



IT Knowledge • Business Results

Content Assisted Search

Centera Seek - Metadata Enabled Search and Index

By Brian Garrett
Analyst, Enterprise Strategy Group

May 2006

Table of Contents

Introduction..... 2
What is Centera Seek?..... 3
Centera Seek in Action..... 8
Conclusion..... 12

Introduction

The storage industry has done a good job over the past 30 years of solving the problem of securely and reliably storing digital information. Storing ever-growing amounts of information has become a relatively easy problem to solve - simply buy more storage hardware. As data center managers throw more and more storage hardware at ever increasing piles of information, the challenge has changed from "How should I protect my company's information assets?" to "How am I ever going to index and search through all this stuff?" As a result, indexing and search are hot technologies that are ripe for innovation in 2006 and beyond.

Centera Seek is an index and search solution from EMC that was released in June of 2005. Centera Seek combines active archive storage hardware from EMC with search and index software from Fast Search and Transfer (FAST). In other words, Centera Seek is an index and search engine appliance designed for an active archive. Centera Seek is a platform for the deployment of a new class of applications that the Enterprise Strategy Group (ESG) refers to as "information asset management." Information asset management applications bring order, meaning, and value to corporate assets residing in digital archives.

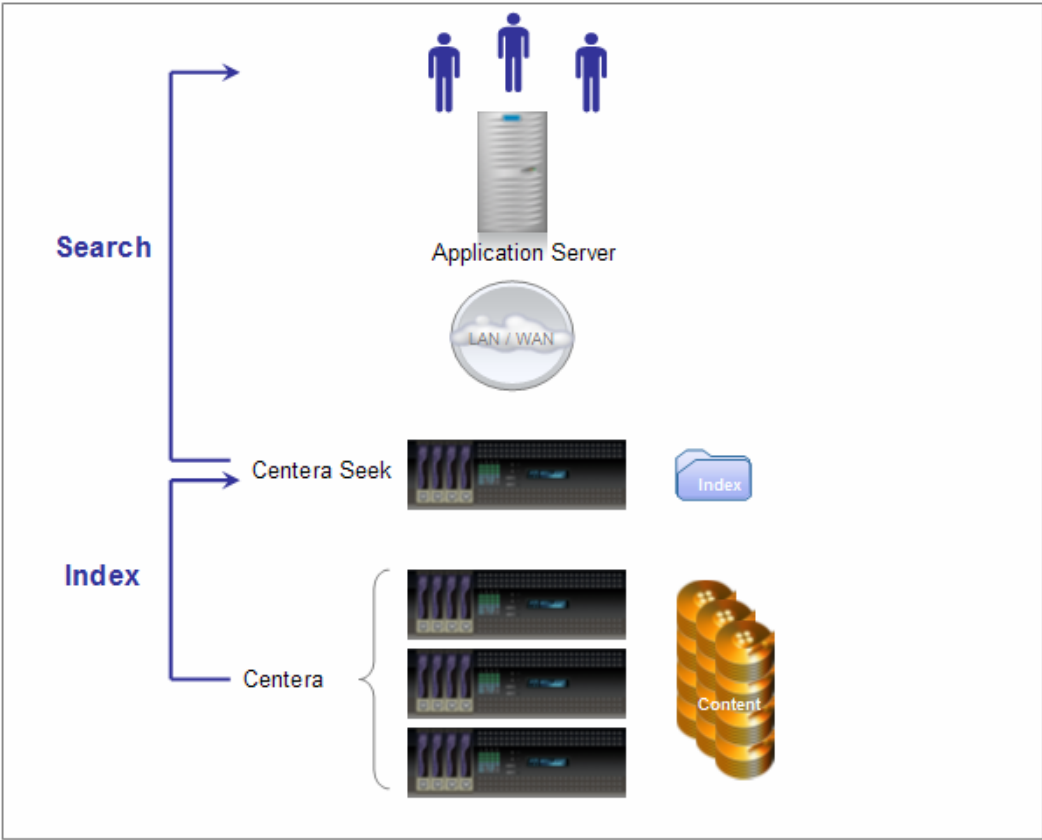
In this paper, we'll explore the value and functionality of the Centera Seek solution. We'll investigate how Centera metadata indexed by software from FAST creates a content assisted search and retrieval platform for the development of information asset management applications. We will present some examples of information asset management applications that can be built using Centera Seek. We'll look specifically at Centera Chargeback Reporter, the first application to be released that uses Centera Seek. And finally, we'll analyze the value of content assisted search and look towards a future where ever increasing information asset stores can be easily indexed, searched, and organized.

