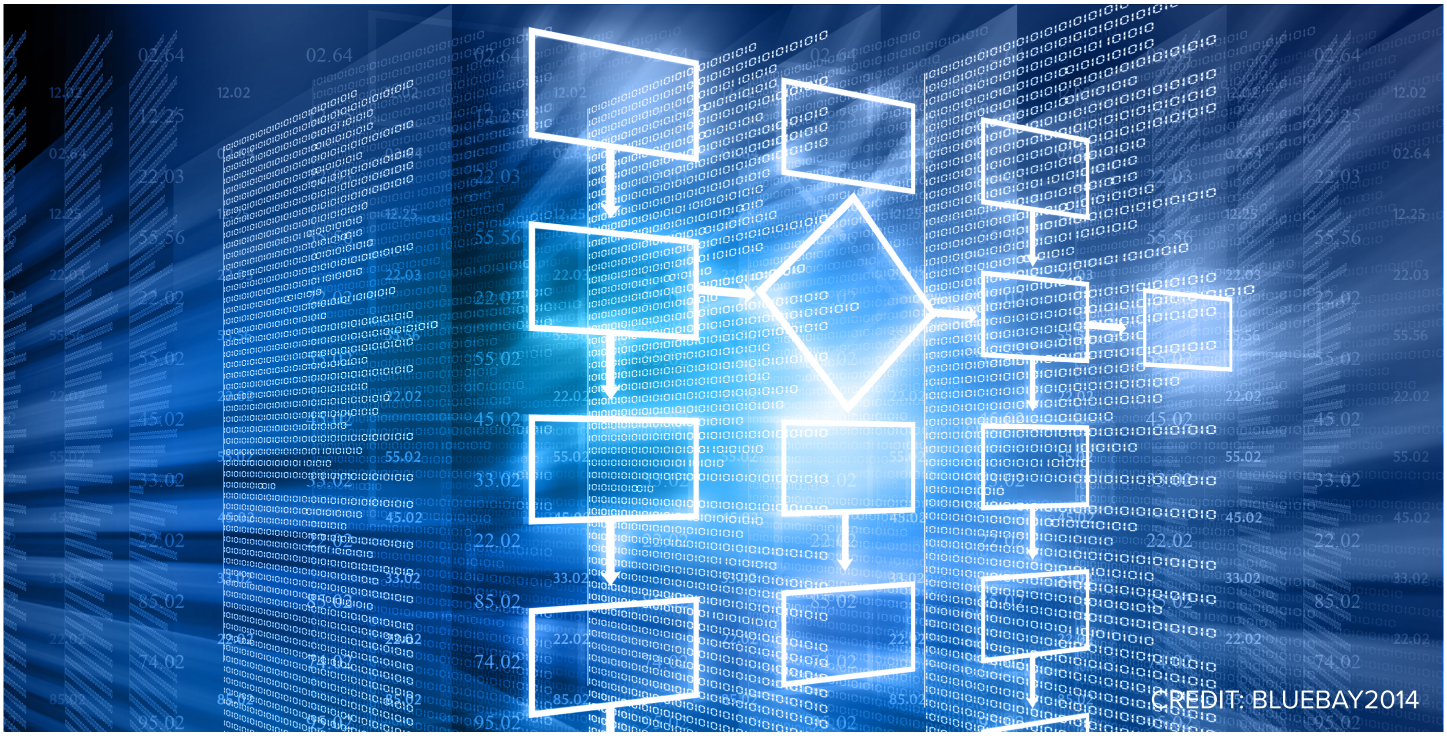
GIGAOM

MARKET LANDSCAPE REPORT

Unstructured Data Management for the Cloud Era

ENRICO SIGNORETTI

TOPIC: CLOUD



CREDIT: BLUEBAY2014

Unstructured Data Management for the Cloud Era

TABLE OF CONTENTS

1. Summary

Unstructured data growth is hardly news anymore. In fact, the challenge is no longer exponential growth, which we are now accustom to and have solutions for, but it is all about keeping data safe while giving access to users, applications, and devices distributed globally, as well as having control over it.

The scenario has been evolving very quickly. In the last ten years we have gone from storing data mostly locally on premises, to storing it in several repositories which are diverse in nature and in the way they are accessed. This trend is not going to change any time soon. Multi-cloud strategies are becoming quite common and data is created and consumed depending on application or user requirements, no matter where it is stored.

