Leveraging the Power of Metadata to Solve Big Data Problems

WHITE PAPER

By Floyd Christofferson

Metadata as a Rosetta Stone for Data Management

In 1964, The New Statesman magazine first used the term Information Explosion to describe what was thought at the time to be a dramatically increasing deluge of data, and the problems that created. Then, of course, the flood was mainly data on paper, and the increasing proliferation of documents in all industries was becoming a nightmare to manage.

Thus the digital Information Technology era brought a welcome relief from the problems of sorting through, finding, protecting, and archiving increasing mountains of information.

Or did it?

The crux of the problem is how to enable intelligence and unified data management across storage types from different vendors.

In fact, industry analysts consistently show that CIOs and IT administrators rank managing data growth at the top of their concerns. Not only are they dealing with how to manage the cost of storing and protecting increasing volumes of data, but they must also determine which data is worth preserving, and how to extract future value from that data.

This issue extends from small to large organizations, affecting most applications of unstructured data in an enterprise. Big Data is not only about the speed (velocity) and amount (volume) of data created, it is also about the variety of data types and data storage choices that are available, a much harder issue to manage.

The crux of the problem is how to enable intelligence and unified management across multiple data types and storage types from different vendors. It is the practical manifestation of the adage, “the whole is greater than the sum of its parts”. In a data-centric world, anything that limits an organization’s ability to seamlessly manage and analyze its data results in greater complexity, added cost, and limits the ability to drive better business decisions, or achieve new discoveries.

Data Stovepipes Inhibit Leveraging Data Value

The storage industry sees the rapid growth of data as a very lucrative problem for which they position themselves as the solution. Whether it is I/O-optimized flash and disk arrays, capacity-optimized file and object-based, tape or cloud storage solutions, there are plenty of options for storing data to meet the particular needs of the moment.

The problem is that even within a single organization different data types and workflows can drive different storage requirements. Even the particular stage in the data's life cycle results in different performance requirements for the storage it lives on. And inevitably, tomorrow those needs will change and the data would be better placed elsewhere.

Which means storage infrastructures usually end up being a mix of different storage types, of different ages, from different vendors. Costs rise as such data stovepipes emerge. Stale or persistent data can become stranded in expensive high-performance systems by inertia, or simply by the complexity of trying to figure out which data can be thrown out, which can be kept, and how to keep track of where it should live.

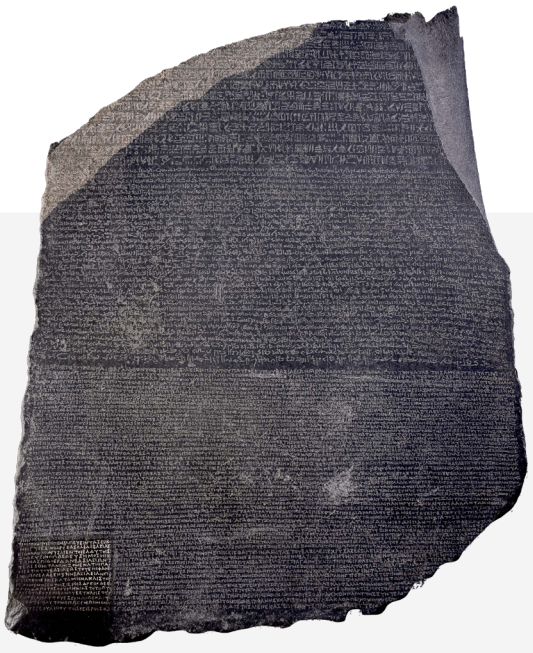
More importantly, as more data stovepipes or silos emerge within an organization, the more difficult it is to manage them and derive useful intelligence from the data scattered across them. Just storing the data somewhere doesn't mean it can be meaningfully exploited. And the complexity this problem causes adds costs as well. That means both storage costs, since data becomes stranded often in the most expensive storage type, as well as operational costs, when administrators must manually move or migrate data to other storage across silos.

Simply building more and better storage containers does not solve the problem of managing data variety and really getting the most value from the data. It would be like car manufacturers adding more fuel tanks to vehicles in response to decreased engine efficiency.

Metadata is the Rosetta Stone

The storage industry is naturally focused on housing the data, since that is what takes up all the space, and drives the cost and design decisions of the infrastructure. But within all data are multiple sorts of metadata that hold the keys to solving all of the problems outlined above.

Metadata is literally data about the data. Think of it as a roadmap that gives you a bird's eye view of everything, without actually needing to access it directly. The traditional infrastructure-based approaches are like planning a trip by first driving all of the available routes before deciding which is best. With a roadmap, the decision is simple and ensures you select the best possible route. Metadata can be leveraged to provide that roadmap to better manage storage resources.



The correlation of three languages on the Rosetta Stone provided a clear understanding that was impossible when each was looked at alone. This is an apt analogy to the power of aggregating metadata to get new insights to your data and storage.

Storage-centric solutions to data management problems simply cannot provide the intelligence about the data they store, nor were they designed to so. But when you can aggregate multiple metadata sources, such information is readily available at any time across all storage. In this way, administrators can now have an intelligent roadmap to manage their data AND their storage resources.

