

## **DATA AWARE STORAGE**

**MARCH 2015**

---

### **Taming Data Sprawl Using Real-Time Analytics**

---

Let's face it: Storage is dumb today. Mostly it is a dumping ground for data. As we produce more and more data we simply buy more and more storage and fill it up. We don't know who is using what storage at a given point on time, which applications are hogging storage or have gone rogue, what and how much sensitive information is stored, moved or accessed, and by whom, and so on. Basically, we are blind to whatever is happening inside that storage array. Am I exaggerating? Of course, I am, but only to a degree. Overall, these statements are true. Can we extract information from the storage array today? Yes, we can. But one has to use a myriad of tools, from a variety of vendors, and do a lot of heavy lifting, to get some meaningful information out of storage. The information is buried deep inside and some external application has to work hard to expose it. This activity is generally so cumbersome that most users simply don't use it, unless it is required by law. In such cases (compliance or governance, for instance), external software is used to pull relevant information at great expense and time. Of course, over the past decade, technologies such as auto-tiering, have helped in moving less active data to lower cost storage and one may even find software that automatically deletes files, when their retention period has expired. But these are all one-off solutions and the basic premise still stands: storage today is basically dumb.

What if storage was aware of the data it stored? What if all data was catalogued upon creation, indexed and analyzed? What if analytics were built-in and real-time? What if storage was aware of all activity taking place inside? Who, where, what, how? What if data protection was an inherent part of storage and there was no need for media servers and tapes and separate disk systems? What if search and discovery were an integral part of the array? Wouldn't smart storage like this be a paradigm shift? Wouldn't it fundamentally change how we manage, protect and use storage?

Of course, it would...

Welcome to the new era of data aware storage. This could not have come at a better time. Storage growth, as we all know, is out of control. Granted the cost per GB keeps falling at about 40% per year rate but we keep growing capacity at about a 60% growth rate. This causes both the cost and capacity to keep increasing every year. While cost increase is certainly an issue, the bigger issue is manageability. And not knowing what we have buried in those mounds of data, if anything, is a bigger issue. Instead of data being an asset, it is a dead weight that keeps getting heavier. If we didn't do something about it we would simply be overwhelmed, if we are not already.

The question we ask is why is it possible to develop data aware storage today when we couldn't yesterday? The answer is simple: flash technology, virtualization, and the availability of "free" CPU cycles make it possible for us to build storage today that can do a lot of heavy lifting from the inside. While this was possible yesterday, if implemented, it would have slowed down the performance of primary storage to a point where it would be useless. So we simply let it store data. But today we can build in a lot of intelligence without impacting performance or quality of service. We call this new type of storage Data Aware Storage.

When implemented correctly data aware storage can provide insights that were not possible yesterday. It would reduce risk for non-compliance. It would improve governance. It would automate many of the storage management processes that are manual today. It would provide insights into how well the storage is being utilized. It would identify if a dangerous situation was about to occur, either for compliance or capacity or performance or SLA. You get the point. Storage that is inherently smart and knows what type of data it has, how it is growing, who is using it, who is abusing it, and so on.

In this article we will define the attributes of Data Aware Storage, provide the business benefits of deploying these systems and finally we will provide an industry landscape of up and coming storage companies that are introducing these pioneering products.

---

## DATA AWARE STORAGE DEFINED

### *Requirements*

All storage systems are getting smarter with each new generation but to be categorized as Data Aware Storage we believe they must meet most, if not all the criteria described below:

- **Increased Awareness:**

The storage must store and understand more about the content or attributes of the data stored on the device. Examples are enhanced metadata about quality of service, file attributes, application-aware metrics; actually scanning the data real-time looking for contextual patterns or keywords for security and regulatory compliance.

- **Real-Time Analytics:**

It is not enough for these storage systems to gather enhanced metadata without making it useful in real-time. Therefore these systems must provide instantaneous updates of the enhanced analytics such that administrators or policy engines can react real-time before issues become critical. Example would be the detection and suppression of a rogue application before it can sap IOPS from a more important application Another example would be unlocking the value of data by understanding who is accessing what files, their relationship to others accessing the same files; this would help a business understand what type of data is more important and to which groups of people.

- **Advanced Data Services:**

In addition to reporting the advanced analytics about what is being stored we envision the storage system to have additional data services that enable better business outcomes, based on the increased awareness. Examples would be the availability of archiving functions for dormant data, bursting the application to cloud once a threshold has been met, or balancing QoS across different application workloads. Other examples: triggering compliance workflows or alerts. Or even built in intelligent data protection.