Moreover, in this multi-cloud scenario, new regulations like GDPR demand a totally different approach to data management and, without the right tools, it is practically impossible to comply with changing laws.

Most common questions asked by CxOs and IT managers now include:

- What type of data is stored in our systems?
- What are we really doing with this data?
- What are the access patterns?
- Who has access to it?
- Who is really using data storage? And why?
- Is it all protected according to our policies?
- Is data stored in the right place?
- Can we reduce the infrastructure costs?

And the list keeps growing. Data has no value if you can not take advantage of it. Indexing it properly and making it searchable radically changes its value. Making it reusable and consumable by a larger group than the individuals that created it, or just their teams, open up several opportunities including:

- Big Data analytics
- E-discovery
- Compliance and security audits

- And much more

The right strategy and modern tools can help take back control of data and exploit its value, transforming it from a liability to an asset and contributing towards increasing competitiveness.

Furthermore, by analyzing stored data in all its aspects, you can have a better understanding of the logical organization of large and complex IT infrastructures and what area to invest or intervene in to cut costs. The goal is to achieve a better overall TCO and faster ROIs.

In this report we will analyze several aspects of unstructured data management including:

- Different approaches to unstructured data management
- Index, search, and metadata augmentation
- Security and compliance
- Data and metadata management
- Chargeback and Showback
- How data management helps to improve business and infrastructure processes
- Cloud Vs. On-premises solutions

Key findings:

- Unstructured data management is key to having a better understanding of what is saved in growing storage repositories. It can help cut costs and also augment its value.
- The benefits of unstructured data management affect everyone; even small organizations now work with petabytes and need to deal with scenarios that are becoming more complex.
- Understanding content and how it is organized helps to improve data protection, retention, compliancy, security policies, and more. In fact, unstructured data management is key to enabling big data analytics and next generation applications, powered by machine learning and artificial intelligence.

2. Market Framework

Even smaller organizations are now managing large amounts of data, and most of it is unstructured. Petabyte scale infrastructures have become very common among small and medium size enterprises; new business requirements, along with stricter laws and regulations, are posing challenges when it comes to how data is stored and managed.

There are several aspects to take into account when looking at unstructured data now:

Cost: Even though storage for structured data (or primary storage) is still one of the most relevant items of expenditure in any IT budget, unstructured data storage counts for 80 – 90% of the total capacity now, and is growing quicker than any other form of storage.

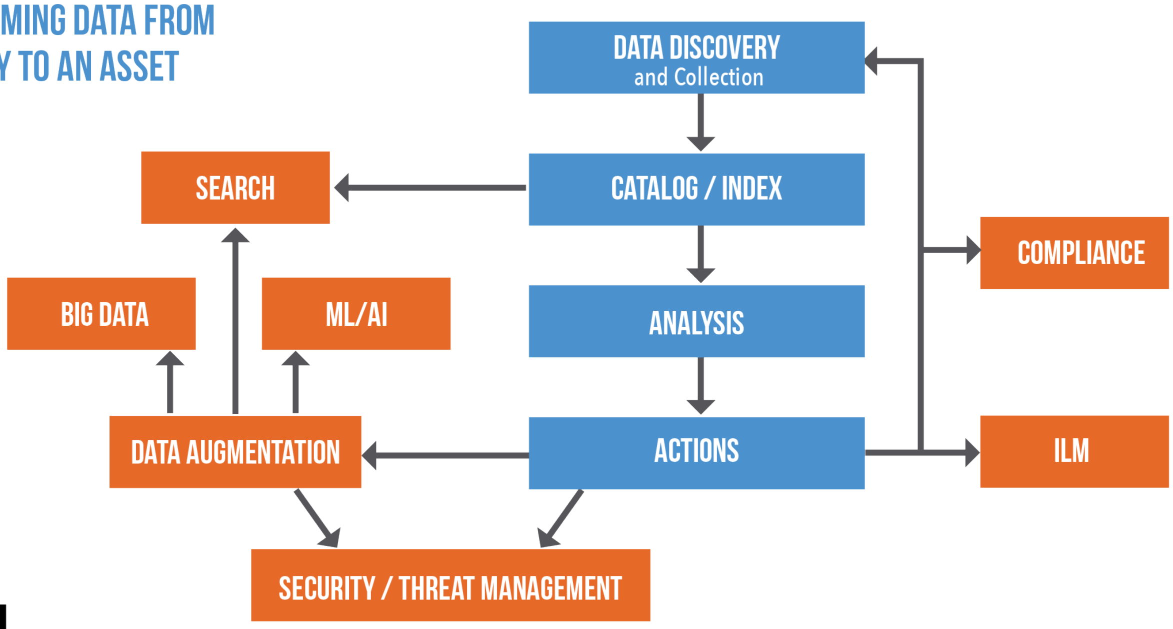
Workloads: The variety of workloads that are now insisting on this storage type is growing steadily, incrementing both its strategic value and importance for day-to-day business operations. In one way or another, Big Data analytics, AI/ML, IoT, and many other workloads take advantage of this class of storage.

