

Design Tip #126 Disruptive ETL Changes

By Ralph Kimball

Many enterprise data warehouses are facing disruptive changes brought about by increased usage by operations and the exploding interest in customer behavior. My impression is that many shops have implemented isolated adaptations to these new forces but haven't lifted their gaze to the horizon to realize that the data warehouse design landscape has changed in some significant ways, especially in the ETL back room.

The overarching change is modifying the EDW to support mixed workload operational applications where low latency data coexists with historical time series and many other customer facing applications and data sources. Yes, of course we talked about operational data even before Y2K, but the early ODS implementations were restricted to tiny answer set fetches of highly constrained transactional questions such as "was the order shipped?" Today's operational users are drastically more demanding.

Here are seven disruptive changes coming from operational requirements. I wouldn't be surprised if you were facing all seven of these at once:

1. *Driving data latency toward zero.* The urge to see the status of the business at every instant is hard to resist. Not everyone needs it, but for sure someone will claim they do. But as you approach zero latency data delivery, you have to start throwing valuable ETL processes overboard, until finally you only have the vertical retrace interval on your display (1/60th of a second) in which to perform useful ETL work. Although this extreme is admittedly ridiculous, now I have your attention. And keep in mind, if you have true zero latency data delivery, the original source application has to provide the computing power to refresh all the remote BI screens. The lesson here is to be very cautious as your users tighten their requirements to approach zero latency delivery.

2. *Integrating data across dozens, if not hundreds, of sources.* Customer behavior analytics is the rage in the operational/BI world, and there is a lot of money chasing behavior data. Operational and marketing people have figured out that almost any data source reveals something interesting about customer behavior or customer satisfaction. I have seen a number of shops struggling to integrate dozens of not-very-compatible customer facing data collection processes.

3. *Instrumenting and actively managing data quality.* After talking idealistically about data quality for fifteen years, the data warehouse community is now turning actively to do something about it. Within the data warehouse, this takes the form of data quality filters that test for exceptional conditions, centralized schemas that record data quality events, and audit dimensions attached to final presentation schemas. Things get really interesting when you try to address this requirement while simultaneously driving data latency toward zero.

4. *Tracking custody of data for compliance.* Maintaining the chain of custody for critical data subject to compliance means that you can no longer perform SCD Type 1 or Type 3 processing on dimension tables or fact tables. Check out [Design Tip #74](#) available at www.kimballgroup.com to understand how to solve this problem.

5. *Retrofitting major dimensions for true Type 2 tracking.* Organizations are revisiting earlier

decisions to administer important dimensions, such as customer, with Type 1 (overwrite) processing when changes are reported. This provokes a significant change in the ETL pipeline, as well as a serious commitment to the use of surrogate keys. Surrogate keys, of course, simplify the creation of conformed dimensions that combine data from multiple original sources.

6. *Outsourcing and moving to the cloud.* Outsourcing offers the promise of having someone else handle the administration, backing up, and upgrading of certain applications. Outsourcing may also be combined with a cloud implementation which may be an attractive alternative for storing operational data that is subject to very volatile volume surges. The right cloud implementation can scale up or down on very short notice.

7. *Harvesting light touch data that cannot be loaded into an RDMS.* Finally, a number of organizations that have significant web footprints are capable of collecting tens or hundreds of millions of web page events per day. This data can grow into petabytes (thousands of terabytes) of storage. Often, when these “light touch” web events are first collected as stateless microevents, their useful context is not understood until much later. For instance, if a web visitor is exposed to a product reference, and then days or weeks later actually buys the product, then the original web page event takes on much more significance. The architecture for sweeping up and sessionizing this light touch data often involves MapReduce and Hadoop technology. Check out Wikipedia for more on these new technologies.

I think that this list of what are today disruptive changes to your ETL pipelines are actually a glimpse of the new direction for enterprise data warehousing. Rather than being exotic outliers, these techniques may well become the standard approaches for handling customer oriented operational data.