Think Global, Act Local!

Data-aware storage. The difference between a data repository and an active information asset

Sponsored by DataGravity
February 2016

TABLE OF CONTENTS

EXECUTIVE SUMMARY

IT organizations are facing many new challenges at the data and storage infrastructure level, but, when it comes to primary storage, two trends are common to everyone:

Capacity growth. Not only are we dealing with larger data sets, but data retention policies extend much longer than in the past and many organizations are now adopting never-delete policies for most of their data.

Data and workload diversity. The number of applications and access methods have radically changed in the last few years. Now we have many more data types stored in a single storage system, accessed by a larger number of people and devices.

These problems are relatively easy to solve when they are treated individually, but the sum of the two introduces a new level of complexity and it becomes much harder to fully understand and control what is actually stored as well as to exploit the value of data and hidden insights. Furthermore, primary storage has to continue to deliver consistent performance while crawling through stored data for these insights.

The growing number of applications, different data sources, lifetime-long retention periods, and users creating and accessing data from anywhere and any device are heavily impacting the effectiveness of traditional data management and auditing mechanisms while increasing infrastructure TCO and all sorts of security risks!

Data-aware storage systems can be the answer to analyze infrastructure and workloads as well as the data involved, giving a complete picture of what is really happening to your data while empowering business, organization and security.

However, infrastructure is only the tip of the iceberg.
Some of the most critical problems are usually seen at the business level. In fact, managers are not aware of the real state of security, user behaviors, compliance and so on. And when they do have a clue of what is really happening, it's because complex and expensive external solutions are in place. These software solutions usually offer a siloed view of data and limited access to it (through external agents deployed on servers and virtual machines), adding even more complexity, negatively impacting the performance of the production environment, and solving only part of the problem.

Having full access rights to data, without the ability to take advantage of it is another concern in many organizations. Providing the ability to search across the entire data domain, would improve business efficiency and competitiveness while enhancing many other organizational aspects. But, if it means building a specialized infrastructure, the cost could outweigh the benefit.

Next generation data-aware storage systems can do more than just save data safely. In fact, they can be the answer to analyzing infrastructure and workloads as well as the data involved, giving a complete picture of what is really happening to your data while empowering several business, organizational and security processes. In fact, **data security** is a big concern for any organization now, and it has already been proven that traditional security mechanisms are no longer effective on modern attacks and data-leak prevention. Data-

aware storage systems can easily help to mitigate this issue by enhancing security policies through automated search and discovery capabilities and by detecting and reporting unusual user behaviors.

In order to be effective, data-aware storage should have some basic characteristics:

- The analytics engine should be seamlessly integrated with the infrastructure, easy to use and shouldn't impact overall performance of the production environment.
- Data insights and visualizations should be accessible to anyone in the organization who needs to analyze and leverage information coming from stored data.
- It should be based on a no-compromise modern design with all the software features and integrations we have come to expect in traditional storage systems (like snapshots, remote replication, VMware integration, etc.)

When well implemented, data-aware storage improves efficiency of the entire infrastructure by consolidating and offloading data analytics from the application level while enabling advanced and insightful information discovery that can dramatically improve both TCO and business intelligence.

UNDERSTANDING DATA-AWARE STORAGE

Data-aware is a new concept and as such it can be misinterpreted or have different meanings for different end users and vendors. It unleashes the power of data analytics applied to primary (active) data but depending on implementation, it can bring totally different results and serve different use cases.

In order to better understand what data-awareness really means, we can divide modern primary storage systems in three categories according to their analytics capabilities:

INFRASTRUCTURE ANALYTICS

The storage system has the capability to collect sensors and logs and send them to an external analytics system (usually to the cloud). This data stream can be analyzed to get plenty of information about the state of the single storage array and compare it to all the others present in the field.

If the analytics engine is well implemented, the system understands application workloads and can give insights on real system usage and potential problems and automatically open support calls when needed. This approach is very useful in order to have full control over the infrastructure and it can also drive down TCO. But on the other hand, the quality of information collected is not sufficient to dig into data content. In fact, most of the activity is confined to the data container (for example a Data volume or a Virtual Machine).

Local resources needed to implement this type of feature are fairly limited. All the analytics are usually done externally and this has no impact on performance whatsoever.

The list of vendors implementing this type of functionality on their storage arrays is getting very long: Nimble storage InfoSight was the first, but similar products are now available from Pure Storage, Solidfire and many others.

