

Kimball Design Tip #4: Super Fast Change Management Of Complex Customer Dimensions

By Ralph Kimball

Many data warehouse designers have to deal with a difficult customer dimension that is both wide and deep. A customer dimension may have 100 or more descriptive attributes, and may have millions of rows. Sometimes the "customers" are health insurance policy claimants, and other times they are owners of motor vehicles. But the design issues are the same.

Often in these situations the data warehouse receives a complete updated copy of the customer dimension as frequently as once per day. Of course, it would be wonderful if only the "deltas" (the changed records) were delivered to the data warehouse, but more typically the data warehouse has to find the changed records by carefully searching the whole file. This comparison step of each field in each record between yesterday's version and today's version is messy and slow.

Here's a technique that accomplishes this comparison step at blinding speeds and has the added bonus of making your ETL program simpler. The technique relies on a simple CRC code that is computed for each record (not each field) in the incoming customer file. More on CRC's in a moment. Here are the processing steps:

1. Read each record of today's new customer file and compute that record's CRC code.
2. Compare this record's CRC code with the same record's CRC code from yesterday's run, which you saved. You will need to match on the source system's native key (customer ID) to make sure you are comparing the right records.
3. If the CRC codes are the same, you can be sure that the entire 100 fields of the two records exactly match. **YOU DON'T HAVE TO CHECK EACH FIELD.**
4. If the CRC codes differ, you can immediately create a new surrogate customer key and place the updated record in the data warehouse customer dimension. This is a Type 2 slowly changing dimension (SCD). Or, a more elaborate version could search the 100 fields one by one in order to decide what to do. Maybe some of the fields trigger an overwrite of the data warehouse dimension record, which is a Type 1 SCD.

If you have never heard of a CRC code, don't despair. Your ETL programmer knows what it is. CRC stands for Cyclic Redundancy Checksum and it is a mathematical technique for creating a unique code for every distinguishable input. The CRC code can be implemented in Basic or C. Most introductory computer science textbooks describe the CRC algorithm. Also look on the Google search engine (www.google.com) for "CRC code" or "checksum utility".

I would like to hear about any interesting techniques you have developed like this for your dimensional data warehouses. Send me an e-mail describing your technique.