TORI
Experience. The difference.

# *Addressing the Complexity of Enterprise Data Management*

## Introduction

*In the infrastructure world storage has for many years been treated as an exception when it comes to cost. Its share of the IT Budget has historically been relatively small, and it has always appeared too hard to get a grip on: storage is complicated, it is difficult to have direct impact, and there are always other priorities. However, with storage volume growing at >40% per year (according to IDC and Veritas[1]) it can no longer be ignored, despite underlying unit costs dropping by 15% per year[2].*

Storage is a catch-all terminology which covers several different technology sets: SAN, (or Block) which tends to be used underneath databases such as Oracle; NAS, (or File), which tends to be used for filesystems (such as personal drives or shared areas) and Backup (including Archiving), which takes copies of both. Each of the three areas has a different challenge set when it comes to cost.

SAN is a foundation component of an application stack. Typically SAN storage is connected to servers, which support one or more databases, which then support one or more applications. The amount of storage space, (or "disk") required by an application is affected by many factors - such as business volumes, database choice and standards, and recovery strategies. Tracking the relationships between applications and underlying storage involves tracking multiple layers, some dedicated and some shared, and once the number of SAN devices increases to tens or hundreds the complexity of the environment restricts end-to-end reporting.

Alternatively, NAS is used by almost everyone in large organisations. Personal drives and shared directories are basic tools required in any modern office. The challenge is that shared drives diffuse ownership - individuals keep multiple copies of files, or files they don't need, and there are few incentives to encourage the sensible use of available storage. Add to this organisational change, leavers and movers, and very quickly large areas of filesystems are untouched, but cannot be purged as nobody is able to determine who owns them, and therefore who has the authority to delete them. Layered over this is the scale of the challenge: in large organisations there will be tens of billions of files. Simply trying to collate the information on all of these files can require terabytes of storage, and a huge amount of processing power.

Backups are at the mercy of both of the two other technology areas. Logically, the more data which is kept in a database or a filesystem, the more needs to be backed up. Reducing backup frequency requires taking on additional risk which is often unpalatable to the business. While the volumes of SAN and NAS usage grow, the volume of backups must grow too.

1 https://www.cio.co.uk/cmsdata/whitepapers/3596151/brocade-scaleout-it-infrastructure-wp.pdf, and Veritas growth prediction of 39% - https://www.veritas.com/product/backup-and-recovery/netbackup/global-data-visibility

2 Cost Predictions (ZDNET/Wikibon) - https://www.zdnet.com/article/enterprise-storage-trends-and-predictions/

Given the difficulty of producing detailed end-to-end reporting, costs tend to be rolled up and reported to the business units as large, opaque 'chunks'. This prohibits the business from understanding the data, with a resulting drop in engagement, accountability, and understanding of the actual cost of their technology decisions. The feedback loop between business demand and the capacity of the underlying storage is broken, which makes for a dysfunctional deployment and fulfilment process, and poor cost management.

Ultimately these technologies are simple while small, but are challenging at scale. Given that they are growing so fast, the problem is only going to get bigger, and organisations need to get a handle on them now to avoid facing disruption and significant cost increases over the next few years.

# Current solutions

*The storage industry in general approaches the management of storage from a capacity only perspective: new storage is cheaper, there is better compression and deduplication technology, data can be moved to the Cloud. While these all have their place, and are required in a complete storage plan, they do not address the real elephant in the room: the demand of storage is uncontrolled. The typical result of adding capacity is that the previously unfulfilled, pent-up demand in the system consumes much of the new capacity, frustrating any plans for refresh, and undermining any predictions of growth and demand forecasts. Storage requirements are challenged at a capacity level only, (the consumer is asked to reduce their request based on available capacity, or the request is denied) not on the merits of the request itself.*

Storage capacity is managed from an aggregate disk perspective: how much is available, how much has been given out and how much is being used. Determining who is using the capacity, and how, is much more difficult to answer - and not required for a capacity only management model. Trying to build reports to answer these questions soon runs up against the issues of scale and complexity highlighted in the introduction to this paper. The reports end up compromised by the number of exceptions which exist in the environment, and require a tremendous amount of human interaction and analysis, which does not scale with the size of the storage estate.

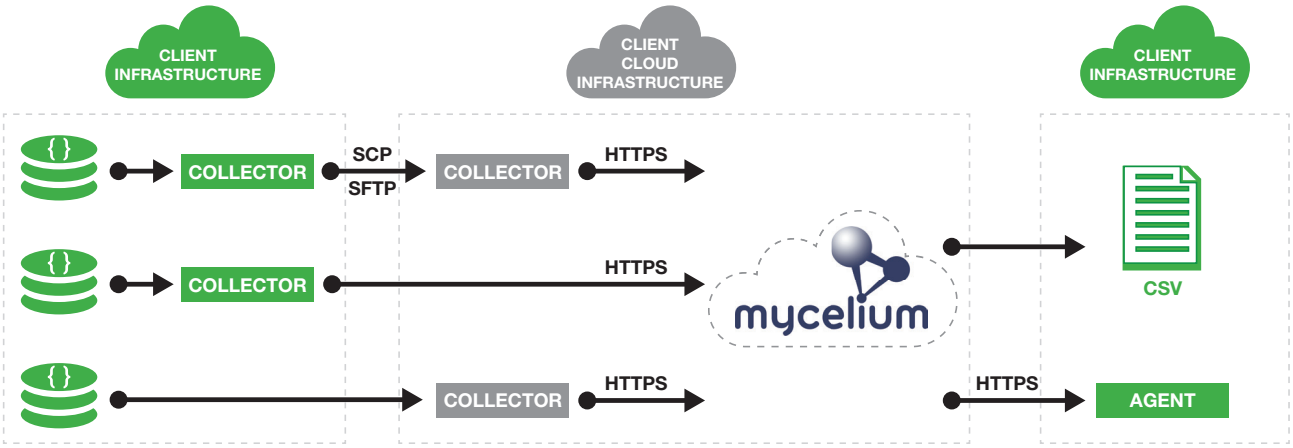
Trying to determine the demand profile usually starts in asking the business groups to provide a demand forecast. This is a reasonable step, but will always face an initial hurdle: in order to predict the future requirements, the business need to know what they already have, as well as their historic growth trends. This is the crux of the challenge facing storage teams. In order to manage capacity they need to understand business demand, and in order to get a demand forecast the business need to understand their capacity. The toolsets and processes currently in place in most large organisations do not address this challenge.