Data Dispersion: Most organizations are now looking at multi-cloud strategies favourably. And while multi-cloud is beneficial from many points of view, it makes data management more complex and harder to have a complete view of available data resources, find the right information, or manage security properly.

Data Discoverability, Availability and Access: Finding and taking advantage of what is already available in the organization is becoming more and more difficult. At the same time, it is harder to ensure that the right data is being saved in the right place while keeping retention and security policies in place. These issues become even more pressing when data has to be accessed remotely from everywhere, at any time.

In the process of transforming data from a liability to an asset, data management is key. There are several steps that must be performed, and constantly repeated, in order to ensure the expected return on investment (ROI).

TRANSFORMING DATA FROM A LIABILITY TO AN ASSET



3. Maturity of Categories

There are at least three common approaches to modern data management. All of them have pros and cons and; depending on the size of the infrastructure and its complexity, they also offer different capabilities.

3.1 Storage System-centric Approach

Many modern file and object storage systems offer embedded data management features. Usually they are not particularly sophisticated, and are aimed at minimizing the Total Cost of Ownership (TCO). The most common examples include information Lifecycle Management (ILM), automated tiering based on the age of files stored in the system, and index/search capabilities. This approach is relatively inexpensive and is quite effective for small organizations, or for infrastructures designed to manage a single use case. It does not exploit the full potential of real data management though, but it is simple to implement and use while giving a good, easily reachable ROI. Some of these systems also offer analytic tools to analyze data that highlight many aspects of the storage infrastructure including:

- types of data stored in the system
- how it is stored and used by users
- data workflows
- and more…

This helps to have a better understanding, shows management how storage resources are effectively used, and how to implement chargeback or showback policies.

Unfortunately, this type of solution only works for single system and small scale infrastructures. Data management features are implemented differently by various vendors to manage different systems. With different types of policies, UIs, and results it could undoubtedly become problematic.

3.2 Data Protection Approach

An increasing number of data protection suites include features aimed at indexing data, augmenting it, and also performing additional operations such as providing analytic insights and search. While improving compliance and security, some of these solutions can also provide data/copy management services to simplify and speed up provisioning of data to test/dev environments or analytics applications.

Leveraging backup operations (which has to be done anyway and, likely, already has hooks to a good part of the infrastructure) simplifies data/metadata collection and other operations. This approach is more scalable than the previous one and gives a higher, organization-wide visibility of data.

Data collection made through the backup process creates a copy of the data, enabling out-of-band operations. This does not directly impact data in production, while allowing several aspects of data management to improve:

- Analytics. Data classification reports identify the value of data including orphan/stale files that can be moved or archived; improving storage utilization and infrastructure TCO.
- Risk and security. By aiding the process of identifying threats, potential data leaks, or tampering and better auditing tools. Pattern matching, file scans, or detection of unusual activities on files can all contribute to identify intrusions, malware activities, and other suspect activities on files.
- Data governance. Content scanning allows a better understanding of what is stored and raises alarms to enforce compliance with organization policies or regulations such as GDPR.
- Indexing and searching. By taking advantage of Machine Learning (ML) and Artificial Intelligence (AI), a much more relevant database search can be built which enables semantic search over rich media files or complex documents.
- Legal eDiscovery. Extracting and creating specific data sets outside of the usual system retention for legal purposes becomes easier and less time consuming.
- and more…

A growing number of solutions are now able to protect hybrid and multi-cloud infrastructures, expanding the reach and; therefore, the quality of the results along with the ROI. When comparing these rich features to the previous storage-centric approach, actions to move data across systems, or to force policy compliance, are not usually automated.

