Bringing Dark Data to Light: How to Handle the Next Great Business Resource

George Kobakhidze

Jun 18, 2020

Dark data is a hot topic in the field of data management. Many perceive it as scary and aren't sure where to start in making it something of value. To back up, dark data is defined as data collected during business operations that otherwise goes unused. This unmanaged content is difficult to monitor, meaning it's hard to notice when information has been replicated, leaked, tampered with, lost, or stolen. It's easy to understand the ominous nature of the discussion around it.

"Dark" sounds foreboding, but only serves to highlight the fact that it's not understood. Similar to dark matter in astronomy, it isn't that dark data is to be feared, but rather that we haven't fully realized its potential.

That said, much of the discussion around dark data treats it as little more than an indeterminable nuisance. While there are difficulties in dealing with dark data, it would be foolish to continue ignoring it, as it is both possible to "tame" dark data and to utilize it as an asset, provided a thoughtful, concise, coherent plan is put in place. How can an organization tame something that, by definition, requires a lack of understanding?

IDENTIFY THE MONSTER UNDER THE BED

Dark data is data that isn't currently understood. "Currently" is the operative word. The first step is identifying what exists within the dark data. This is easier said than done and, in fact, is very difficult to do. As such, it's best to separate dark data into smaller pieces, and then break it down.

The goal is to know the repositories of data possessed by an organization, and then to understand the footprint left by that data. Hypothetically, let's say that an organization has a petabyte of data. Upon breaking it down, it's revealed that one-third of this is comprised of file shares, one-quarter is SharePoint sites, and the rest is email. That's about 80% of the way to understanding dark data, because now, at least, there's an idea of what's out there. This allows further segmentation, enabling things to be further broken down into an examination of the distribution of data.

If one-third of that dark data is comprised of file shares, activity is a good metric to measure. If, of the 200 file shares present, 100 of them are actively managed and modified by users, while the other half are dormant, then the organization has discovered the dangerous part of its dark data and can begin to tame it. If data is sitting unassessed, unattended, or otherwise unoccupied, it's a ticking time bomb—both wasting storage space and containing potentially compromising information. Repeating this procedure allows the taming process to be completed.

Classifying, separating, and codifying dark data into something understandable gives an organization the ability to say, "I don't know everything about this data, but it's being actively used." The danger comes from unknown data that isn't actively used.

After an organization has finished codifying its dark data, it can become an invaluable asset. At the end of the day, dark data is data. It's unstructured data, meaning that instead of traditional 1s and 0s, it's all other content; but it's data, nonetheless. Rather than being processed by simple computers though, dark data is produced by the most sophisticated computers to exist: humans. In the dialogue around data, it's described as "the new oil" or "the next gold rush." That data though is incredibly processed and created by simpler computers. It's easy to place value on it, and therefore clear to call it the next hot commodity. Unstructured data is created by humans, which makes it more difficult to value but potentially worth more than oil and gold combined.

TRAIN THE BEAST AND MAKE IT WORK FOR YOU

Whereas regular, structured data is only useful after processing, dark data is useful in its raw state. It was created by a human for a reason. It can tell a story about a human, the team the human was working with, the state of the team, and the work that was being performed. The inherent value is much greater than structured data. Yes, it's a huge task to analyze it, but there's an incredible amount of value in a raw state. To understand this value though, there must be a plan in place: What is the organization trying to identify?

Ironically, the easiest way to do that involves examining the most structured parts of unstructured data—the metadata. Metadata, such as dates created, accessed, and modified, can help further break things down. It allows context to be determined. Say that out of a 100TB hard drive, most of the data was accessed or otherwise modified around 2007 or 2008.

Dark data could be pertinent to understanding historical events such as the crash of 2008. A financial institute could examine that dark data and answer important questions such as these: How have we evolved, if at all? How responsible were we in this issue? What can we do to ensure this never happens again? Dark data can certainly have value extracted from it, but there must be an understanding of what is being sought. Otherwise, there's a giant content index sitting unused.