# A different approach

*Modelling storage capacity and demand is ultimately a data problem, and requires a different approach. Existing reporting relies on tables and sets, which are excellent at managing large simple datasets, but incapable of handling complex relationships and exceptions. The solution is to model the storage in a graph database. A graph in this case can be thought of as a tree, with a root, and then some branches, which may branch further and have leaves. Which is exactly the basic structure of a filesystem. Using the structure of the data being modelled allows far more complex analysis of the relationships within the data.*

This philosophy of building the data around its relationships is the core of a new product built by TORI Partner Tech Marionette, called Mycelium. Built on top of the Neo4J Graph database, Mycelium provides a flexible data model which uses both explicit and implicit links within the data to maximise the value of the data consumed.

| | | | | |
|---|---|---|---|---|
| Data sources can be anything. How data is retrieved from them and frequency of retrieval is highly dependent on the source of data. Collection can be scheduled, or manual depending on requirements. | Data is pulled from data sources using collectors. These submit data to Mycelium in a specific format. Collection can be done in two phases with data being collected locally (either as JSON or CSV) and that file copied and uploaded. | Data is inserted into Mycelium over HTTPS via a REST API. Access can be configured to match the mode of operation the collectors are using. | Queries can be run over Mycelium over HTTPS via a REST API. Access can be configured to match the mode of operation the agents are using. | The result of queries can either be displayed using agents via the Mycelium API, or provided as CSV documents, Power BI, Tableau, or similar reports. |




Mycelium is designed to take in data from as many datasets as possible, including AMDB's, CMDB's etc., which are often imperfect. This breadth of data, however incomplete or compromised, provides multiple points for reconciliation which Mycelium can leverage.

This also feeds back to data quality. As a principle Mycelium will accept any data which is not corrupted or malformed (e.g. a word instead of a date). This means there may be 'bad' data in the system, but this then allows data quality reports to highlight the issues, and flag where source systems can and should be tidied up. Given these sources should be the actual configuration systems, this should help clean up the environment.
In addition, the process helps to reduce risk by identifying configuration and security issues (such as incorrect user-name permissions.

Modelling storage systems requires collating several key datasources, organisational as well as technical. Mycelium can take data directly from application APIs, its own scripts or CSV files. As new data is added it will be mapped to the existing model, and extend the completeness of the answers which can be provided. It will also identify gaps and further data requirements to support ongoing continuous improvement in the data set.

# Analytical process

*While using Mycelium provides a powerful toolset to analyse the storage estate, it is not a magic wand for the issues identified earlier, it still requires business context and analysis.*

Starting with a wide data sweep, a high-level dataset is used to build a summary view of the environment, and then identify where deeper data analysis is justified. For example, a simple statistical view of the filesystem environment might identify that ten percent of the files actually consume half of the storage volume. The toolset would highlight this, and as additional data is collected on these targeted files, it will then be added to the model. Where user ID's cannot be mapped to existing people within the organisation, it will be possible to apply methodologies to identify ownership, such as tracking up the directory tree, cross checking Active Directory groups, or providing a list of files for deeper scanning.

As data is added to the model, the more complete and holistic the reporting becomes, iteratively providing greater insight into the business teams. These will include basic capacity views, and also lists of files, or databases, which should be reviewed for purging. At least initially any deletion of files, or reduction of databases, should be specifically authorised by the business.

The ultimate goal of a good enterprise data strategy is to enable the business groups to manage their own storage. They have the authority to delete files they don't need, by providing a detailed capacity view, they will also have the responsibility.

Once the business have this level of control, they will be in a better position to make demand forecasts, and be held accountable to them. This will free up storage teams to manage the underlying platform, optimising the disk costs and performance and storage purchasing decisions.

In addition, as the data model is built in Mycelium it can be extended to support other use cases. For example, once Mycelium has a model of the organisational hierarchy, and the Windows AD environment, it is possible to start investigating logical access management, toxic combinations and data classification.

# Conclusion

*Overall enterprise storage costs will continue to increase as business demand continues to rise. This will not be off-set by the falling unit prices of storage infrastructure. As organisations seek to get to grips with the storage cost base, it is clear the only route to fundamentally reducing this expense is through a detailed understanding of what is stored, where and by whom to empower the organisation to make the right decisions around effective data management.*

Delivering this requires compiling data from various sources to build a holistic view of the infrastructure estate, which in turn should drive a functioning feedback loop between the business and the storage teams. Ultimately, the shift in mindset required is a move to managing capacity effectively, not by simply buying more storage.

Building such a dataset has been prohibitively complex and time consuming. It is for this reason that we developed the Mycelium product. The tool can consolidate imperfect data from multiple sources to provide the insight organisations require to manage their storage demand. The output provides a single, reliable and topical view which can be leveraged across the complex stakeholder network as a basis for addressing the true root causes of the issues explored in this paper.

**TORI London** *+44 (0) 20 7025 5555*
**TORI New York** *+1 212 618 1970*

**toriglobal.com**
*info@toriglobal.com*

TORI
Experience. The difference.