Operation Clean Data

CIO Magazine, 2004

Malcolm Wheatley

Cleaning dirty data is not just a matter of mastering the technical challenges. It requires making sure your staff is working closely with the business every step of the way.

In the early hours of March 20, 2003, British soldiers, sailors and airmen joined US forces in the invasion of Iraq and the toppling of Saddam Hussein. Thus far, they have played a vital role in rebuilding Basra and the critical Persian Gulf port of Umm Qasr. Massive shipments of military materiel were essential to their success, and basically, anything that wasn't a vehicle, live ammunition or fresh provisions (which have different supply lines) began its journey to the Gulf from England's military warehouses. In the few weeks prior to the invasion of Iraq, these depots sent by ship or air 3169 6-metre shipping containers to the Gulf, along with almost 22,000 1-metre pallets.

Getting these shipments to the Gulf was a logistical nightmare that would have been far more fraught had the British defence ministry not embarked four years ago on a £6 million effort to pull together three separate supply chains: This involved reconciling some 850 different information systems, and integrating three inventory management systems and 15 remote systems.

The biggest foe in this massive integration effort was not Saddam Hussein, but dirty or disparate data. To one system, stock number 99 000 1111 was a 24-hour, cold-climate ration pack. To another system, the same number referred to an electronic radio valve. And if hungry troops were sent radio valves instead of rations, the invasion and rebuilding of Iraq wouldn't have gone very far.

Dirty data has long been a CIO's bugbear. But in today's wired world, the costs and consequences of inaccurate information are rising exponentially. Muddled mailing lists are one thing, missing military materiel quite another. Throw in the complications arising from merging different data sets, as in the aftermath of a merger or acquisition, and the difficulties of data cleansing multiply. For this article, we interviewed seasoned data-cleaning veterans from organizations as diverse as the British Ministry of Defence, the US Census Bureau and Cendant, a real estate and hospitality conglomerate. But the lessons learned contain two common themes: How to surmount the technical challenges of cleaning data, and how to align IT staff with the business side to ensure that the task gets done right.

Know Your Enemy

When Britain's defence department began its data-cleaning project in early 2000, it faced a huge task, says Lieutenant Colonel Andrew Law, head of The Cleansing Project. (It just so happens that the acronym TCP is also a well-known British brand of antiseptic.) The department's IT team was using three main systems to sort through 1.7 million records, which each had literally hundreds of attributes. Each record referred to an item that troops might require, and many of these items were to be dispatched from the ministry's widely dispersed warehouses in Bicester, England, and other locations. (The Bicester warehouses are far apart because they were built in 1942 with the idea to make it hard for German bombers to deliver a knockout punch.)

Law's mission was to review all the data, but he had to concentrate his team's energies on cleaning six critical data fields: the NATO item identifier, the NATO supply classification, the unit of issue, the supplier code, the packaging code and the hazard code. These six fields were chosen based on which ones would have the biggest impact on the supply chain if they were wrong.

"The first step was to identify homonyms and synonyms," says Paul Nettle, manager of data cleaning for TCP. Homonyms, he explains, are two or more different items with the same identifier, such as rations and radio valves. Synonyms are the same items with more than one identifier - the same radio valve kept in two places in a warehouse under two different numbers, for example.

"Synonyms are merely inefficient," Nettle observes. Overstocking and overbuying result from such data mistakes, rather than troops being shipped the wrong gear.

Next, the IT team employed data-profiling software to crawl though the data, checking it for valid NATO numbers. The troubling finding: 119,000 numbers (about one in 10) weren't valid. The radio valve, it turned out, was a valid NATO part number, but the rations came from a satellite system where non-standard rules had been used. Every one of them had to be sent to a NATO office in Glasgow for codification, and then corrected in each system in which it occurred. Nettle and his team also discovered they had quite a bit of relabeling to do at the depot, since much of the inventory sitting on the shelves was now incorrectly labelled.

The next step was "fuzzy matching", using software to look for duplicates and errors introduced by keyboard entry. "The ability to ignore [minor mistakes in] punctuation and figure out when a 3 had been erroneously substituted for an 8 was important when dealing," Nettle says. Such numerical errors, after all, could change the entire meaning of the text, while punctuation mistakes merely provided Nettle's team with much needed amusement.