3.3 Specialized Data Governance and Management Solutions

The third way to approach the problem is to adopt tools specialized in data governance and management. Depending on the main focus of these solutions, characteristics and results could be similar to what can be achieved from the previous category, but they could also have some features that are more common to the first group, for example, ILM.

These kinds of tools are usually configured to access data stored in several repositories, both on the premises or the cloud. They scan the content to index data and give insights quickly. Data governance, risk and security assessment, auditing, indexing and search, as well as analytics and e-discovery are the most common areas in which these solutions operate best; and, as already mentioned, some of them offer mechanisms to move data across the storage system and between on premises and cloud locations.

Not depending on backup, or specific storage systems for data collection and operations, is an advantage in large and distributed systems that span several locations. This type of solution usually works in an out-of-band fashion, and is less invasive than others when adopted in existing infrastructures; which allows most processes already in place to remain unchanged. Furthermore, these solutions are usually designed from the ground up, with data management and data governance in mind. This makes them more effective in large scale scenarios while being more focused on business process efficiency than infrastructure.

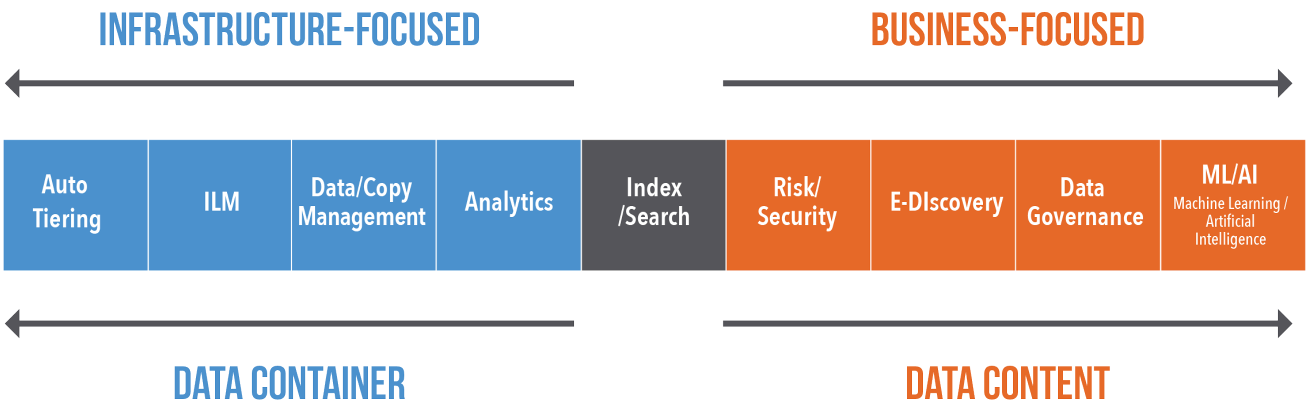
4. Considerations for Using Unstructured Data Management Solutions

In order to find the appropriate solution for each organization it is necessary to understand that data management has different meanings in the IT industry, depending on the person's role in the organization. Depending on who or what the driver to implement it is, solutions may be very different from each other, and it is not unusual to see different solutions competing on the same infrastructure. Unfortunately, many of these tools do not talk to each other and, even if there are overlaps in their scope, they can not take advantage of the work done by others.

If we consider the infrastructure and most technical roles, data management is concentrated on solving TCO, efficiency, data placement, and flexibility challenges. In this context, we find that most common solutions are based on the first two categories we analyzed previously. The focus is not on the content, but on its containers.

- **System analytics:** primarily focused on understanding data storage utilization and activity to improve capacity planning and provisioning, as well as chargeback and showback.
- **ILM and automated tiering mechanisms:** thanks to information collected through system analytics, they help move data and find the right location for it based on user-defined policies. Helping to save on precious, faster storage resources by moving cold data to low cost, low performance, storage or cloud.
- **Copy Data management:** these features are again focused on the management of multiple copies (including creation, clone, retention and clean-up) of the same data for test/dev, analytics, DR, and other use cases while keeping track of them.

