Give Your Unstructured Data The Meyers-Briggs (TM)

Don MacVittie, 2010-09-06

For those who don't know, according to the Meyers & Briggs Foundation, part of the Meyers-Briggs Assessment is defined as: *The essence of the theory is that much seemingly random variation in the behavior is actually quite orderly and consistent…* The same can be said about your data. Much that is seemingly random is consistent and predictable. One of the problems currently facing the enterprise is to properly categorize that data so that its "personality" is well known. You cannot sort (or tier) what you don't know, and this is a simple proposal for how you might begin such a categorization.

No matter what your organization does, it has a variety of data in a variety of types with a variety of attributes that can be built into indices to help you understand not only what you have, but how much of it you have, what its relative importance is to the organization, and how you can make use of all of this information to help you move data about in an intelligent manner.

Meyers-Briggs uses initials to give your average Joe (or Jane) an easy-to-access summary of a given individual's personality type, and by extension how to interface with that person. This little tool aims to do the same type of thing for your data, so I kept their initials and mapped them to information about your data that will help you figure out what to do with it. Perhaps a bit gimmicky, but it's valid, so let's get started defining your data

We can break data information into two categories – physical and extended. Physical information can be readily accessed and utilized by automated tiering systems like the one built into our own ARX, but extended information is unique to your organization and some of it will fluctuate over time. That is the hard part that interviews and intelligent data analysis will be required to determine. Not insurmountable, but certainly a task, and if you're a geek that doesn't like to play with "squishy" data, not an envious task at all. Though knowing this stuff will help you come to logical conclusions about where and how the data will be stored.

Physical	Extended
Extension	**I**nterest
Timestamp	**J**urisdiction
Size	**P**ermissions
Filesystem	**N**ecessity

First, clear definitions of each data type.

- **Extension** is the file extension. It will not only tell you type of file (generally), it will also tell you aggregate type. An AVI falls into the video category, for example. Yeah, it can be audio too, but most organizations treat the two media types similarly when making decisions strictly on extension.
- **Timestamp** the last time that the filesystem shows this file as written. If you have a tool (like ARX) that allows you to accurately track last access date/time also, then you could use that information much more intelligently than last save time.
- **Size** Let's face it, the multi-gigabyte file is going to be treated differently than the 10K file just because it is a big win to get it off of tier one storage and on to something cheaper.
- **Filesystem** Files on the SAN generally take more money to keep there than those on the NAS. Now if your SAN is low-end and your NAS high-end, it is possible that this is untrue for you, but either way, knowing what filesystem a file is hosted on helps you to understand what the impact of moving that file will be.
- **Interest** How much interest would this file be to ne'er-do-wells that got access to the storage medium it is currently stored on?
- **Jurisdication** Who is the ultimate owner of this data? The person who can make decisions about its use, distribution, and access rights?
- **Permissions** Who has access to this file, and is it by user or group, is access to this file managed on the file itself, or the file system it is stored on?
- **Necessity** How is this file used within the organization? If it went away tomorrow, who would be impacted and

The idea is to collect all of this information about your files so that you can make intelligent decisions about how to move that data around and store it in the most appropriate place. As I said above, there are tools to help you with the physical stuff, and some of them help with Permissions also. But you'll still need to collect the other data, and that's a lot of work. If you just plain don't have time to interview director-level people about their team's data usage and specific files, then start with directories. Something is better than nothing after all, and behaviorally most groups put like data into folders as far as usage, permissions, jurisdiction, and necessity. After all, the fantasy football spreadsheet isn't generally stored in the new product development folder.

Using these values, you can properly categorize your data, which is the first step to both understanding it and organizing it – and tiering it.

Unlike Meyers-Briggs, these attributes can have multiple non-numeric values, so your tracking will be a little bit more complex than a Meyers-Briggs score, but it will be highly valuable in helping you figure out what to do with your data. Data whose necessity is high will obviously take pride of place on your tier one storage systems – unless it is almost never accessed, which the better version of timestamp could tell you.

If you just don't have the manhours, cooperation, or desire to work through all of this, then invest in an automated tiering product, let it learn, and turn it on. It will get you 50% there, maybe 5/8ths of the way there, with no significant effort on your part. You'll have to install and configure it, and monitor it… But the investment is small compared to interviewing business owners and asking them to make definitive statements about all of the data they own. And it gets you started.

In the end, you can't send stuff that is of high interest unprotected into the cloud, you can't run stuff that is frequently accessed into an archival format, and you want to check how many movies and audio files you have, where they're stored, and how much space they take up. So the more you know, the more power over your storage environment you will have.

Meyers-Briggs is a trademark of the Meyers and Briggs Foundation.

Related Articles and Blogs:

- Wikipedia Meyers-Briggs Type Indicator
- Create A Smarter Storage Strategy (pdf)
- Storage Industry Tackles Making Sense of Metadata
- Tiering is Like Tables or Storing in the Cloud Tier