Metadata as the Rosetta Stone to Data Management

Leveraging the power of metadata to solve big data problems

The problem

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?

In fact, studies from Gartner and IDC 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 what data is worth preserving, and how to extract value from that data.

This issue extends from small to large organisations, affecting most applications of data to 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 types, 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. 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 organisation’s ability to manage and analyse all data (seamlessly) results in greater complexity, added cost, and limits the capability to drive better business decisions, or achieve new discoveries.

Data Stovepipes Limit 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-optimised flash and disk arrays, capacity-optimised file and object-based, 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 organisation different data types and workflows can drive different storage requirements. Even the particular stage in the data life cycle results in different performance requirements for the storage it lives on.

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 emerge within an organisation, the more difficult it is to manage them and derive useful intelligence from the data scattered across them. Just having the data doesn't mean it can be meaningfully exploited.

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 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 to data management are like planning a trip by first driving all of the available routes before deciding which is best. A much better approach is to use a roadmap, which simplifies the decision process, and ensures that you select the best possible route.

Storage-centric solutions to data management simply cannot provide the intelligence about the data they store, nor were they designed to. Coalescing multiple metadata types can be done without even moving the data, and provides an intelligent roadmap to data management and useful intelligence to enable new insights without needing to fundamentally alter the underlying infrastructure.

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. Text-based contents of standard office files contain information that can provide greater insight when correlated with all of the other types of metadata available. Even the absence of metadata can be significant. Everything about the data leaves a digital fingerprint that can be analysed, and which can extract maximum value.

So rather than trying to physically normalise 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 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. And it can all be done without needing to physically move the data, or alter the underlying infrastructure.

A prime example of this solution in action on a massive scale is the Research Data Storage Infrastructure (RDSI).

Using Mediaflux, the Research Data Storage Infrastructure Project ties together Australia's many research communities into a massive collaborative data network providing access to over 11 petabytes of nationally significant research content.

Arcitecta is delivering scalable and customised data-connected research platforms that will provide the Australian research community shared with access to nationally distributed data centres (or Nodes), which are expected to grow to over 55 petabytes of research content funded by the RDSI project.

These data sets cover a broad range of specialties, from high-energy physics to the humanities, from climate change to cancer research, and much more. By comparison, the RDSI data repositories will soon contain the equivalent of over 55 times the volume of the entire data store managed by the U.S. Library of Congress.

One outcome of the RDSI project is that researchers will be using and manipulating significant collections of data previously unavailable or difficult to access, driving innovation by enhancing collaboration between researchers nationally and internationally.

The powerful metadata management capabilities of Arcitecta's Mediaflux software are a key component of enabling the RDSI project. By simplifying rapid collaboration across different data types and data repositories, Mediaflux breaks down the barriers between disparate data, enabling researchers to focus on their work.

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 Mediaflux, 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 nationwide research environment such as RDSI, Mediaflux enables organisations to create a virtual metadata lake to get the advantages of Big Data methodologies 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.

-- o0o --

CONTACT US

If you would like more information about how Arcitecta can help you and your business, please contact us.

Suite 5, 26-36 High Street
Northcote, Victoria 3070
Australia

555 California Street
Suite 4925
San Francisco, CA 94104
USA

sales@arcitecta.com

The Americas	+1 303.800.5999
Asia Pacific	+61 3 8683 8523
Europe, Middle East and Africa	+61 3 8683 8523

Facsimile: +61 3 9005 2876

info@arcitecta.com

Arcitecta Pty Limited
Registered Office
11th Floor, 446 Collins Street
Melbourne, Victoria, 3000
Australia
ABN 83 081 599 608