By moving to the business and strategic side of data management, other roles in the organization may be involved. And the discussion is always focused on the content or, more often lately, its value and how to take advantage of it.



Traditional data governance solutions (for compliance, auditing, and security) are now often accompanied by sophisticated index and search, as well as e-discovery features. At the same time, new features aimed at increasing the value of data are increasing in popularity. In fact, through ML and AI techniques, it is now possible to understand the content and augment metadata fields, or enable semantic search to have a better understanding of context and user-intent, with the goal to find more relevant information for the business or to feed other applications. Data augmentation can also be useful to query.

There are several benefits that come from unstructured data management, but there is no single solution that can cover all needs.

5. Vendors to Watch

As already mentioned GigaOm has identified three major categories where interesting solutions have been placed for unstructured data management: Storage system-centric, data protection, and data governance. Although most of these solutions have features that span in more than one category, the classification was necessary to better understand how the products interact with data and metadata in terms of collection and analysis.

5.1 Modern System-Centric

Commonly focused on infrastructure TCO and features aimed at optimizing data placement and provisioning.

Komprise

Intelligent Data Management solution works with file system based storage and analyzes the contents to understand how data is used and moves it to secondary object storage or cloud repositories when necessary. This is a SaaS based solution with data analysis and movement agents called Observers installed locally in the datacenter which are in charge of seamless data movements between tiers.

The first goal of Komprise is TCO reduction thanks to its ability to place data on the right storage tier, but it also generates comprehensive reports aimed at improving capacity planning, chargeback, and showback policies. Even more, its analytics features can be used to improve and enforce policy and regulation compliance (such as GDPR).

5.2 Data Protection

Cohesity

Cohesity offers an end-to-end solution designed to tackle all secondary data and apps challenges now present in modern enterprises. It is available as a scale-out physical or virtual hyperconverged appliance, software-defined platform on leading 3rd party servers, as well as through service providers and major public cloud providers.

Often starting with integrated data protection, users can consolidate disparate workloads including backup, archiving, file shares, object stores, test/dev, and analytics onto one software-defined platform, simplifying the ability to store, protect and manage massive volumes of secondary data to unleash its value. On top of the efficient web-scale distributed file system, this solution offers index and search capabilities, copy data management for test and development, automation, and extensive reporting capabilities. Even more so, the user can take advantage of the native MapReduce powered Analytics Workbench to run customized jobs against data with the goal of analyzing content for a growing

number of use cases, including e-discovery as well as finding and neutralizing security threats or improving compliance.
Cohesity integrates its solutions with an increasing number of technology partners, including HPE, Cisco, and Dell, service providers, and major public cloud providers, offering customers compelling ecosystems worldwide.

Druva

Druva inSync, part of the Druva Cloud Platform, provides centralized protection and management across end user data sources, and is offered as a SaaS service. By unifying distributed data across endpoints, cloud applications (Office 365, G Suite, Salesforce, Box) organizations have a single place to manage backup and recovery, archiving, legal hold and compliance monitoring to minimize data risks and ensure continuity for employee productivity.

Druva inSync can also check for unusual activity patterns to minimize the impact of ransomware as well as enforce data privacy and sovereignty policies based on employee region. The solution offers a complete and highly configurable auditing and reporting system to help system administrators and business managers to act quickly on insights and detailed reports.

Hitachi Vantara

Hitachi Vantara has an end-to-end data management strategy for unstructured data that finds its execution in the integration of HDID (Hitachi Data Instance Director), for data protection and copy management, with HCP (Hitachi Content Platform) object store. Once data is copied and organized in HCP, HCI (Hitachi Content Intelligence) can further optimize data and metadata and make it available to other tools like Pentaho (data analytics suite).

This solution can be optimized for several use cases including indexing and search, data governance and compliance, auditing, e-discovery, ransomware and other security threats detection. HDID can be utilized with a broad variety of sources, including non Hitachi storage systems, while HCP and Pentaho are designed for high scalability and can be deployed in hybrid cloud environments.

