

Design Tip #131 Easier Approaches For Harder Problems

By Ralph Kimball

Data warehouses are under intense architectural pressure. The business world has become obsessed with data, and new opportunities for adding data sources to our data warehouse environments are arriving every day. In many cases analysts within our organizations are discovering "data features" that can be demonstrably monetized by the business to improve the customer experience, or increase the conversion rate to purchase a product, or discover a demographic characteristic of particularly valuable customers. When the analyst makes a compelling business case to bring a new data source into our environment there is often pressure to do it quickly. And then the hard work begins. How do we integrate the new data source with our current environment? Do the keys match? Are geographies and names defined the same way? What do we do about the quality issues in new data source that are revealed after we do some sleuthing with our data profiling tool?

We often do not have the luxury of conforming all the attributes in the new data source to our current data warehouse content, and in many cases we cannot immediately control the data collection practices at the source in order to eliminate bad data. As data warehouse professionals we are expected to solve these problems without complaining!

Borrowing a page from the agile playbook, we need to address these big problems with an iterative approach that delivers meaningful data payloads to the business users quickly, ideally within days or weeks of first seeing the requirement. Let's tackle the two hardest problems: integration and data quality.

Incremental Integration

In a dimensionally modeled data warehouse, integration takes the form of commonly defined fields in the dimension tables appearing in each of the sources. We call these conformed dimensions. Take for example a dimension called Customer that is attached to virtually every customer facing process for which we have data in the data warehouse. It is possible, maybe even typical, for the original customer dimensions supplied with each source to be woefully incompatible. There may be dozens or even hundreds of fields that have different field names and whose contents are drawn from incompatible domains. The secret powerful idea for addressing this problem is not to take on all the issues at once. Rather, a very small subset of the fields in the Customer dimension is chosen to participate in the conforming process. In some cases, special new "enterprise fields" are defined that are populated in a consistent way across all of the otherwise incompatible sources. For example a new field could be called "enterprise customer category." Although this sounds like a rather humble start on the larger integration challenge, even this one field allows drilling across all of the data sources that have added this field to their customer dimension in a consistent way.

Once one or more conformed fields exist in some or hopefully all of the data sources, then you need a BI tool that is capable of issuing a federated query. We have described this final BI step many times in our classes and books, but the essence is that you fetch back separate answer sets from the disparate data sources, each constrained and grouped only on the conformed fields, and then you sort merge these answer sets into the final payload delivered by the BI tool to the business user. There are many powerful advantages to this approach, including the ability to proceed incrementally by adding new conformed fields within the important dimensions, by adding new data sources that sign up to using the conformed fields, and being able to perform highly distributed federated queries across incompatible technologies and physical locations. What's not to like about this?

Incremental Data Quality

Important issues of data quality can be addressed with the same mindset that we used with the integration problem. In this case, we have to accept the fact that very little corrupt or dirty data can be reliably fixed by the data warehouse team downstream from the source. The long-term solution to data quality involves two major steps: first we have to diagnose and tag the bad data so that we can avoid being misled while making decisions; and second, we need to apply pressure where we can on the original sources to improve their business practices so that better data enters the overall data flow.

Again, like integration, we chip off small pieces of the data quality problem, in a nondisruptive and incremental way so that over a period of time we develop a comprehensive understanding of where the data problems reside and what progress we're making in fixing them.

The data quality architecture that we recommend is to introduce a series of quality tests, or "screens" that watch the data flow and post error event records into a special dimensional model in the back room whose sole function is to record data quality error events. Again, the development of the screens can proceed incrementally from a modest start to gradually create a comprehensive data quality architecture suffusing all of the data pipelines from the original sources through to the BI tool layer.

We have written extensively about the two overarching challenges of integration and data quality. Besides our modeling and ETL books, I've written two white papers on these topics if you want to take a deeper dive into the detailed techniques. While these papers were commissioned by Informatica, the content is vendor agnostic. [Links](#) to the white papers (*An Architecture for Data Quality written in October 2007* and *Essential Steps to an Integrated DW written in February 2008*) are available from our website, however be forewarned that you'll need to register with Informatica to download them.