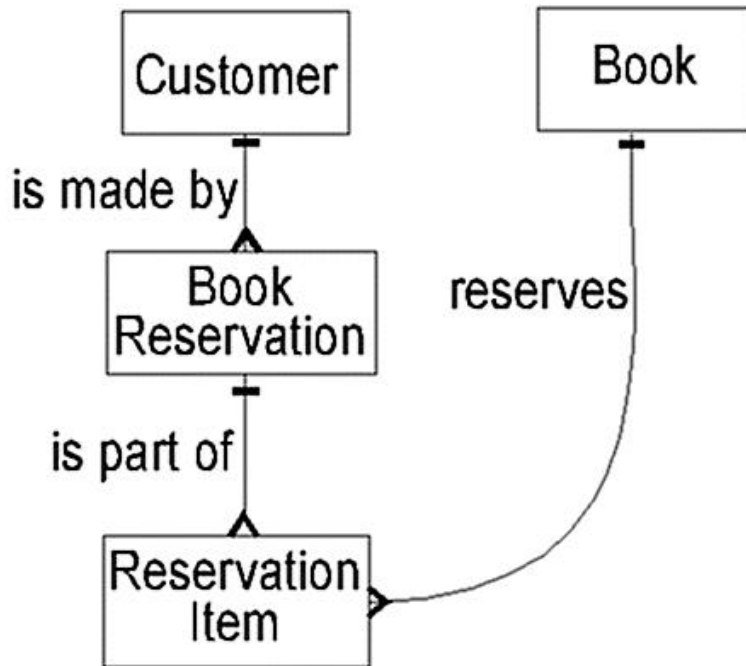


# Thoughts about Data (3): Conflicting sources for the Data Warehouse

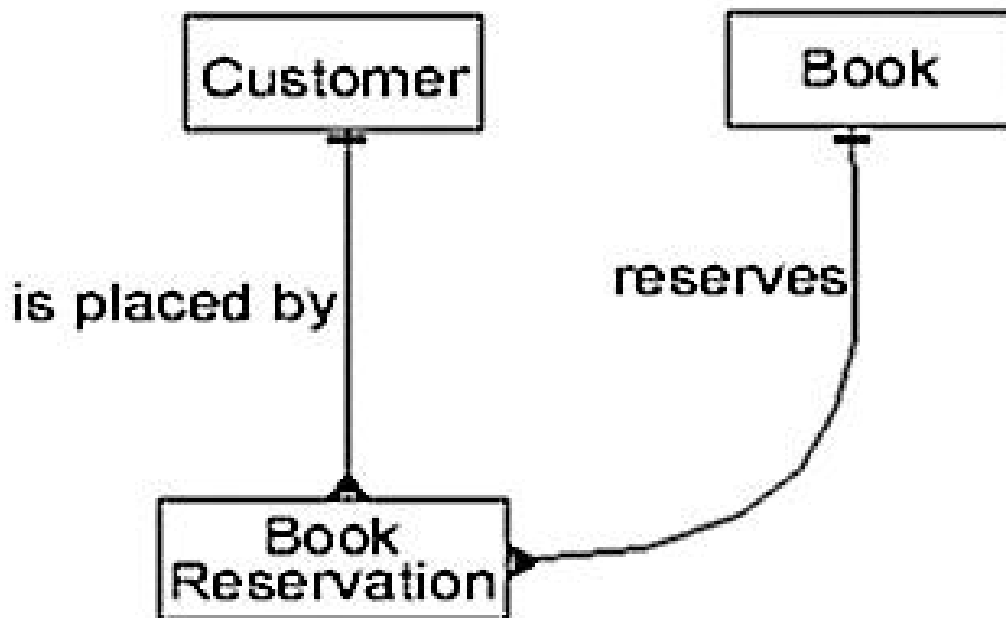
By Mike King

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

---



Model A



Model B

---

The data warehouse is supposed to collect all the organization's business information together to support analysis for business decisions and actions, but what if the information sources are inconsistent?

If the organization has never reached the stage of using a corporate entity model to coordinate the use of its information resource, then the same business data is very likely to be held in more than one place, without consistency of structure.

Why is this so likely? The answer is found in practical experience of entity modeling, where it is observed that two different teams who set out to model the same business situation, rarely construct exactly the same entity model. Sometimes the models are even radically different!

One can perceive two reasons for this divergence. Firstly, business people do not all have the same detailed perceptions of how the business is run (in fact one of the benefits of group modeling sessions is to supply a sufficiently clear and precise discussion framework, that these detailed differences can be perceived and resolved) and secondly, even when there is agreement about the business requirements, there is still more than one way to represent them.

One should not be surprised at this subjectivity. After all, data structures are artifacts rather than pre-existing natural objects. Once we accept this inevitable subjectivity as part of the modeling process, it is clear that we need criteria to help us choose a 'good' model from alternative structures.

The following criteria have been found useful in practice. The model must:

- be correct in a business sense
- be non-redundant except for controlled redundancy
- be complete (able to support all information requirements)
- be as simple as possible subject to the above conditions
- be able to accommodate changes in the business while as much as possible retaining its present form (stability)

A data model which fails the first three conditions must be discarded or improved. Whenever there is more than one model that satisfies the first three conditions, then the simplest and most stable is chosen.

An example shows the case of two groups who have the same view of the business process, but different models. A model of library data is being constructed. Both groups accept that the library must be able to accept book reservations over the telephone, and that during one telephone call, a customer may reserve more than one book.

To record a phone call using model A, one Book Reservation containing two (say) Reservation Items would be recorded in the database. With Model B, two Book Reservations would be recorded, each for one book.

Both models are able to record full details of the phone call and are non-redundant. In this case it is not clear whether either model is simpler or more stable than the other.

A Data Warehouse which combined two such sources (different branch libraries, say), would have to rationalize the data into a single coherent view, to support cross branch analysis.

In this case it can be done, however in some cases the structure incompatibilities might prevent any rationalization.