

The perils of dirty data

How important is data cleansing and validation? Read these tales of horror, and beware

By Dan Tynan

October 29, 2007

Few IT projects are more frightening than data integration and reconciliation. Actually, let us rephrase that. One thing is more frightening -- when data integration goes bad.

Sometimes it's a problem of starting out with bad data, through user error or even deliberate sabotage. Sometimes the data starts out good but gets lost, truncated, or altered when it moves from one system or database to another. Your data may go stale, or it may become collateral damage in a turf war inside your organization -- everyone clinging to their own little piece of the data store, nobody willing to share. The task certainly isn't helped by the overwhelming volume of data companies generate each day.

Data projects can go bad in many ways. Here are five of the most common: what went wrong, what happened as a result, and what you can do to avoid having the same thing happen to you. The names of the companies involved have been obscured to protect the guilty. Don't let your own project become someone else's horror story.

1. The "Dear Idiot" letter

Be careful where you get your data -- it may come back to haunt you. This tale of terror comes from the customer call center of a large financial services institution. As in nearly all help desks, service reps take calls and enter customer information into a shared database.

This particular database had a salutation field that was editable. Instead of being constrained to Mr., Ms., Dr., etc., the field could accept 20 or 30 characters of whatever the rep typed. As service reps listened to the complaints of angry customers, some of them began adding their own, not entirely kind, notes to each record, like, "what an idiot this customer is."

This went on for years. No one noticed because no other system in the organization pulled data from that salutation field. Then, one day, the marketing department decided to launch a direct mail campaign to promote a new product. They came up with a brilliant idea. Instead of purchasing a list, why not use the service desk database?

So the letters went out: "Dear Idiot Customer John Smith."

Strangely, no customers signed up for the new service. It wasn't until the organization began examining its outgoing mail that it figured out why. The moral of this story?

"We don't own our data any more," says Arvind Parthasarathi, vice president of product management and data quality for data integration specialists Informatica. "The world is so interconnected that it's likely someone will pick up your information and use it in a way you never anticipated. Because you're pulling data from everywhere, you need to make sure you have the right level of data quality management before you use it for anything new."

What constitutes the "right level" will vary depending on how you use the data. "In the direct mail industry, getting 70 to 80 percent of your data correct is probably good enough," he adds. "In the pharmaceutical industry, you want to be at 99 percent or better. But no company really wants, needs, or will pay for perfect data; it's just too expensive. The issue always is, how will it be used and at what point is it good enough?"



2. Dead men cast no votes

Data cleansing can be a matter of life and death -- literally. PR specialist Nancy Kirk was volunteering in the congressional elections of 2006, calling registered voters to get them to the polls, when she noticed something odd: Three out of ten voters she dialed were deceased and thus ineligible to vote (except in certain precincts in Chicago).

The problem of having data that is literally dead is not uncommon in the commercial world, and it has real consequences for the living.

Jim Keyser, president of The Keane Organization's Investor Retention and Communication Solutions division, has spent the past year rolling out an investor data quality program for Keane's clients, which include major insurance companies, mutual funds, and Fortune 500 firms.

Keyser says they often find 8 to 15 percent of clients' data records contain anomalies such as mistyped Social Security numbers or outdated addresses. But about one in five of those anomalies is a shareholder who's been dead for more than five years. In one case, a client had an "active" account for a shareholder who last drew breath more than 72 years ago.

"This isn't client negligence, it's just a naturally occurring problem," Keyser says. Private companies go public, change names, get acquired, or spun off, and their shareholder data follows along, often for decades.

But the consequences can be greater than just money wasted on unnecessary mail. The biggest concerns are fraud and identity theft. Some stranger could be cashing the late shareholder's dividend checks, the rightful heirs could be denied their inheritance, or confidential company info could leak out.

The solution? Software such as Keane's Score application can identify data anomalies across different systems and flag them for review. But all companies must exercise due diligence, have good internal controls, and scrutinize their data on a regular basis, says Keyser.

"Virtually every business has this problem to some degree," he says. "From a risk management point of view, the best practice is to make sure you're keeping it in check. Understanding how this natural phenomenon impacts you is a good first step."

3. Duped by duplicates

User error is bad. User ingenuity can be worse. Take the case of the major insurance carrier that kept most of its customer data within a mainframe application from the 1970s. Data entry operators were instructed to first search the database for existing records before entering new ones, but the search function was so slow and inaccurate that most operators gave up and entered the records from scratch.

The result? Individual companies ended up in the database 700 or 800 times, making the system even slower and less accurate.

Unfortunately, the application was so deeply embedded in the company's other systems that management was reluctant to spend the money to rip and replace. Finally, the carrier's IT department made the business case that the company's aging data app would ultimately prevent it from being able to add new customers, costing it \$750,000 a day in new premiums.

At that point, the company used SSA-Name3 by Identity Systems to clean the data, ultimately weeding out 36,000 duplicate records.