- **Open And Accessible APIs:**

Open API's must exist to unlock the value of the advanced data awareness and capabilities. In order for this new category of storage to flourish all the capabilities of these new systems must be open and available to enable a rich ecosystem of integrated applications and tools to come along side and complement the data aware storage. There are far too many vertical application requirements that could take advantage of unique

data aware features such that no one company could provide it all. Over time natural de facto industry standard APIs will emerge for the most popular enhanced capabilities, similar to how the Amazon S3 data protocol became a standard.

### ***Business Benefits***

The key business benefits that come from this new style of storage system can be described as follows:

- **Customized Business Outcomes:**

These advanced data awareness features should be tailored to a business' needs and can be customized through open APIs. Data Aware Storage can provide this capability by enabling easier integration into a business' unique process management and allow for a business to get more value out of the storage.

- **Mitigate Business Risk:**

Data Aware Storage systems can provide compliance and risk mitigation features. Systems can provide alerts if the wrong type of information is stored in the wrong place or contain wrong types of data. Other systems can provide fine grain user access tracking, who, what, where, when and how often. In addition, these systems can enforce retention policies based on compliance or security needs. Yet other systems can provide advanced data protection policies based on unique data awareness.

- **Bending The Storage Management Cost Curve:**

Storage management needs to be radically simpler in order to bend the cost curve for this important task. Data Aware Storage flips this on its ear by providing the extra data aware information in real-time such that storage issues can be mitigated immediately and other storage management tasks can be automated through open APIs. Having a real-time pulse on all storage activity along with advanced analytics provides a more proactive management approach to storage and actually increases the value of storage as it grows.

- **Cost-optimized Storage:**

This long overdue promise will now become increasingly true. The key to solving this problem is having the real-time metrics on the entire storage environment which a data aware system provides. Knowing what storage is being consumed by what business application and at what quality of service is critical. In addition, tools are needed to move data to the appropriate type of storage to optimize cost. The only way to solve this problem is to have a data storage systems that provides additional storage metrics in real-time all the time. In addition, some systems can provide very unique data optimization, archiving, compliance, and protection schemes based on being more aware of the data stored.

<b>Real World Data Aware Storage Examples</b>		
<b>Storage Issue/Opportunity</b>	<b>Before Data Aware</b>	<b>After Data Aware</b>
Tracking usage and performance patterns for PB sized global file system	1.1 PB system with 790M files doing genomic sequences cannot track disk usage. (“du like command” takes hours and days)	Disk usage tracking done real time including access patterns and IOPS hot spots. (“du like command” completes in less than a millisecond for 1+ billion files)
Security officer wants to know who accessed particular file and when.	Would have had to buy industry specific compliance and archive storage device.	Primary Storage device provides fine grain audit log by file by user and time.
Compliance officer want to know if any PII is accidentally stored in home directories or accidentally added by employees’ BYOD devices.	Would have to buy separate e-discovery engine to monitor and peruse files.	Real-time searching and indexing on content by data aware platforms can mitigate this risk in near real-time and alert compliance officer of suspicious files.

**DATA AWARE VERSUS APPLICATION-AWARE STORAGE**




We should differentiate between Data Aware Storage and Application-Aware Storage. Both terms are used in the industry to position products with any level of intelligence. We believe Application-Aware Storage has some similarities with Data Aware Storage but also some differences. Application-awareness implies that the storage array is aware of some application attributes and/or the application is aware of some storage attributes in such a way that a) makes the interaction more intelligent and b) triggers some actions automatically to optimize/improve application performance and/or storage performance/utilization. Note that while Data Aware Storage attributes apply to all applications that are being served by that storage, Application-Aware Storage is application specific.

Perhaps the clearest example of Application-Aware Storage in the industry is that of Oracle’s FS1 and ZS3/4 product lines. These products are general purpose and offer standard interfaces to support all applications. However, for Oracle applications they invoke special features and procedures that make those applications perform better than they would without them. For example, Oracle implements a special protocol called Oracle Intelligent Storage Protocol (OISP) that enables Oracle Database specific performance and capacity optimizations and assist with provisioning and management in the storage array. This allows Oracle to sell these storage products in the open market for all applications but enjoys enhanced performance, cost and manageability advantages for their own applications.

Data awareness and application-awareness can, of course, coexist. A Data Aware Storage product can also be Application-Aware for certain applications. But an Application-Aware storage product may or may not be Data Aware as we have defined it here.

