



Data Lakes: Lessons From the Junk Closet

By Drew Ramm

*How that special closet in
your home can prevent you
from creating chaos in your
data storage solution.*





everyone has *that* closet in their home. The one that started out with the best of intentions, meticulously organized with future needs considered, but quickly turned into what can be best described as pure chaos. It happens to everyone. And it happened to my wife and me as we packed in preparation for moving to a new home.

As we started to go through that one closet, I quickly realized it had succumbed to the same fate as many data lakes implemented by companies: It had become a swamp.

If you have had conversations about business intelligence and analytics in the past year, you have likely heard the term “data lake.” A data lake is supposed to be a single central repository where the enterprise can store and analyze every kind of data relevant to its business—data stored in its native format.

Rather than a traditional data warehouse, which is highly structured and takes significant time to build and integrate new data sets, a data lake offers fast, iterative analytic experimentation with a variety of data sets.

While there is no secret formula to avoiding data swamps, my experience with our junk closet offers four valuable lessons about data lakes.

The potential to store huge amounts of data and use it to uncover deep correlations has led many organizations to implement data lakes of their own. And, much like your closet at home, they start out with the best of intentions.

As my wife and I packed the contents of our closet, I recalled when we first moved in together. After we finished putting our belongings away, we looked at one remaining empty closet and realized we now had the space to keep some things we previously could not. Just as the cost of data storage has become cheaper and cheaper, the cost of storing things in our condo had as well.

So, we got to work. We installed new shelving, bought and labeled several storage containers, and added a filing system—complete with labels—to organize everything. Everything that could go into the closet had a specific place to go. It was the perfect solution—or so we thought.

Over time, our closet went from a well-organized area where we could quickly retrieve anything we needed to a dumping ground where our belongings went to die. What started out as a place for important documents, sporting equipment, and office supplies became a place where we stored my bacon costume from Halloween 2015, never to be seen again.

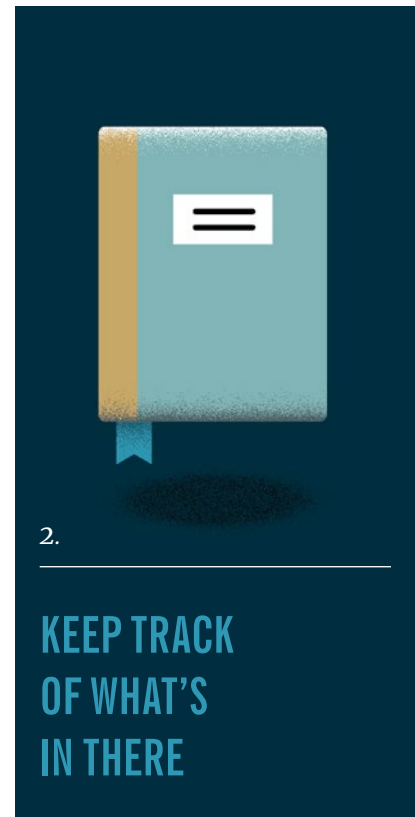
You may ask: What in the world does a bacon costume in a closet have to do with data lakes? Just like our closet, many of the data lakes implemented in the past two years have deteriorated into what are now data swamps—massive repositories of data that are essentially inaccessible to end users. While there is no secret formula to avoiding data swamps, my experience with our junk closet offers four valuable lessons about data lakes.



The second we had an empty closet for extra storage, we completely shifted our thinking. Instead of questioning whether that old computer cable had any use, we just added it to the closet—with no regard for whether we had any use for the cable.

Limiting the data you store in your data lake may seem counterintuitive. After all, isn't that what a data lake is for? And it's true: One of the benefits often touted for a data lake is the ability to store anything until you need to find value in it.

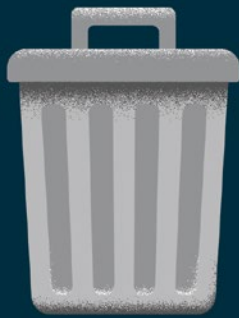
Storage is cheap, right? Not always. Just because the cost of storing data seems cheap doesn't necessarily make it cheap to use. Finding and accessing the data can make it even more difficult and expensive to use.



The early days of our closet—with well-defined and easily accessible folders—was short-lived. We quickly defaulted to putting everything into an “other” folder, which soon became the only folder we used.

Various solutions have emerged that allow organizations to clearly understand what is available in their data lakes. Not only do they provide key metadata—such as where the data came from and how old it is—they also provide insights into the data quality.

There are even collaborative tools that act almost like Yelp, allowing users to vote on the data and add comments. Keeping an updated catalog of data sources is critical to a successful data lake.



3.

DEFINE RETENTION POLICIES

Going through our closet was like going back in a time machine. We were amazed by how many BlackBerry and old iPhone cables we had amassed over the years. Guess how many of those cables have any value to us? That's right: zero.

Just like old cables, data loses its value over time. Some, such as data from "internet of things" sensors, can be prohibitively expensive to store. It's important to not only define data retention policies, but also to continually evaluate and define them as regulations and technology options continue to change.



4.

IMPLEMENT STRONG SECURITY

The final and most important lesson we learned from our closet was related to security. Since we moved right before the holidays, I went ahead and purchased a Christmas gift for my wife and hid it—where else?—in our closet. I left it there unwrapped, open for any eyes, specifically hers, to see.

For data lakes, it is important to have data usage policies that outline who can access, distribute, change, or delete data in the lake. Otherwise, you risk your data finding its way into the wrong hands.

All these lessons have one thing in common: They are not about technology.

Companies that want to succeed in analytics also need to invest in organizational structures and processes. Data lakes are a promising source for companies to tap and extract actionable data.

However, companies must carefully consider how to ensure their data lakes are implemented properly and kept clean. With the right vision and governance, data lakes, and even that one closet, can provide companies with easy access to the data they need, when they need it. ≈

Drew Ramm
drew.ramm@jabian.com