

Design Tip #78 Late Arriving Dimension Rows

By Bob Becker

Your ETL system may need to process late arriving dimension data for a variety of reasons. This design tip discusses the scenario where the entire dimension row routinely arrives late, perhaps well after impacted fact rows have been loaded.

For example, a new employee may be eligible for healthcare insurance coverage beginning with their first day on the job and be issued a valid insurance card with a valid patient ID. However, the employer may not provide detailed enrollment information to their healthcare insurance provider for several weeks; it may take several more weeks before the new employee is entered into the insurer's operational systems. Of course, the new employee may require health care during this time and submit claims using their patient ID. In this case, the insurer's data warehouse ETL system will receive claim fact row input with a valid patient ID that doesn't have an associated row in the patient dimension – yet. The timing of the enrollment business process and patient setup in the insurer's operational systems naturally runs slower than the claims submission business process.

There are several possible ETL reactions to this problem. The ETL system can reject the fact rows to a suspense file that is reworked on a frequent basis to load the fact rows only after complete patient information has been received and the missing dimension row created. The downside to this approach is the claims fact table will not fully represent the insurance company's true financial exposure.

Alternatively, the ETL system can designate a single dimension row with a description such as "patient unknown" and load the impacted fact rows with the foreign key pointing to this common dimension row. After the patient dimension rows have been created, the affected fact rows need to be revisited and updated with the appropriate patient foreign key. This approach requires retaining the incoming claims fact input in a suspense file to help determine which fact rows need to be attached to the new patient rows. This is an appropriate solution when the incoming natural keys in the fact data input aren't reliable or need to be researched/corrected and thus can't be used to create valid, unique dimension rows.

In the case of the healthcare insurer, the fact row data almost certainly contains a valid patient natural key – it just hasn't been reflected in the patient dimension table, but most likely will at some point. In this situation, we prefer an ETL solution that creates and inserts a new row in the dimension table with only the surrogate key and valid natural key as a placeholder for each unique patient. All incoming fact rows will be loaded (rather than put into a suspense state) with the foreign key linking to the placeholder dimension row for each unique patient natural key. Later, as the business activities and operational source systems catch up, the complete dimension row attributes are populated for the patient. At that time, the placeholder row simply becomes the permanent dimension row. No changes are required to the fact rows as the patient foreign keys in the claims facts will not need to change.

Of course, there are complications to be considered. A detailed discussion is outside the scope of this design tip, but briefly, there may be type 2 version changes to the placeholder dimension row that warrant destructive changes to the foreign key in the associated fact rows. Likewise, if it is determined that a placeholder row was not required for some reason, the placeholder row will need to be deleted or expired and affected fact rows updated. As always, if we change data in the data warehouse, we need to assure the organization's audit and compliance requirements are satisfied.