

Design Tip #89 The Real Time Triage

By Ralph Kimball

Asking business users if they want “real time” delivery of data is a frustrating exercise for the BI system designer. Faced with no constraints, most users will say “that sounds good, go for it!” This kind of response is almost worthless. One is left wondering if the user is responding to a fad.

To avoid this situation we recommend dividing the real time design challenge into three categories, which we will call Daily, Frequently, and Instantaneous. We will use these terms when we talk to end users about their needs, and we will design our data delivery pipelines differently for each of these choices.

Instantaneous means that the data visible on the screen represents the true state of the source transaction system at every instant. When the source system status changes, the screen responds instantly and synchronously. An instantaneous real time system is usually implemented as an EII (Enterprise Information Integration) solution, where the source system itself is responsible for supporting the update of remote user’s screens, and servicing query requests. Obviously such a system must limit the complexity of the query requests because all the processing is done on the legacy application system. EII solutions typically involve no caching of data in the ETL pipeline, since EII solutions by definition have no delays between the source systems and the users’ screens. EII technologies offer reasonable light weight data cleaning and transformation services, but all these capabilities must be executed in software since the data is being continuously piped to the users’ screens. Most EII solutions also allow for a transaction protected write back capability from the users’ screens to the transactional data. In the business requirements interviews with end users, you should carefully assess the need for an instantaneous real time solution, keeping in mind the significant load that such a solution places on the source application, and the inherent volatility of instantaneously updated data. Some situations are ideal candidates for an instantaneous real time solution. Inventory status tracking may be a good example, where the decision maker has the right to commit inventory to a customer that is available in real time.

Frequently means that the data visible on the screen is updated many times per day but is not guaranteed to be the absolute current truth. Most of us are familiar with stock market quote data that is current to within 15 minutes but is not instantaneous. The technology for delivering frequent real time data (as well as the slower daily real time data) is distinctly different from instantaneous real time delivery. Frequently delivered data is usually processed as micro-batches in a conventional ETL architecture. This means that the data undergoes the full gamut of change data capture, extract, staging to file storage in the ETL back room of the data warehouse, cleaning and error checking, conforming to enterprise data standards, assigning of surrogate keys, and possibly a host of other transformations to make the data ready to load into a ROLAP (dimensional) star schema, or an OLAP cube. Almost all of these steps must be omitted or drastically reduced in an EII solution. The big difference between frequently and daily delivered real time data is in the first two steps: change data capture and extract. In order to capture data many times per day from the source system, the data warehouse usually must tap into a high bandwidth communications channel such as message gram traffic between legacy applications, or an accumulating transaction log file, or low level database triggers coming from the transaction system every time something happens. As a designer, the principal challenge of frequently updated real time systems is designing the change data capture and extract parts of the ETL pipeline. If the rest of the ETL system can be run many times per day, then perhaps the design of these following stages can remain batch oriented.

Daily means that the data visible on the screen is valid as of a batch file download or reconciliation from the source system at the end of the previous working day. A few years ago a daily update of the data warehouse was considered aggressive, but in 2007 daily data would be the most conservative choice. There is a lot to recommend daily data! Quite often processes run on the source system at the end of the working day that correct the raw data. When this reconciliation becomes available, that is the signal that the data warehouse can perform a reliable and stable download of the data. If you have this situation, you should explain to the end users what compromises they will experience if they demand instantaneous or frequently updated data. Daily updated data usually involves reading a batch file prepared by the source system, or performing an extract query when a source system readiness flag is set. This, of course, is the simplest extract scenario, because you take your time waiting for the source system to be ready and available. Once you have the data, then the downstream ETL batch processing is similar to that of the frequently updated real time systems, but it only needs to run once per day.

The business requirements gathering step is crucial to the design process. The big decision is whether to go instantaneous, or can you live with frequently or daily. The instantaneous solutions are quite separate from the other two, and you would not like to be told to change horses in midstream. On the other hand, you may be able to gracefully convert a daily real time ETL pipeline to frequently, mostly by altering the first two steps of change data capture and extract. If you would like to study all 34 steps of the ETL pipeline including the ones mentioned in this design tip, please check out the [Data Warehouse ETL Toolkit](#) book and our [ETL in Depth](#) class, both of which are described on our website www.kimballgroup.com.