



EMC Centera as a Production Archive — Archiving Data Sooner and Accessing It Frequently

Analyst: Anne MacFarland

Management Summary

When it was launched back in 2002, EMC's *Centera* was an unusual product. The design point – long-term disk-based retention of information that probably would be accessed infrequently – drove many innovations. Since the information might outlive the storage platform on which it originally resided, support for long-term retention demanded a simpler access mode. Thus, the object was to be accessed, not as an assigned block, or a file at the end of a file path, but by a hashed *content-addressed* (CAS) tag. Since the information it housed was not of the traditional “mission critical” variety, *Centera* had to have costs taken out – and so it was designed to store only one copy of any unique information. The information as a whole is replicated only once to another machine. Cost containment also dictated that *Centera* be self-managing and self-healing. It had to scale huge, and insure the authenticity of the information it retained. With all these parameters in mind, EMC architected *Centera* as a self-sufficient black box. It was very different from typical high-end transaction-oriented storage arrays.

With *Centera*, EMC was first to market with a 21st Century disk-based *electronic archive* to meet the needs of information-rich customers. Interestingly, over time – and particularly recently – **customers have expanded the use of their Centeras, archiving more information sooner, thus creating what EMC calls a *Production Archive*. This new use contravenes some of the early assumptions for *Centera*.** The frequency of reads, which were expected to be a thin sliver of the *Centera* workload, is growing to equal writes in a production archive. And, in these production archives, information is being archived as soon as three weeks after it has been created or captured.

Today's *Centera* is meeting Production Archive demands while addressing the challenges of both the increasing volume of business data and of government regulations about that data. It also addresses several more general storage challenges.

- **Backup Mitigation** – Electronically stored information is fragile and must be replicated (“backed up”) to ensure ongoing availability. In the past, users have used this fragility to justify keeping a local copy, which often is inadequately categorized and identified. Properly archiving information (with a redundant copy) removes it from the back-up queue. **The rate of growth of business information is now such that reducing the amount of information to be backed-up must be optimized. Some have seized on deduplication within backup as the appropriate discipline to do so. Archiving is more pre-emptive – and effective.**
- **Securing Information Sooner** – Regulations aside, no business wants to rely on bogus data. Prompt archiving narrows the window of time when someone or some process can alter a value.
- **Wider Information Access** – Sharing information between people or processes works best when the information comes from a sharable source of well-documented information that is protected from common sources of corruption.

When a vendor is surprised by the uses to which its product has been put, it is a signal of profound change – change worth studying. For a closer look, please read on.

IN THIS ISSUE

➤ The Retrieval Imperative	2
➤ What Has Changed	2
➤ Why a Production Archive?	3
➤ How <i>Centera</i> Has Changed	4
➤ Conclusion	4

The Retrieval Imperative

Traditionally, information retrieval is done by the application as part of its execution. Information can also be stored in and controlled by databases and data warehouses (each with extensive disciplines and practices) and, in the *ad hoc* collections of spreadsheets and other shadow IT mechanisms that have no control and are the bane of IT administration.

Consider today's prototypical organization with Internet and self-service capabilities. The basic order-to-cash process of commerce is often highly automated, removing much of the repetitive work of yesteryear. It is possible for even small organizations to have a large and global value net of customers, partners, and other stakeholders. This would seem to be good news.

However, supporting all parties in the style to which they have become accustomed is another story. **Good support relies on recalling all the information necessary to render an experience that is satisfactory to all parties.** This information seldom sits neatly in one application, or is in one place (or even a handful). Enterprise search¹ continues to struggle to address the enterprise domain in an adequate fashion.

Information retrieval, or recall, grows more challenging as new information sources provide additional potentially relevant elements to be used. The introduction of maps, for instance, has made formerly obscure business patterns glaringly obvious. **What once were neatly stove-piped sub-processes, often now re-characterized as services, now borrow and feed information to better optimize the larger process.** (See Exhibit 1, below.)

Transactional vs. "Other" Information

As businesses have digitized more of their information assets, a bifurcation has developed between transactional workloads and all the other workflows that surround and support those transactions. Transaction workloads continue to be optimized for faster time to completion. However, **most of business information growth is in unstructured (non-transactional) information. For that information (medical images, contracts, information to be analyzed, etc.) sub-second retrieval of information meets business requirements. Faster is not as important as secure and genuine.**

What Has Changed

The file cabinets that assisted recall in the past have morphed into electronic filing systems – but the folders in electronic filing systems are not as searchable as the paper ones. Copying is far too easy and, since corporate naming and versioning policies are often contravened, the right version to use is not always clear. The early success of search engines, such as Google, was with HTML documents (Web pages) – which was stunningly effective – but tagging originally was constrained to HTML's roots. Search has greatly improved, but better context documentation in the data sources is needed to make business search truly adequate.

