

Ensuring Data Quality in Data Consolidations

Data consolidation is the process of merging data from several databases. Data consolidation projects are a common occurrence in the data integration landscape. They take place regularly when old systems are phased out or combined. They are key to the successful implementation of data warehouses, operational data stores, and information integration hubs; of course, they always follow company mergers and acquisitions.

Database consolidations after corporate mergers are especially troublesome because they are usually unplanned, must be completed in an unreasonably tight timeframe, take place in the midst of the cultural clash of IT departments, and are accompanied by inevitable loss of expertise when key people leave midway through the project.

Data Consolidation Challenges

An old man once rode his Pontiac three miles in the oncoming traffic before being stopped. He was very surprised why everybody was going the wrong way. That is exactly how I feel when involved in a data consolidation project.

In my column in the July 26 issue, I discussed why data conversions cause data quality problems. Consolidation of data from multiple systems poses the same challenges, though magnified to a great extent, and adds a completely new dimension of complexity, for two reasons:

1. The data among consolidated systems often overlaps. There will be duplicates, partial overlaps in subject populations, and data histories. Most importantly, there will be numerous data conflicts. Deciding which data elements to trust is never trivial and inevitably backfires at the data quality.
2. When the data is merged into an existing nonempty database, the target data structure can have few or no changes. Often, however, the new data simply does not fit. Efforts to squeeze square pegs into round holes are painful, even to an outside observer.

What is even more troublesome is that data consolidation specifications are usually built without deep understanding of the actual structure, content, and quality of the data in each source. Comprehensive data profiling and quality assessment are key here, even more so than for simple data conversion. A priori understanding of overlaps and conflicts between the databases allows you to navigate a safe route rather than march into oncoming traffic.

The Traditional Approach

The traditional approach is to setup a winner-loser matrix indicating which source data element is picked up first. For instance, date of birth will be taken from System A if present, from System B otherwise, and from System C if it is missing from both A and B.

This approach rarely works because it assumes that data on System A is always correct – a laughable assumption.

To mitigate the problem, the winner-loser matrix is usually transformed into a complex conditional winner-loser hierarchy. Now we take the date of birth from System A for all males born after 1956 in California, except if that date of birth is January 1, 1970, in which case we take it from System B, unless the record on System B is marked as edited by John Doe who was fired for playing games on the computer while doing data entry, in which case we pull it from Spreadsheet C...

At some point the winner-loser hierarchy is so complex that nobody really understands it. It becomes impossible to manage even for simple indicative data elements, and reaches mind-boggling complexity for historical data stacks, especially event and state-transition histories. An even more serious issue in this model is that we absolutely cannot cleanse the data before conversion, because it is impossible to determine which data elements will really be used. It is time to scrap the approach and start over.

The Correct Approach

The correct approach to data consolidation is to view it in a similar light as data cleansing! We start with a comprehensive set of tests, comparing the data across all sources. We now have a full list of discrepancies. These data inconsistencies are conceptually very similar to the errors found by the data quality rules in the process of data quality assessment. While some of these discrepancies may be legitimate, without proper care they will most likely turn into true data errors after consolidation.

The next step is to analyze the discrepancies and look for patterns. Suppose we conclude that any time values of a certain attribute in Systems A and B coincide, they can be trusted (regardless of the values in database C). We can mark those values as trusted and eliminate the discrepancies from the list. We can also make corrections to the mismatching values on C. For each group of discrepancies, we make an individual decision using a variety of conditions. Every time a decision is made, good data must be marked as trusted and bad data can be corrected. With this technique, we decompose the list of all discrepancies into a set of simple groups and derive a straightforward solution for each group. With every step, we move closer to the ultimate data quality objective.

When the number of data sources is large (more than three), the effort to compare data among all sources increases exponentially. We end up with a rather complex algorithm to pick up data marked as “trusted” from all sources. A simpler (though conceptually identical) technique is to start by selecting for each target data element a primary data source (PDS), which will be used in conversion. Then we design a comprehensive set of tests, comparing the selected primary data sources against other data, and use these additional data sources for “data cleansing.”

In a general case, different databases will serve as PDSs for different data elements. Even for a given data element, different PDSs can be elected for different subject populations

and time periods. Once data in each PDS is validated and cleansed, the data consolidation is performed by data conversion algorithms from various PDSs.

Since data consolidation is performed piecemeal from multiple primary data sources in multiple source databases, it may be difficult to ensure consistency of all the data at their sources. This is especially true when historical data is consolidated from various PDSs for different time periods. In this case, it is critical to assess consistency of the overlapping data. The solution is to implement additional data quality rules either at the “connection points” or at the target database after consolidation.

Summary

Data consolidation is the process of combining data from several databases. The traditional winner-loser approach typically fails to achieve acceptable data quality levels, making data consolidation one of the main causes of data problems. The correct attitude is to view data consolidation in much the same way as we view data cleansing. The method is certainly not simple, especially for complex state-dependent and time-dependent data. However, it ensures success.