to navigate without a map. It can be done, but it is very prone to trial and error. ...'

It is not always realized that the entity model, far from being a technical thing, is the clearest and most precise vehicle for understanding and communicating about the business. It should be the property of the USERS to a greater extent than it belongs to the Information Technology Division. It really shows the business concepts, whether or not computers are used.

Without a data model, it is not possible for different groups of users to even realize that they do not have identical perceptions of how their business operates, let alone arrive at a true and precise consensus.

How then could the users of a Data Warehouse properly interpret the business reports they obtain, if the warehouse is not based on a data model? Clearly, the interpretation would often be wrong, and the resultant semantic disintegrity could result in strategic decisions being made for the wrong reasons.

A Data Warehouse is the target of queries which often involve huge numbers of records from many tables, and so it needs to be specially organized to respond quickly to such heavy loads. In the jargon of data analysis, the tables need to be denormalized.

The perception has grown that normalization is therefore outdated in the Data Warehouse environment.

However, once we remind ourselves that normalization is intended to remove the duplication of business facts, then clearly an unnormalized data warehouse will contain duplication, and will be bound to contain contradictions - i.e. data full of corruption - unless the duplication is controlled through special consistency checks.

This means that normalization cannot be ignored. It must serve as the initial basis from which we can consciously deviate while applying special compensating checks.

It is a deep understanding of normalization which informs us what standard denormalizations to apply, and what standard integrity checks must accompany them.

Thus the essential basis for a Data Warehouse is a fully normalized entity model, from which a denormalized model (the data warehouse model) is consciously constructed.