Use of XML-metadata has blossomed and "find all" is improving, but it still is accomplished most effectively when the information target is a repository, not a file system or LUN. Repositories include elements of control that other alternatives do not. (See Exhibit 2, on the next page.)

Exhibit 1 — Data Use Then, Now, and in the Future

Category	Then	Now	Future Imperatives
Transactions	Rapid. Deft locking was the focus	Faster, leveraging more CPU memory. Contention moves to I/O	Real-time streaming and in-line analysis provide more characterized data for better targeted use
Non-Transaction Workflows	Manual configurations and tuning	Virtualization eases configurations, scale out replaces tuning	Geographic dispersion of workforce may make caching relevant again
Decision and Customer Support	Best effort	Mission Critical	More data sources and targeted analysis enable customer intimacy, but require pervasive virtualization and sharable access
Operations and IT Management	Limited automation	More automation, virtualization	Predictive analytics more fully characterize exceptions for appropriate action

¹ Web Search, by contrast, is easy and straightforward – a piece of cake!

Exhibit 2 — Venues for Data at Rest

Characteristic	Blocks (SAN)	File System	Archive Repository
Stored as:	LUN	File	Object
Access Method:	Through an application	File path (note: search is stymied by file trees)	Directory of objects, search and seek
Access Control:	Permissions as part of object	ACLs, as part of file system	ACLs, as part of application
Optimal Use:	Transactions, distributed updates	Presentations, workflows	Presentation, secure preservation
Features:	Features are storage efficiencies like snapshots and thin provisioning. Block storage is focused on performance and security.	File systems provide basic access control and locking	Check-in, checkout, versioning, auditable as well as controlled access, audit trails of use.
Vulnerabilities:*	Disk corruption, slow rebuild of large disks	File system corruption, inability of basic search approaches to fully parse file trees	Slower access

* All vulnerabilities can be mitigated or solved, but not for all use cases.

Data Has Changed

- **Businesses have access to more data sources, both internally and externally.** More data and information elements (not just documents and values) are destined for real-time display on role-specific dashboards.
- **Social software supports interactive conversations critical to business** – conversations that must become part of the business record, not just for each contributor but also as a whole.
- **Blobs** (large binary objects, such as security and medical images) **abound**, and often must be shared.
- **Sensor data**, often used in the aggregate or as streams, **is critical to optimize operations**, and for post-event forensics.

Data Use Has Changed

Analytics leveraging these new data sources has become distributed throughout key business processes and user dashboards, and the retrieval and use of information beyond the context of its generating application has become more sophisticated. (See *Future Imperatives* in Exhibit 1 on previous page.)

- **The pace of business** has driven parallelization as a method of shrinking time to completion – of analysis or of information retrieval. This architecture, in turn, requires clustering and coherent caching strategies to keep data quality high. This becomes inherently difficult with unorganized data sources. The efficiencies and control of a central repository have become

more important.

- **New disciplines**, such as eDiscovery, extend the scope and particularity of data access. They also mandate a degree of *ad hoc* enforced retention never anticipated by hardware or software vendors. In contrast, issues of privacy have been a part of business and their operational systems (paper or electronic) for a long time.
- **New initiatives**, such as social software, customer self-service, and the ambiguous domain of ratings and recommendations challenge the notions of privacy and control for both businesses and consumers. (By contrast, the notion of business confidentiality persists).

Why a Production Archive?

A production archive becomes appropriate when certain business conditions occur.

- Most, if not all, business elements are digitized.
- The information in this “more” of information sources can be characterized or tagged to enhance findability as well as retrieval (the terms are related but not identical).

A decade ago when Centera was introduced, neither of these conditions was at a point where archiving, as described on page one for the original Centera design point, could also pertain to most business information. Storage management and automation were not at a point where off-loading something other than structured data made sense, and the costs of that management had not yet become a pain point.

How Centera Has Changed

Centera² has changed in many ways. Its processors are faster, and its disks are both faster and of higher capacity. Its read and writes performance has increased from 50 objects per second³ at the initial launch to a read of 900 objects per second and write speed of 650 objects per second. The number of objects supported by each node has grown from 5 million to 100 million. Hundreds of business applications now support Centera.

The following enhancements are those most relevant to supporting wider, more opportunistic use.

Seek (2004)

Relatively soon after Centera's introduction, EMC contracted with FAST Search and Transfer (now a part of Microsoft) to use its each product, renamed Seek, to search metadata stored in Centera specific to each application/user's use of the single stored copy of unique information.

The more diverse the use of a repository, the more it needs comprehensive finding tools in addition to the basics of classification schemes and/or registry. This becomes especially important when the needed information spanned applications. In such a case, a horizontal search of an archive would be faster and more accurate than individual stovepipe searches through a myriad of applications. Seek searches all metadata in the archive returning to the requestor all information that meets the criteria.