Igneous

Starting from data protection for file-based sources, Igneous provides a comprehensive data management solution for large unstructured data environments. Making up the data management portfolio is a set of services that include data protection for file-based sources including public cloud integrations. Optimizations services solving the capacity management problems of aged data, chargeback and showback. And lastly at scale indexing to provide search capabilities across your enterprise environment.Igneous provides a comprehensive data management platform for large scale environments. Starting from data protection, or file-based sources, Igneous offers solutions ranging from optimization of storage utilization and chargeback, down to metadata tagging, file classification, indexing and searching, and sophisticated customized data views.

Furthermore, thanks to the features described above, the company is now previewing a new API-based powerful tool, DataFlow, that can find relevant data across several storage systems and build new data sets on the fly which can be moved, or presented, next to the compute resources and applications that need them. Big data analytics, HPC and any other application operates on data available on large scale multi-vendor distributed infrastructures are the most common use cases for this technology.

Veeam

DataLabs, part of Veeam Backup & Replication suite, is continued evolution of a solution first introduced in 2010 that allows customers orchestrate restores in a sandbox and run applications against them. First use cases showcased by Veeam include ransomware protection, secure restores, and GDPR compliance;, but the vendor is planning to develop more out-of-the-box templates and involve partners to provide a broader range of use cases.

>At the moment, DataLabs has to be considered as a workbench that the end user can customize for specific needs but, in the future, the same tool could be utilized to run analytics jobs, indexing and full content search, e-discovery applications, and more.

5.3 Data-Governance

Aparavi
Aparavi is a SaaS-based multi-cloud data management application that provides strong metadata handling and data classification capabilities for indexing and search, including the ability to search within file content. This makes this solution particularly interesting for data management and governance use cases, like compliance, e-discovery, legal holding, intelligent archives, smart data lakes, analytics, and so on.

Aparavi policy-based data management engine enables the end user to set data lifecycle rules and dynamic migrations seamlessly across several on-prem and public cloud backend storage systems, limiting egress fees and increasing data availability while making the protocol translation transparent to front-end clients and applications, which can access data through standard interfaces like S3, file (SMB/NFS), and native APIs.

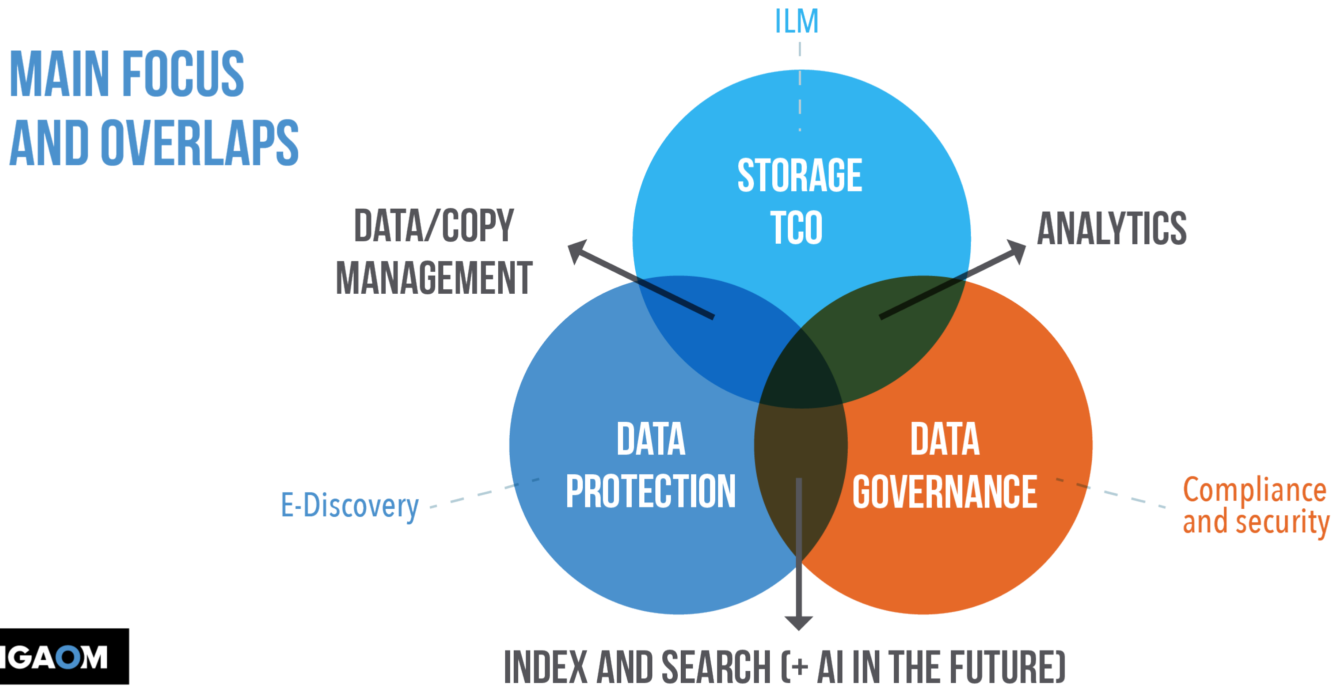
Aparavi has partnerships with all major cloud providers and on-premises object stores. Data is saved in an open format and the product is available through a subscription-based licensing, enabling end users to limit the initial investment and avoid the creation of cloud silos or lock-ins.

