Print Article Close Window

From: www.cio.com

How to Solve CRM Data Deduplication Dilemmas

– David Taber, CIO

October 01, 2013

Unless you have an explicit data deduplication strategy in place, your CRM system is almost certain to have some level of duplicate records. There are a few reasons why:

- Humans may not search for records before adding new contacts, leads or accounts. Even if your CRM system has some sort of duplicate-alert system, not everyone will pay attention to it — and it may not work on mobile devices anyway.
- Data import tools may not successfully identify duplicate records or may import them anyway, though some use a "potential duplicate" flag.
- Integrations with outside sources such as website registration forms, partner portals, message brokers or other applications may not query the CRM data before inserting new records. In some situations, even if the external source detects a duplicate, it can't update the existing record and must create a new one.
- Several types of administrative errors and software bugs — both inside the CRM system and in associated applications — can rapidly produce thousands of duplicate records.

There's nothing wrong with having 2 percent duplicate records, as long as they are short-lived and your tools and processes detect and correct them. Get beyond 5 percent or so, though, and users will start to complain. Reports will become misleading. Data updates will get lost, since users won't be able to find the changes they made just the other day. System credibility will start to plummet. Systems with 25 percent duplicate records, moreover, can threaten careers.

Detect, Correct Data Duplicates Takes Time

As with any type of data pollution, the correction cycle will take a while. You need to develop a methodical get-well plan with expectations set correctly for budget and schedule.

Start by using the best available duplicate detection tools that warn users while they enter data. Before you deploy any tool, test that it doesn't throw such severe errors that your integrations with external systems are blocked.

Next, analyze the systemic sources of duplicate records. Narrow them down to one or two external systems, and get part of the team working to cut down the inflow of new duplicate records. Rinse, lather and repeat as necessary.

Commentary: Who Owns CRM Data at Your Company?
Also: Why the Last Mile of CRM Implementation Is the Hardest

While that's going on, analyze the heuristics of the duplication. Hopefully only one or two tables are involved, but remember that each table will have its own (set of) patterns. Since data deduplication requires an iterative approach, you'll want to get the team members familiar with the tools and processes on the lowest-risk tables. "Lowest risk" varies a lot, but leads or activities are typically the best candidates, as they tend to have the fewest number of other records pointing to them.

In analyzing the heuristics of duplicates, you need to understand four things:

- The most reliable way to detect a potential duplicate "pair;"
- The best way to identify the "winner" in the merge cycle;
- What parts of the "loser" record need to be preserved, and
- How to deal with pointers to and from the loser record.

Most data deduplication tools these days operate on the data while it's in your database. This is typically less intrusive than a full export/clean/import cycle (although I'm sure I'll get several angry emails from vendors that use that approach). Even so, deduping is hardly zero impact. In fact, it's a major pain if done incorrectly, because there's no "undo." Extra care is mandatory.

CRM Data Deduplication Requires Methodical Approach

Basically, the data deduplication cycle looks like this:

- Do a full system backup. (Yes, every time.)
- Do your work in the system sandbox first, if you can. Once you've validated the approach, the tools and the results, you'll have to do the entire cycle again in the production instance.
- Keep a thorough log of every step (including when you did it) so that you can precisely repeat it (and troubleshoot in case things go wrong).
- Normalize all the records of the table in question, particularly for things such as state and country codes, picklist values and other items where clean field data will improve the quality of duplicate detection and winner/loser determination.
- Identify fields that you'll want to preserve both the "winner" and "loser" values, such as phone numbers, email, stage/status, owner and record type. Both records might be correct, even though they are different values. Once you've identified these vulnerable and valuable fields, concatenate them together and put them in a new text field on for each record in the table. The concatenation can be done with an ETL tool or with code inside the system. Just make sure this step is complete and correct before you move on.

Related: Why Location Is a Growing Issue for CRM Systems
Also: When Your CRM System Passes 1 Million Records

- Set up your data deduplication tool with the fields and matching criteria that best identify potential duplicate records. In the first iteration, you want the criteria to be very tight. Run this part of the tool and see how complete are the duplicate candidates.
- Set up the merge rules and processing scenario that will identify the "winners" within each set of duplicates. Make sure the text field you created (with those concatenations) will be merged with "append mode" so that the loser's data is preserved. Review the merge results with users to identify any gotchas.
- Perform a test merge with these tight criteria. See if there are any unintended consequences. You may need to wait a day to make sure that daily batch updates don't have problems with the merged records or re-create the duplicates in the course of external-system synchronization.
- When you're satisfied with this pass in the sandbox, run through the process in production. Do it once. (Yes, once.) Again, look for side effects after one day.
- Assuming all goes well in production, back up the entire system again.
- Repeat this process with looser criteria. Typically, each deduplication cycle will process about half the total duplicates. You may need to run through the entire process four times to get duplication down to an acceptable level.

High-Risk Data Will Cause the Most Dedupe Headaches

Typically, high-risk records have more things pointing to them. If that record disappears, there are more orphaned references. In most CRM systems, the account is by definition the highest-risk table — and several outside systems likely point to its records. You may determine that accounts really can't be merged. Then what?

Most CRM systems have a concept of a parent account. That can be the basis of a solution strategy for accounts that can't be merged. The duplicate accounts all become children of a new master account, which acts as a "holding company" of sorts for roll-ups.

No doubt about it, duplicate records in CRM systems are the most popular form of data corruption. Unfortunately, data deduplication isn't an event. It's a process. Even if you get them all out of the system now, there's going to be a software or process change that will cause a new pattern of duplicates to be created. Since it will be easier to clean them up if there's only one error pattern, a thorough duplicate detection cycle will be needed on a monthly basis.

David Taber is the author of the Prentice Hall book, "Salesforce.com Secrets of Success" and is the CEO of SalesLogistix, a certified Salesforce.com consultancy focused on business process improvement through use of CRM systems. SalesLogistix clients are in North America, Europe, Israel and India. Taber has more than 25 years of experience in high tech, including 10 years at the VP level or above.

Follow everything from CIO.com on Twitter @CIOonline, Facebook, Google + and LinkedIn.