METADATA ANALYTICS

In this case, depending on the implementation, additional metadata is created and maintained by either the storage array or externally. Rich metadata can be created during the ingestion process (as it could happen in object storage systems with file indexing) or it could be embedded in the file system itself (with much more info about files and their utilization over time).

Use cases vary, but this type of indexing and metadata management entails strong search capabilities which are very helpful in managing large repositories (in the order of several Petabytes) or for data discovery.

Infrastructure TCO is still the main target but data discovery or more sophisticated activities for specific applications in vertical markets are becoming more common. In the latter case, it's important to note that pre-packaged solutions are very rare. API and third party visualization tools are still the most common way to

interact with these kinds of systems. Examples of large scale-out file-based solutions that have implemented this type of functionality for vertical applications can be found in Qumulo and Caringo FileFly with Kibana integration.

FULL DATA-AWARENESS

This is the most advanced solution and its scope reaches well beyond the traditional infrastructure TCO, helping to take full advantage of stored data by building an information asset out of it. The storage array allocates specific resources to a full featured analytics engine to provide detailed information about file contents and related workloads. This has to be embedded in the system to capture logs, sensors and real data activity as well. Maintaining a separate metadata-enriched copy of data and its associated workload allows for a better understanding of what is happening to the storage system as well as to the data saved in it.

Data-aware storage can be thought of as a sort of auto-generated data lake, where all data can be searched and analyzed to exploit all the hidden value in it.

This is the only way to analyze the array and its content simultaneously. Complex queries can be run without impacting production while advanced data discovery and full content search represent two other key aspects. Depending on the implementation, this type of analytics can be leveraged on the contents in file shares (NAS volumes), as well as file systems embedded in block-based volumes and Virtual Machines for the maximum granularity. Furthermore, thanks to integration with Active Directory and other user authentication services, it is possible to match users and groups to get a deeper understanding of user behavior and data access patterns.

Reducing TCO is certainly a big benefit of this approach. But data-aware storage can be thought of as a sort of auto-generated data lake, especially for SMB organizations, where all data can be searched and analyzed to exploit its hidden value and potential risks in it. This infrastructure component can automate and offload many tasks from different business units in the organization while improving processes and reducing costs.

Potential use cases range from advanced data recovery and discovery to Auditing, Security and Compliance.

At the moment, the most advanced solution in this space is offered by DataGravity. Other solutions, like Cohesity for example, are focused on secondary storage and their architecture doesn't grant consistent performance for primary data.

WHAT DATA-AWARE STORAGE CAN DO

The ability to search and analyze all enterprise data opens up a new world of possibilities, especially if it comes in a transparent fashion and without additional costs, directly embedded in your primary storage.

At the beginning, enterprise storage systems were designed around resiliency, data protection and performance. Later, data services (like snapshots for example) became table-stakes. Now all these features are taken for granted and organizations, of any size, are demanding more control over data, users and workloads. Data-aware storage systems are the answer.

Organizations, of any size, are demanding for more control over data, users and workloads. Data-aware storage systems are the answer.

A traditional, layered software-based approach is possible but installing and managing external tools is expensive and often implies specific skills that have to be maintained over time. It's also important to note that building an external analytics platform increases costs and complexity.

Implementing a data-aware storage infrastructure enables the adoption of new and smarter strategies in data management, allowing a complete rethinking of many organizational processes. The use cases are endless but the most common can be found in areas like data discovery, infrastructure TCO, proactive security, auditing, compliance and chargeback.

DATA SEARCH AND DISCOVERY

The ability to search the entire contents of the storage system through simple Google-like interfaces allows users and administrators to find relevant information quickly. Thanks to this capability it is much easier to re-use content that is already available and leverage the organization's knowledge.

INFRASTRUCTURE TCO

Thanks to the rich metadata maintained by this kind of system, it's possible to have a complete view of what is really stored in your primary storage, which users use the largest amount of capacity, for what purpose, and what the access patterns are. Modern user interfaces help to visualize this data very quickly, accelerating decision making about what data is worth leaving on primary storage and what data is better to move to cheaper archiving systems or even delete. This helps keep primary storage lean and efficient while optimizing storage spending for the right resources. Eventually, SysAdmins will have a powerful tool to help them in taking defensible decisions on how to manage to "keep everything forever" storage policies.