What is Centera Seek?

Centera Seek is a combination of active archive storage hardware from EMC with InStream™ search and index software from Norway-based Fast Search & Transfer (FAST) ASA. The EMC Centera is a storage system built using clusters of commodity servers, with each one full of large capacity affordable hard drives.

As shown in the following diagram, Centera Seek runs continually in the background indexing metadata on a server appliance known as the Centera Seek Appliance. Later as users and applications issue search requests, Centera Seek returns Content Addresses for data that matches the search criteria.

Figure One: Centera Seek - Centera Seek search and index



Metadata Assisted Search and Index

Centera Seek is a metadata indexing engine. Metadata, or data about data, is stored in a Centera archive in human and machine readable form (XML). Metadata is like a wrapper or a container for data. Metadata can be used to understand and classify data without needing to know how to open and read the content. Centera Seek software accesses Centera metadata and creates an index for future reference.

Automation-friendly Centera metadata is what makes Centera Seek work. Consider the challenge of accessing data that has been sitting in an archive for years, or even decades. The applications and people that placed the data in the archive may have long since departed. Centera Seek can turn this data into a searchable and

useful information asset using human readable and automation friendly metadata that is saved by the Centera as data is archived.

The operating systems, business applications, and personal productivity tools we use daily all deliver value using metadata. For example, consider the metadata associated with this report. The file name, file type, and the date it was last saved are three vital pieces of metadata associated with the report that are managed by the operating system. The application I am using to write the report enables me to associate additional metadata using a file properties dialogue box. The metadata properties that I can associate at this level include the subject, author, comments, and keywords. When I save the document, a desktop search application running on my laptop associates another set of metadata characteristics to the report. And later, when the report is rendered in PDF format on our web site, metadata will be associated, including an abstract, keywords, and a technology coverage area.

There are three types of Centera metadata. Each of these three types of metadata can be indexed and searched using Centera Seek:

1. Centera generated
2. Storage Administrator generated based on access profile or environment variables
3. Application Specific

Centera generated metadata is automatically maintained by Centera software. Like a file system, the metadata maintained in this manner includes the name of the content and the date and time of archival. Unlike most file systems, the Centera maintains one crucial additional field at this level - the retention period. A file with a defined retention period in a Centera archive can not be changed or deleted until the retention period has expired. Retention is one of the key differences between a Centera and a traditional file system. Additional Centera generated metadata fields that can be classified using Centera Seek include the total data size (data plus metadata) and the retention class.

Storage Administrator Generated metadata can be optionally wrapped around data based on an environment variable or an application/user profile. Two methods are available for storage administrators. The CENTERA_CUSTOM_METADATA environment variable lists named value pairs to be saved as metadata during each archive operation from a specific server. Access profiles defined on user or group basis are used to achieve a similar result. Examples of metadata defined at this level include a company division, a cost center, or an application server ID.

Application specific metadata is optionally added by applications integrated with the Centera API. Applications like e-mail archiving, which typically maintain metadata both internally and in databases, can now store custom metadata within the Centera using this method. For example, the application specific metadata that an e-mail archiving application adds to an archived e-mail record includes the sender, receiver, and date sent.