**MEET THE DATA AWARE STORAGE PLAYERS**

Over time we expect many more vendors to embrace data aware product capabilities as they re-architect their products, however at the time of the writing we consider the following companies at the forefront of Data Aware Storage.

Company	Key Data Aware Focus Areas	Target Markets
	<ul style="list-style-type: none"> <li>• Data Aware through a new Scale-Out File System with massive scalability.</li> <li>• Real-time analytics with extended metadata on a massive global scale (multiple billions of files and petabytes of storage).</li> <li>• Focus on Quality of Service and ability to pinpoint issues by access point and users.</li> <li>• Self-building and self-describing API's for every operation.</li> </ul>	<p>Companies with very large, fast growing unstructured data environments. Commercial HPC workloads in M&amp;E, Life Sciences, Oil &amp; Gas, EDA, Manufacturing, High Frequency Trading, Homeland Security, and the web Internet of Things.</p>
	<ul style="list-style-type: none"> <li>• Data aware through global name space and object access on a distributed object store backend.</li> <li>• Use case specific data aware platform. Optimized to fit the data centricity of the application specific environment.</li> <li>• Intelligent Backup and Archive capability with E-Discovery and compliance features built in.</li> <li>• Content based indexing with search along with extensive metadata analytics.</li> </ul>	<p>High end mid-market to large enterprises focused primarily in Financial services, Oil and Gas and Healthcare providers.</p>
	<ul style="list-style-type: none"> <li>• Data aware built on a primary storage array with unified block and file access.</li> <li>• Expanded metadata on who is accessing the data, frequency and interactions for security and compliance.</li> <li>• Visualization tools to gain business insight through data aware analytics.</li> <li>• Cost optimized data protection built in.</li> </ul>	<p>Mid-market companies with a need for simple all in one storage that can unlock the value of the data stored while also optimizing the TCO of the storage.</p>

While the definition of Data Aware Storage is purposely broad what we are finding is each of these companies is taking a unique perspective of where they want to apply data aware methods to solve very real business storage issues. They do this while also creating business value through data analytics not previously available. For instance, Qumulo is focused on solving the problems for the largest media, life sciences and oil & gas companies (initial markets) with petabytes of data. They emphasize scalability into many billions of files. DataGravity, on the other hand, is more focused on the mid-market and perhaps solving a broader set of problems for such customers. Tarmin is focused on use case specific capabilities, an example would be a data aware storage platform focused on archiving or backup optimization that can simultaneously perform E-discovery, compliance and archive. We fully expect that each will add more data aware capabilities as they evolve their products to meet unique customer demands.

## SUMMARY

Storage has been dumb long enough. We believe the time is ripe for storage to become data aware and thereby radically reduce administrative costs while unlocking the value of the data stored. All the right key technologies are now readily available to make storage smart. As exemplified by Qumulo, Tarmin and DataGravity, data aware storage is not only possible but already delivering serious benefits to customers, especially those that were otherwise buried under mountains of data and losing control fast. We believe the data aware category of storage is in the early stage of development. These companies are all pioneers. They have put a stake in the ground. But a lot of learning is ahead of us. We are sure we will all learn together, as we experiment and see what plays, what works and what sticks. But lest there be any doubt, the time to seriously look at data aware storage is now. Waiting for perfection is a fool's paradise, as we have learned again and again in this industry. These companies represent enough of a leapfrog that they are worthy of consideration. Go ahead, get started. We think you will reduce management costs, improve business insights and reduce business risk, all at the same time. That's more than you could say about any technology to come in the past two decades.

In the future we believe that there will be multiple players who emerge clustered around unique data aware features that resonate most with customers. Look for 2015 to be the year that the Data Aware Storage category starts to take shape as a key emerging technology. We also expect most, if not all existing storage vendors to embrace data awareness; however they would have to significantly re-architect their current products to create offerings equal to those from the pioneers mentioned above.

*A version of this article originally appeared at [www.InfoStor.com](http://www.InfoStor.com) on March 30<sup>th</sup>, 2015*

NOTICE: The information and product recommendations made by Taneja Group are based upon public information and sources and may also include personal opinions both of Taneja Group and others, all of which we believe to be accurate and reliable. However, as market conditions change and not within our control, the information and recommendations are made without warranty of any kind. All product names used and mentioned herein are the trademarks of their respective owners. Taneja Group, Inc. assumes no responsibility or liability for any damages whatsoever (including incidental, consequential or otherwise), caused by your use of, or reliance upon, the information and recommendations presented herein, nor for any inadvertent errors that may appear in this document.