PROACTIVE SECURITY

Another advantage of having rich metadata and file analytics is being able to quickly discover files based on their content or user-defined rules. Every single file that lands in the primary storage system is tagged and indexed, even if it is saved into a VM, and alarms can be raised to identify potential data breaches or leaks. Advanced reporting can help to identify where sensitive data is (PII, PHI, PCI, etc.) and ensure that it is properly secured and not being accessed inappropriately.

AUDITING

The same technology described in the previous point can be applied to verify compliance and perform real-time auditing over users and data. The storage system has all the information about every file, user and access pattern. By relating them, and by visualizing the results with simple graphs, it becomes very easy to understand how data flows in the system and who is doing what.

COMPLIANCE

What was explained for file access patterns is even more true for content. The system can easily discover user-defined sensitive information in any file, including files stored in VMs, and give a complete map of their content. This information can also be exported and used to create reports or produce legal evidence when necessary.

CHARGEBACK AND SHOW-BACK

Thanks to all the information available from the system, it is also easy to export specific and detailed reports on storage usage, and build chargeback and show-back mechanisms for budgeting and providing evidence on how resources are actually used.

With data-awareness, the role of the storage system changes from being a mere data repository to smart and active information asset.

These are only examples. Once the analytics engine is in place, thanks to its seamless integration with the rest of the system, modern UI, ease of use and advanced query capabilities, the type of reports that can be created are endless, making the primary storage much smarter than in the past. With Data-awareness, the role of the storage system changes from being a mere data repository to a smart and active information asset.

BOTTOM LINE

Full data-aware solutions are the first of a new class of infrastructure components that are smarter and much more proactive. Compared to the past, these new building blocks are capable of interpreting what they are doing and depending on the level of the implementation, give insightful information not just to the System Administrator but to everyone across the organization.

Data-aware storage helps to move the needle from traditional paradigms to next generation data-integrated infrastructures that are finally capable of meeting growing user expectations.

This is happening not only in storage, but similar examples can also be found in networking and computing. This new advanced approach dramatically changes the role of IT regarding business intelligence and enables organizations to answer many more questions faster. It's helping to move the needle from traditional paradigms to next generation data-integrated infrastructures that are finally capable of coping with growing user expectations.

With data-aware storage, the overall infrastructure is simplified and becomes more agile, while more key organizational roles now have access to an unprecedented set of insightful information about all aspects of one of the most valuable assets for any modern enterprise: its own data. And it is made possible without implementing any expensive and complex additional infrastructure to maintain.

Depending on the use cases and user needs, solutions can vary in terms of depth of information that can be carved out from data but its potential and value for the end user is unquestionable.

At the moment, one of the most interesting solutions in the market comes from DataGravity. It has a unique and complete solution covering primary storage needs while providing a wide spectrum of pre-configured analytics tools to various roles across the organization. It's ease of use allows users to quickly get answers without having specialized skill sets.

The DataGravity Discovery Series product, thanks to its ease of use and appliance-based deployment, can be adopted by a wide range of end users who need to do more than just store and protect files. It caters to the kind of user that wants to exploit and capitalize on their data.

Think Global, Act Local!

JUKU

WHY JUKU

Jukus are Japanese specialized cram schools and our philosophy is the same. Not to replace the traditional information channels, but to help those who make decisions for their IT environments, to inform and discuss the technological side that we know better: IT infrastructure virtualization, cloud computing and storage.
Unlike the past, today those who live in IT should look around themselves: things are changing rapidly and there is the need to stay informed, learn quickly and to support important decisions, but how? Rely on us because of our support, our ideas, the result of our daily interaction that we have globally on the web and social networking with vendors, analysts, bloggers, journalists and consultants. But our work doesn't stop there, the comparison and the search is global, but the sharing and application of our ideas must be local and that is where our daily experience, with companies rooted in local areas, becomes essential to provide a sincere and helpful vision. That's why we have chosen: "think global, act local" as a payoff for Juku.

AUTHOR

Enrico Signoretti, Analyst, trusted advisor and passionate blogger (not necessarily in that order). Having been immersed into IT environments for over 20 years, his career began with Assembler in the second half of the 80′s before moving on to UNIX platforms (but always with the Mac at heart) until now when he joined the "Cloudland". During these years his job has changed from deep technical roles to management and customer relationship management. In 2012 he founded Juku consulting SRL, a new consultancy and advisory firm highly focused on supporting end users, vendors and third parties in the development of their IT infrastructure strategies. He is constantly keeping a vigilant eye on how the market evolves, and is constantly on the lookout for new ideas and innovative solutions. You can find Enrico's social profiles here: http://about.me/esignoretti