XAM Support (2009)

To facilitate use by disparate parties and applications, it makes sense to use a standard data access. SNIA's *eXtensible Access Method* (XAM), was developed as an access method that could be used by *all applications* to address data stored on *all kinds of storage devices* (not just CAS). EMC is working with the ISV community to migrate their integrations from the Centera API

² For a historical look at Centera, see **The Clipper Group Navigator** entitled *Retrieving the Needle in the Haystack – EMC's Centera Manages by Content*, dated May 20, 2002, and available at http://www.clipper.com/bulletins/2002/EMC_Centera_final.pdf. **The Clipper Group Navigator** entitled *The Value of Guaranteed-Authentic Information- The Expanding Role of EMC's Centera*, dated April 18, 2003, and available at <http://www.clipper.com/research/TCG2003015.pdf> and **The Clipper Group Voyager** entitled *Archive Before Backup – EMC Centera's Prescription for the Smaller Enterprise*, dated July 7, 2005, and available at <http://www.clipper.com/research/TCG2005042.pdf>.

³ Remember object can be large- and now have grown larger (think MRIs and videos).

to the industry-standard XAM *where and when appropriate*. This change is consistent with Centera's leveraging of standard components where they exist. It also opens new markets and use cases – not only for Centera, but for the ISVs as well.

As an open standard interface for reference (archivable) information, XAM provides the needed functionality and supplants proprietary methods that hampered data federation and made qualification of information-centric solutions massively difficult across the myriad of combinations of archiving applications and storage platforms. XAM can support millions (and, in theory, billions) of objects. It comes as an API and also facilitates the data migrations that are inherent to very long-term retention.

Future Directions for Archiving

EMC customers' repurposing of Centera as a production archive indicates future requirements for archive repositories. To be successful, information policy management must become deft and consistent throughout an organization's archive environment. The future archive environment must allow seamless content migration within that virtualized archive based on individual customer needs. Because of these requirements, virtualization will become a central and compelling strategy.

Conclusion

In the context of the frenetic but frugal present, archiving is appropriate for more than end-of-life data. That EMC customers have discovered this, and are leveraging Centera's gains in speed, ease of use, and reduced cost in new ways, is not surprising, but it is significant.

It is hard to abandon old habits. However, there are now new business requirements, particularly in the area of information presentation in the right time, and context. The challenge presented by these requirements is compounded by the growth and variety of business data sources. Thus, it is time to re-examine data and information strategies.

That EMC customers have revised their strategic use of Centera is one example of such thinking. Consider your enterprise. A production archive offers many value opportunities. An archival rethink just might optimize or improve the entirety of your future operations.



About The Clipper Group, Inc.

The Clipper Group, Inc., is an independent consulting firm specializing in acquisition decisions and strategic advice regarding complex, enterprise-class information technologies. Our team of industry professionals averages more than 25 years of real-world experience. A team of staff consultants augments our capabilities, with significant experience across a broad spectrum of applications and environments.

- ***The Clipper Group can be reached at 781-235-0085 and found on the web at www.clipper.com.***

About the Author

Anne MacFarland is Director of Information Solutions for The Clipper Group. Ms. MacFarland specializes in strategic business solutions offered by enterprise systems, software, and storage vendors, in trends in enterprise systems and networks, and in explaining these trends and the underlying technologies in simple business terms. She joined The Clipper Group after a long career in library systems, business archives, consulting, research, and freelance writing. Ms. MacFarland earned a Bachelor of Arts degree from Cornell University, where she was a College Scholar, and a Masters of Library Science from Southern Connecticut State University.

- ***Reach Anne MacFarland via e-mail at Anne.MacFarland@clipper.com or at 781-235-0085 Ext. 128. (Please dial “128” when you hear the automated attendant.)***

Regarding Trademarks and Service Marks

The Clipper Group Navigator, The Clipper Group Explorer, The Clipper Group Observer, The Clipper Group Captain's Log, The Clipper Group Voyager, Clipper Notes, and “*clipper.com*” are trademarks of The Clipper Group, Inc., and the clipper ship drawings, “*Navigating Information Technology Horizons*”, and “*teraproductivity*” are service marks of The Clipper Group, Inc. The Clipper Group, Inc., reserves all rights regarding its trademarks and service marks. All other trademarks, etc., belong to their respective owners.

Disclosure

Officers and/or employees of The Clipper Group may own as individuals, directly or indirectly, shares in one or more companies discussed in this bulletin. Company policy prohibits any officer or employee from holding more than one percent of the outstanding shares of any company covered by The Clipper Group. The Clipper Group, Inc., has no such equity holdings.

Regarding the Information in this Issue

The Clipper Group believes the information included in this report to be accurate. Data has been received from a variety of sources, which we believe to be reliable, including manufacturers, distributors, or users of the products discussed herein. The Clipper Group, Inc., cannot be held responsible for any consequential damages resulting from the application of information or opinions contained in this report.