Commvault

Activate is a new solution from Commvault that, starting from backups or other sources, can analyze and provide insights and extensive reporting on data storage utilization and trends. It offers index and searching, data classification, and content scanning, enabling to look for specific patterns with the goal to identify issues and enforce compliance or solve security threats.

Most sophisticated applications of this tool include the possibility to take advantage of machine learning and artificial intelligence for contextual and cognitive search through content recognition. AI and ML also enables sophisticated content processing, which can include advanced data masking and transformation. API are available for integration with third party tools and custom application development.

6. Near-Term Outlook



The market is evolving rapidly. Unstructured data is polarizing a lot of the budget expense, be it CAPEX or OPEX. Since it is not really possible to control data growth, the only way to face this challenge is by managing it properly.

In most cases, the solutions available in the market are focused on TCO improvement or data governance; they are not really focused on augmenting the value of data stored in the systems. Data and metadata augmentation, alongside machine learning and artificial intelligence, are going to change the way data is managed, conserved, and consumed (or recycled for lack of a better term). These type of tools are still at the early stages of their development and will need some time before becoming main stream. Maturity of these products will also help to reconsider and consolidate data management strategies, now focused on solving single issues in separate teams and silos.

7. Key Takeaways

Unstructured data management is key to understanding and taking decisions regarding an ever-growing part of the infrastructure and the data stored in it. Since it is not possible to resist data growth we have to identify the right tools to keep it under control and make the best use of it. This is particularly true now, with data stored in several different places across on-premises and cloud locations.

Even though we are not there yet, the final goal is to make data reusable and available for new applications across the entire organization. Data and metadata augmentation alongside new enabling technologies, like machine learning and artificial intelligence, will help reach this goal. In the meantime, the solutions available in the market can address a growing number of pain points while delivering an improved TCO and better visibility over data, costs, and its utilization.

8. About Enrico Signoretti

Enrico has 25+ years of industry experience in technical product strategy and management roles. He has advised mid-market and large enterprises across numerous industries and software companies ranging from small ISVs to large providers.

Enrico is an internationally renowned visionary author, blogger, and speaker on the topic of data storage. He has tracked the changes in the storage industry as a Gigaom Research Analyst, Independent Analyst and contributor to the Register.

9. About GigaOm

GigaOm provides technical, operational, and business advice for IT's strategic digital enterprise and business initiatives. Enterprise business leaders, CIOs, and technology organizations partner with GigaOm for practical, actionable, strategic, and visionary advice for modernizing and transforming their business. GigaOm's advice empowers enterprises to successfully compete in an increasingly complicated business atmosphere that requires a solid understanding of constantly changing customer demands.

GigaOm works directly with enterprises both inside and outside of the IT organization to apply proven research and methodologies designed to avoid pitfalls and roadblocks while balancing risk and innovation. Research methodologies include but are not limited to adoption and benchmarking surveys, use cases, interviews, ROI/TCO, market landscapes, strategic trends, and technical benchmarks. Our analysts possess 20+ years of experience advising a spectrum of clients from early adopters to mainstream enterprises.

GigaOm's perspective is that of the unbiased enterprise practitioner. Through this perspective, GigaOm connects with engaged and loyal subscribers on a deep and meaningful level.

10. Copyright