CIO

Print Article ☒ Close Window

From: www.cio.com

Strategies for Pruning Data in the Cloud

– David Taber, CIO

November 09, 2011

Year after year, the cost of disk space has plummeted. Since you can pick up a terabyte for $50, it's often seemed a false economy to be careful with storage.

But in the clouds, the rules are different. If you've got too much low-value data or too many copies of files, it can cost you in two ways. First are the monthly storage charges, and second is the inevitable performance hit when it comes to searches, views, reports, and dashboard updates. In the clouds, it really pays to prune you data set.

The first order of business is assessing the problem: is it documents, or table data? These typically have different storage limits, and the strategies and tools used for pruning are quite different.

Documents typically serve as attachments to records (such as a PDF of a signed contract, pinned to the relevant opportunity), so users may not be able to find them easily. Consequently, the same document may have been attached to three or four different records. You also need to look for cases where people have attached every version of a rapidly-changing document. The first thing to do is export an inventory of every document in the system (including the record IDs they are attached to, plus their last update date) and look for possible duplicates using spreadsheet filters. There are duplicate file detection tools that can do a much better job (by inspecting the contents of the files), but I don't know of any of these file tools that work directly in cloud applications. Unless you are willing to download all the file contents onto your own servers for that deep analysis, you're going to have to live with metadata analysis to identify which files to prune. Since optical storage is cheap, you might as well archive all the files you delete from the cloud, in case somebody complains later on.

Table data is a very different story, with many system-specific tricks and techniques for different kinds of clouds. That said, here's the general workflow:

• Identify which of your cloud systems really have a storage problem. Some systems (e.g., accounting) really can't be pruned very much because they need to be auditable and must hold all the details over long periods. Other systems (e.g., marketing automation or log analytics) rapidly collect enormous amounts of detail that can really slow the system down.

- Identify which tables are consuming more than 20 percent of your total storage. Focus there.

- For each table, understand the value of the individual records. Some tables (particularly accounts or contracts) are almost inviolate because of what they represent and the impact of record removal (particularly when these tables are integrated with outside systems). Other tables, such as "anonymous leads" in a marketing automation system, can be pruned with abandon.

- Before you go any further, do a complete backup of all your cloud's data onto either disk or optical media. I cannot say it any more clearly: this is NOT optional.

- For tables that can be freely pruned, look for the "signal to noise ratio." Is there some time horizon beyond which the information doesn’t matter at all? For example, in a marketing automation or web monitoring cloud, do we really care about anonymous visitors who haven’t returned in 6 months? Is it OK to remove all Leads with a score of less than zero? Make sure you get buy-in from all the affected user groups first, but signal- to-noise based pruning can get rid of millions of records in a hurry.

- Some tables have decent signal-to-noise ratios, but the amount of detail stored just isn't worth it over time. For example, many marketing automation and e-mail blasting systems use the activity table to record important e-mail and Web interactions. These activity tables can represent half of the system's storage. But how much will it matter a year from now whether a person watched video A today versus video B yesterday? Use this litmus test: if a particular detail will not actually change anyone's decision or behavior, it's not "information" any longer. For these situations, we recommend a compression approach: keep the information, but remove most of the details after 6 months or so. The histories are typically stored as custom tables, represented by tallies, token strings, or even bitmaps with tiny storage requirements. This strategy will require some careful thinking, user input, and custom code development, but can provide continuous pruning based on information value.

- Some tables (particularly leads and contacts) can collect duplicates in a hurry, particularly if your firm has process problems in lead collection and handling. If your cloud system has deduping tools (from the main vendor or third parties), buy a good one and really learn it. The best tools have fuzzy-logic algorithms that let you find and merge duplicates without moving the data out of the cloud. The merging process preserves as much of the data as possible, but if you have a lot of data collisions (e.g., two different mobile phone numbers for the same person), you may need to create shadow fields and pre-populate them with divergent data prior to the merge. For a number of reasons, data merging must be done in phases: it takes a lot of CPU time, as well as your think-time, to get rid of 100,000 dupes. Do not rush it, as there is no undo for a merge.

Most of the above is a one-time fix, rather than a process change. If you aren't willing to invest in enhancing your data management processes, you may need to revisit these issues on a quarterly basis. Pretty much forever.

David Taber is the author of the new 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, and David has over 25 years experience in high tech, including 10 years at the VP level or above.

Follow everything from CIO.com on Twitter @CIOonline.