By August 2001, they had completed the relatively easy (if time-consuming) task of examining item identifiers to see, for instance, if an item held the valid NATO number. Now they had to find a way to correct the other data fields. Here, the challenge was more difficult. For things such as unit-of-issue labels, packaging codes and supplier details, hard and fast rules to tell clean data from dirty data didn't exist. For example, supplies of aircraft oil: A military unit in the Gulf might order 250 litres of oil, expecting 250 one-litre cans - only to receive 250 separate 250-litre

drums of the stuff. The reason? On the Royal Air Force system responsible for ordering the oil, 250-litre drums, not one-litre cans, were the unit of issue. Neither label was technically an error, but clearly, such inconsistencies could quickly cripple a supply chain. To make sure such a disaster would never occur, the TCP team turned to a data-profiling tool, which highlighted errors and inconsistencies in the various codes. The software provides easy-to-understand, computer-generated diagrams to spot unusual data formats that could be erroneous.

As Law points out, however, technology only goes so far. The 12-man TCP project succeeded in large part because team members at headquarters worked closely with members in the British Army, Royal Navy and Royal Air Force, who made sure flawed data actually did get cleaned and organized activities such as relabeling inventory on the shelf. So far, Nettle says, the cleansing project has cost £6 million over four years, and has saved the Ministry of Defence £50 million.

Define the Rules in Advance

While the British defence ministry is still cleaning up its data warehouses to generate a more consistent view of military supply items, commercial companies are employing much the same technology to develop an enterprise-wide view of their customers. In some cases, the demand for consistently clean data comes from the customers themselves, who want to see how their business is performing in real time. For example, in these economically constrained times, the corporate customers at Carlson Wagonlit Travel, one of the largest travel agencies in the world, are eager for good quality data on exactly how their travel and expenses budgets are being spent. Indeed, building a data warehouse that can deliver such information has become a competitive differentiator in the industry, says Jay Vetsch, senior director of information delivery at Carlson.

The task for Vetsch and his team was daunting. With annual sales of $US10.5 billion and operations spread over 140 countries, the agency has high data volumes: 14 million airline tickets per year, 12 million hotel nights booked every year and so on. While the raw number of transactions per day (around 60,000) is doable, each record often equates to a trip with several flights, hotels and rental car reservations. Thus, the record size is massive, around 400 fields.

Worse, the data must be extracted from a number of different back-office systems spread across the business. What's more, the data is subject to the inputting vagaries of the front-office operators in those 140 countries - not just human vagaries, but also differences in legal, tax and accounting regulations. And from the point of view of the people generating the data, Vetsch's task is not mission-critical.

"You have to remember that the information is being generated for the purpose of getting a traveler a ticket - not for an MIS system to provide reports to clients," he says.

As a result, the data can contain errors - an invalid supplier code, client code or a fare discrepancy - not major enough to prevent a ticket from being issued, but flawed enough to foul up an analysis. Vetsch relies on software that acts as a gate guardian to the data warehouse. If a record meets defined data quality criteria, it's allowed to proceed. If it doesn't, it's kicked back to the originating office for correction.

Data from Europe, where the company has offices in most countries, is already being used on a limited basis to generate client reports. Company agents in North America and the remainder of the countries in which Carlson has offices should be able to generate such reports by early 2005. Vetsch declined to disclose the projected ROI for the cleansing effort. However, if good quality client reports have become the price of getting corporate business, then it's a bold manager who'd argue that the investment was nothing other than the price of survival.

Buy-In from Owners of the Data

Similar to Carlson, Cendant - owner of car rental company Avis and realtor Century 21 as well as hotel chains Days Inn, Howard Johnson, Ramada and Travelodge - would love a single, enterprise-wide view of all its customers. But five years work on building a data warehouse delivered virtually nothing. That's because no one was using it. By now you can guess the culprit: dirty data.

"Basically, the data warehouse was being used for list generation by two people in marketing," says Vincent Kellett, senior director of data services who was hired in 2002 to see if the project could be revived. "Because of data quality issues, the project was dying on the vine."

To make the system viable, Kellett realized the company would have to throw out a bunch of hard-to-maintain custom code, spend money on cleaning up some truly horrible data and institute formal processes for data maintenance. Even basic procedures such as subscribing to the national change-of-address database maintained by the US Postal Service had been overlooked by the project team. "They'd been so mired down in day-to-day problems that they just hadn't got round to it," he says.