InStream™ Technology from FAST

Centera Seek uses industry leading core search technology from FAST Search and Transfer. FAST technology is used by some of the world's best known companies with the most demanding search problems. Companies like Dell, Reuters, and BEA use FAST technology. FAST works closely with large and well-known customers and partners who integrate InStream™ technology into services and products. As a result, extensive documentation and best-of-breed support is a given.

The Centera Seek engine is mature, scalable, reliable, flexible, and efficient. The software can be run over multiple servers if needed for scalability. Index creation is fast, and space efficient. Multi-threaded Centera Seek searches are serviced quickly in real-time.

Centera Seek indexing and search options are supported and presented through the Centera Seek Application Programming Interface (API). Applications integrate with the Centera Seek API using C++, Java, and .NET. An HTTP interface is also supported for browsing and scripting (e.g., Perl).

FAST technology brings a rich and mature arsenal of indexing and search options to the Centera Seek platform. Complex and powerful queries can be built using logical AND/OR constructs and wild card expressions. Data types including strings, dates, and numbers are supported. Further, the indexing is XML format aware so that searches can be performed to look for data within a specific XML attribute name or value.

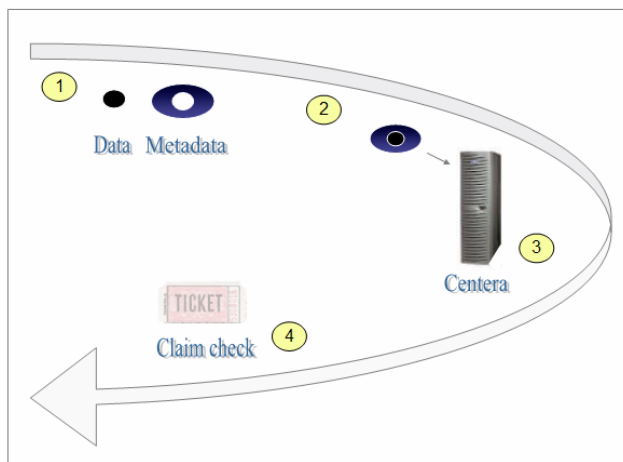
The richness of the metadata of the Centera and the flexibility of Centera Seek API can be combined to build powerful search and classification queries. For example, with Centera Seek you can:

- Find documents that have not been modified since Jan 1, 2002.
- Find all files produced by a particular application that need to be migrated to a new application.
- Escrow all data archived by a division of the company that was recently spun off.
- Find check images created before Sept 1, 1999.
- Find e-mails with an expired retention period to be purged.
- Find e-mails to be retained indefinitely due to litigation.
- Find spreadsheets that are due to expire next month.
- Classify scanned contracts produced by the legal department by division.
- Classify hosted storage archived in the past month by cost center for charge back.
- Order application servers by capacity consumed.
- Order medical records for a patient by admission date.
- Find expired .ppt and .xls files archived through a Windows interface.

How the Centera Archives Data and Metadata

The Centera is an IP-protocol connected storage system that is used to affordably store large volumes of fixed content on disk for long periods of time. To match the behavior of paper records and tapes locked in a vault, the Centera is used to ensure that content does not change for a prescribed retention period. Applications that use the Centera as an archive are freed from the responsibility of managing the location of the data and the metadata associated with each archived object. Instead, the Centera creates a unique address for each archived object and its metadata (Content Addressed Storage). The Centera then manages the storage and retrieval of archived objects and metadata in a very large and flat address space. The diagram and explanation that follow describe the Centera archival/retrieval cycle in more detail.