Dupes are one of those problems that keep IT managers up at night. The larger your database, the worse the problem usually is, says Ramesh Menon, a director at Identity Systems, which provides identity searching and matching software for organizations such as AT&T, FedEx, and the Internal Revenue Service.

Unfortunately, nobody knows how big their problem is, he says. "If anybody tells you 'I have exactly 2.7 percent duplicates in my customer database,' they are wrong."

There's no magic bullet, either. Menon says the solution lies in using data matching technology to isolate "the golden record," a singular view of information across multiple data repositories. Even then, the hardest part may be getting all the vested parties in an organization to agree on what data they're willing to share, as well as what constitutes a match.

"Two different sections of the same organization may have completely different definitions of what a match or duplicate contact is," he says. "These kinds of integrations fall apart because people can't agree about who owns the data or what information can be exchanged with others."

4. When data decays

Remember text-based adventure games such as Zork? Apparently, somebody somewhere is still making these things. Worse, they're using data that's equally ancient.

MailChimp co-founder Ben Chestnut tells the story of an old-school games developer that used MailChimp's e-marketing service to contact 10,000 previous customers, alerting them that he'd finally finished version two. Most of the addresses were at least 10 years old -- some of them Hotmail accounts discarded so long ago that Microsoft was using the addresses as spam traps. Within a day, all MailChimp e-mail was blacklisted by Hotmail's spam filter.

Fortunately for MailChimp, the developer had kept pristine records, down to the IP address each customer had used to download his games. That's what saved them, says Chestnut. "We fired off a quick note to Hotmail's abuse desk -- proved they were legitimate customers, just old. The next day we got delisted. That's pretty rare."

All data ages quickly, but contact data ages faster than most.

"You have to make the assumption that data decays like a radioactive sample," says Informatica's Parthasarathi. "You have to go into every system and periodically update it."

Jigsaw.com, an online contacts database geared toward sales professionals, takes a Wiki-style approach to data cleansing. Its 335,000 members get points for uploading their own contacts to Jigsaw and correcting others. Every record must be complete, and if Jigsaw users enter information that's incorrect or old, they lose points. Members spend their points by buying information for people they want to reach.

Jigsaw CEO Jim Fowler says an Atlanta-based technology company recently asked his firm to compare its contacts databases to Jigsaw's and weed out the bad data.

"They had 40,000 records," he says. "Only 65 percent of them were current and 100 percent were incomplete. We're finding that most of our corporate customers have sets of data so cruddy no one can match to them. Corporations spend millions on CRM, and it's amazing how bad that data is."

The real value is not the data itself, but the ability to keep up with how quickly it changes.

"The power of Jigsaw is complete data and self-cleansing," says Fowler. "If our self-correcting mechanisms don't work, we're just another crappy data company."

5. The war on error

The difference between good data and bad can be as small as a single dot. Penny Quirk, principal consulting manager at Robbins-Gioia's Records and Information Optimization Practice Area, says she once consulted on a major data integration project where everything seemed to go fine. Six months later someone opened a data table and found rows of symbols but no data.

"It was a character coding error," says Quirk. "They used ellipses in some fields, and wherever someone had entered two dots instead of three it triggered the whole line of data to go corrupt."

The company had to painstakingly re-create the database from a backup, searching for the ellipses, then replacing them with the actual data.

More often than not, the problem is more than mere data entry errors or garbage in/garbage out. Most organizations fail to adequately plan when moving data between different operating systems or upgrading from older versions of SQL, says Quirk. They'll do it too quickly, using whatever resources are available now with the hope of cleaning it up later. (A bad idea, she adds.) Worse, their test environments and production environments may not match, or they may test using a small subset of data, only to have big problems arise later with the data they didn't test.

"Organizations making dramatic changes in technology without putting forth the necessary time and effort to manage the data reconciliation, integration, and conversions can become victims of bad data," Quirk says. "As data is moved from one source to another, the number of chances for it to become bad is astronomical."

Quirk's advice? Don't expect IT departments to validate your data set. Get the power users who work with the data to help plan and test the integration. Before you decide to consolidate, look at all your data fields and identify the applications that may be pulling data from them. When possible, test with all your data, not just a subset because even the tiniest errors can send you and your data into a world of pain.

One final horror story illustrates just how big a small error can become.

Peter Teuten, president and CTO of Keane Business Risk Management Solutions, tells of a client that created an application to determine whether corrupt files were circulating in their systems. If the amount of corrupted data exceeded a certain threshold, the company would know to implement data protection processes.

The problem? They accidentally inverted the rule set for the data protection system; the more corrupt data it found, the better their systems appeared in the reports.

"The network was eventually infiltrated by a worm, which corrupted their files," says Teuten. "They had to rebuild most of them from scratch, which cost them millions of dollars. All from a very simple configuration and management error -- two numbers were reversed."

If that doesn't scare you into approaching your next data management project with caution, nothing will.