Aggregated metadata gives admins the information they need to make pro-active policy-based decisions to better manage both data and storage resources.

Every digital file contains multiple types of metadata. There is file system metadata that describes its basic attributes, such as file size, location, name, when it was last modified, and so on. But there is also much richer descriptive metadata contained within the files that can enrich the roadmap, and can give you more information to work with. Whether a satellite image, the output of an MRI scanner, a genome sequence or a medical record. Header metadata in standard office files contain information that can provide greater insight when correlated with any of the other types of file system metadata available. Even the absence of metadata can be significant. Everything about the data leaves a digital metadata fingerprint that can be analyzed, and which together can extract maximum value out of digital assets.

So rather than trying to physically normalize all the data into a giant data lake infrastructure, why not create a virtual lake of metadata, leaving the actual data right where it is? Why not use the metadata roadmap to plot the journey?

In this way, all of the available metadata from the many different data types and data locations can be made globally searchable, regardless of storage type or data location. This can drive decisions on how to manage this metadata lake. It also provides a much stronger base from which new correlations and discoveries can be made to get better utilization of the data. And it can all be done without needing to physically move the data, or alter the underlying storage infrastructure.

In addition, based upon these aggregated metadata sources, storage administrators now have the information they need to make pro-active policy-based decisions. StrongLink software from StrongBox Data Solutions includes powerful data movers and storage resource management policy capabilities to leverage this aggregated metadata to automate cross-platform storage resource management based upon business needs.

In this way, StrongLink ensures data is on the right storage of any type at the right time. This includes automating file copy management and data placement across flash, disk, tape and cloud storage.

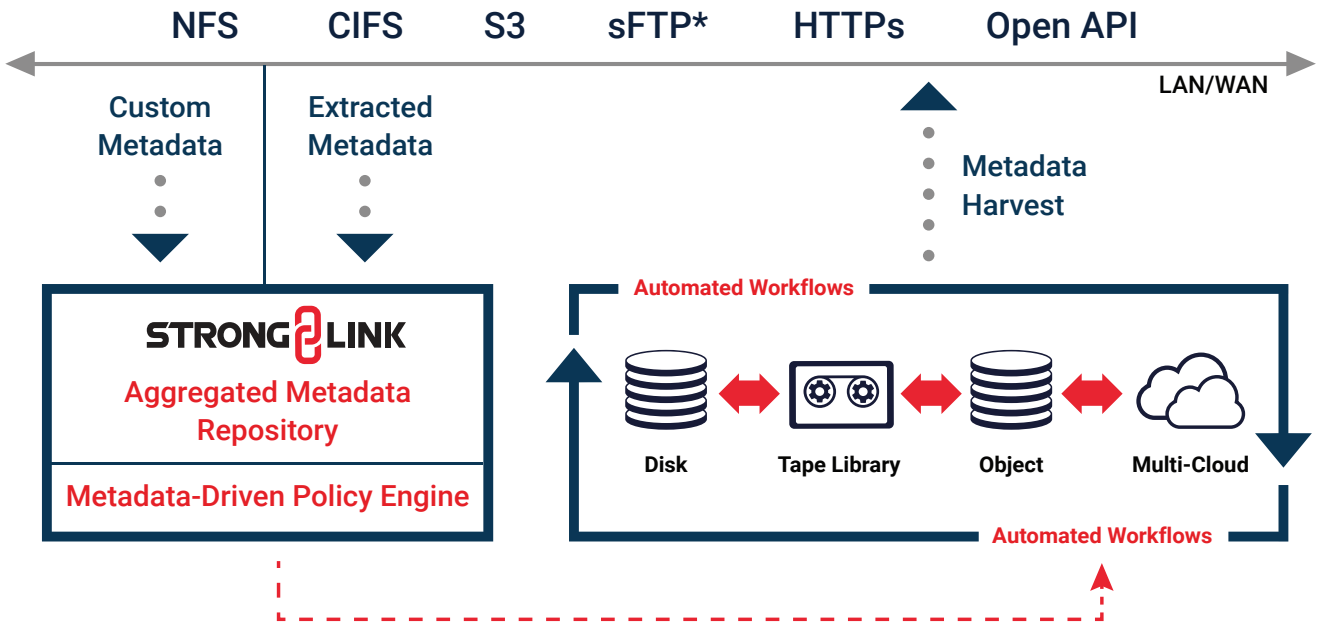


Figure 1. *StrongLink harvests multiple types of file and system metadata from any storage type or file system, which can be used for search, policy-based data migrations, and much more, with a global view across all storage silos.*

Big Data discussions often refer to the concept of needing to create a 'data lake' in which all the disparate data is physically moved to enable this type of collaboration. With StrongLink, the same result is achieved without the expense and disruption of physically moving the data anywhere until it is absolutely needed.

Whether in a small enterprise or a large research environment, StrongLink enables organizations to create a virtual metadata lake to get the advantages of Big Data methodologies based upon the storage they have, without needing to recreate their entire infrastructures. Leveraging the powerful Rosetta Stone of metadata, the whole can indeed be greater than the sum of its parts.