Figure Two: The Centera Archival Process

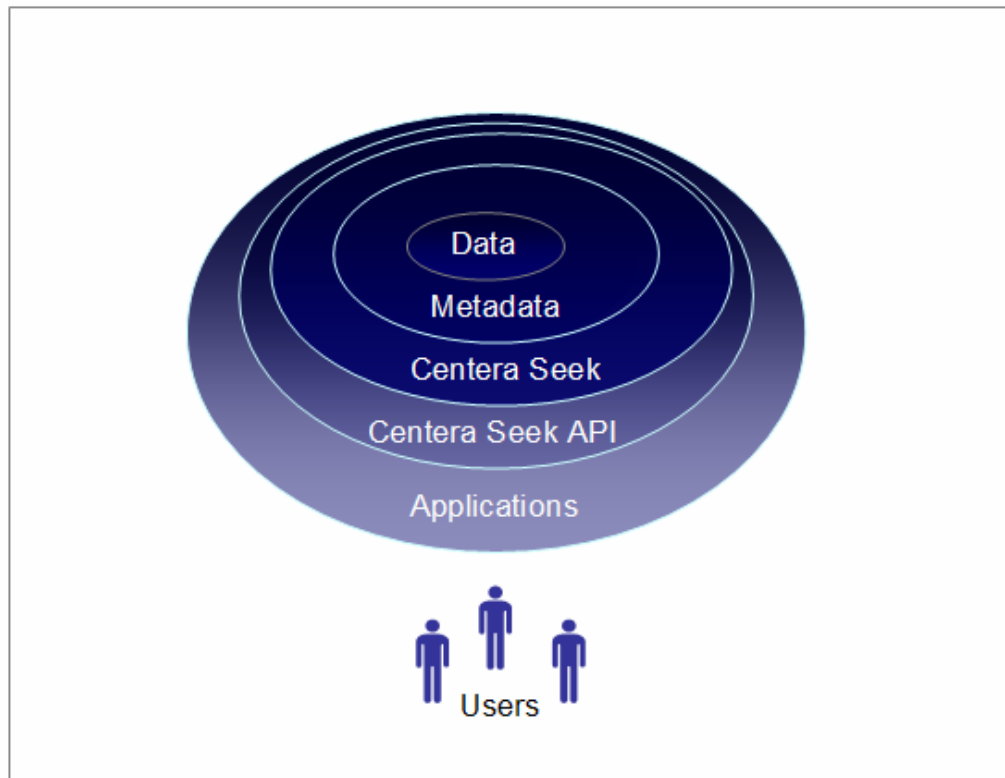


1. An application delivers a data object and an XML file containing metadata that describes that object.
2. The Centera stores mirrored copies of the data and the metadata on disk.
3. The Centera runs the data and the metadata through an algorithm that yields a claim check.
4. The Centera returns the claim check (a big number) to the application.
5. Some time later, the application retrieves a copy of the data and metadata using the claim check.

Putting it all Together

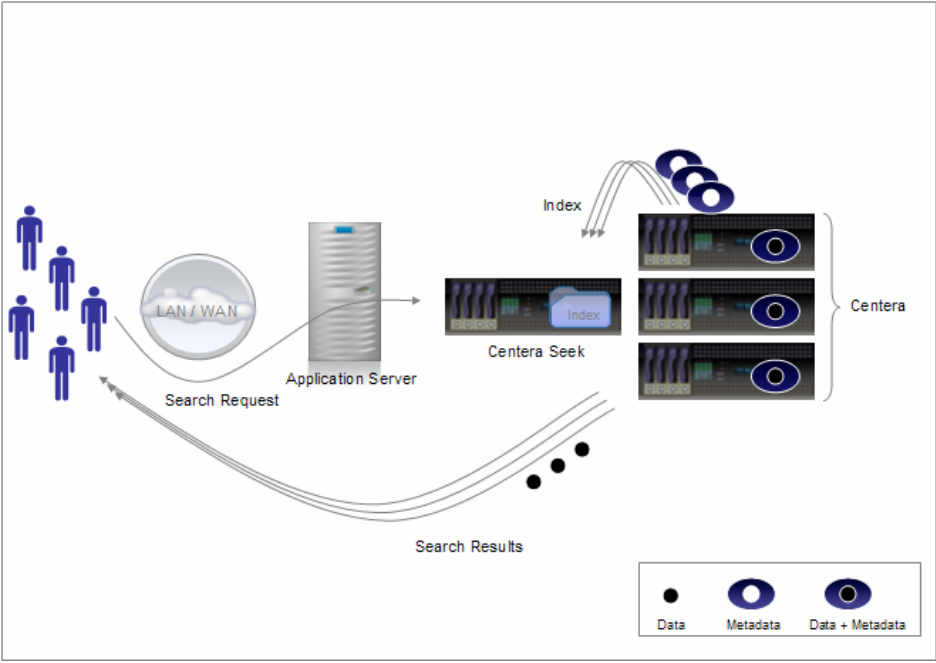
At its core, Centera Seek is built around data records archived within a Centera archive. The Centera wraps metadata around the data and assigns a tamper-proof digital signature. The Centera Seek Appliance indexes metadata and services search requests. Applications are integrated using the Centera Seek API.