Data-cleaning software from Trillium Software was pressed into service. The database originally contained 132 million records, a number that was eventually boiled down to 90 million "that at least had a name and a street address", Kellett says. At each cycle of the data-cleaning process, his team formulated new rules, which were then subjected to a trial experiment to both detect duplicates and correct them. Further winnowing, by matching against the latest information on address changes, eventually reduced the number to closer to 80 million cleaned records that were loaded into the data warehouse.

When a customer checks into a hotel or picks up a rental car and a new record is created, the system asks: Do we know this person? If so, load any new information - such as change of address or phone number - and then update their transaction data with another stay or car rental. And that information is automatically integrated with the rest of the customer database.

Key to the project's completion was a decision to closely involve the business owners of the data (in this case, the marketing department) in developing the data-cleaning standards. "I'd advise [anyone working on data quality] to work closely with the business users to define the matching rule," Kellett says. "What constitutes a match? Last name, or last name and first name? Or these, plus a matching credit card? And when a duplicate is detected, what rules determine which record will be the survivor record? For instance, is Bob Smith the same as Robert Smith? And is

the new address revealed by a car rental due to a house move, or the acquisition of a summer cottage? Or just the wrong address completely?"

The turnaround in the project's fortunes has been so complete, Kellett says, that Cendant has been able to launch a loyalty program across nine of its chains - including Days Inn, Howard Johnson, Ramada, Super 8 Motel and Travelodge. Customers can now collect points (much like frequent flier miles) every time they stay in a Cendant hotel. Such a program would have been impossible without the single customer view that the cleaned-up data warehouse provides.

Editing Out Inaccuracies

Even better than cleaning dirty data is making sure it can't be soiled in the first place. Organizations heavily reliant on accurate information, such as the US Census Bureau, are leading the charge when it comes to building real-time validation into data as it is generated. The bureau undertakes hundreds of surveys a year into demographics, the economy, trade data and much else. And needless to say, clean data is imperative.

To facilitate its work, the bureau has developed an approach of building feedback and validation loops into each survey and questionnaire in order to make sure that human-generated information is as accurate and reasonable as possible, says Richard Swartz, associate director for IT and CIO at the Census Bureau. Whenever the completed questionnaires are returned from businesses and individuals, and scanned into the bureau's computers, checks called "edits" take place that test the responses to make sure they are complete and reasonable, Swartz explains. Are the required fields complete? If not, how should nonresponses be dealt with? - should records be ignored, or should responses be "created" by estimating or putting in an average value so as to avoid throwing out a whole record just because of one odd or missing data item? Are responses reasonable? Can a 96-year-old describe herself as unemployed, and is that 80-year-old man really the father of a new baby? Is data consistent? Could a company with three people on the payroll really have a salary bill of more than $1 million?

At the US Centres for Disease Control and Prevention, such real-time data validation underpins data gathering, according to CIO Jim Seligman. When laptop-wielding field workers quiz 40,000 US households a year for the "National Health Interview Survey", automatic edits make sure that responses are as complete as possible while the survey is taking place. Some edits are "skip patterns", designed to prevent erroneous questions from being asked in the first place. If the respondent is male, for example, he won't get the question about mammographies. Other edits are consistency checks: Respondents are asked their age, but also their date of birth - and the two are compared.

It may sound trivial, but from such small foundations, clean data is built. "Any time a human being has something to do with entering data, there's the potential for error - whether it's misreading something, misinterpreting something or miskeying something," Seligman says. And very often, it takes humans working with machines to clean up the mess.

Four Ways to Make Your Data Sparkle

- Prioritize the task. Cleaning data can be costly and time-consuming, so your first step should be figuring out which data is mission-critical and which isn't. For some companies, it's not worth cleaning data errors like sloppy punctuation when they don't get in the way of business objectives.
- Involve the data owners. Ask the business units that own the data for help defining precise rules for what constitutes dirty data. That includes figuring out in advance whether 98 percent clean is good enough, or whether 100 percent is required or affordable.
- Keep future data clean. Put processes and technologies in place that check every zip code and every area code.
- Align your staff with business. Make sure you have IT people working on the ground with business units to make necessary changes in the data and relabel wrongly tagged inventory.