Figure Three: Centera Seek - A Layered Approach



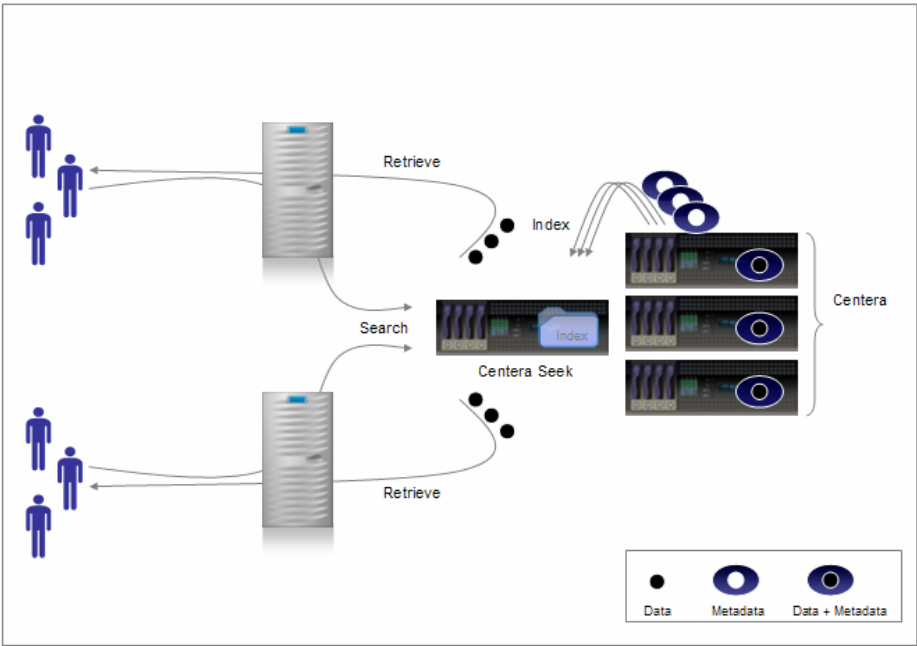
Let's take a look at how this all works when put together. As shown in Figure Four, metadata is wrapped around data and stored in a Centera cluster. The Centera Seek Appliances uses metadata to build an index. When an application server executes a search on behalf of a user, the search request is handled by Centera Seek. Centera Seek uses its index to locate the data which is returned to the users over a local or wide area network. It is important to note that the application server doesn't have to maintain the index.

Figure Four: Centera Seek - Metadata Indexed Classification and Search



Note that Centera Seek can be used to index and search Centera archived data that spans applications. Imagine the challenge of escrowing archived data created by a company division that was recently sold. Up until the sale, the Centera was used as an archive by multiple application servers. To find all of the archived data belonging to the recently divested division, a storage administrator would have to log into each of the application servers and run queries to find the data to be extracted. Using Centera Seek queries can instead be directed to a shared seek and indexing platform.

Figure Five: Shared Metadata Classification and Search



Centera Seek in Action

How Centera Seek works as explained in the previous section matters to application developers and EMC partners, but is a bit of an esoteric subject for end users. What matters to end users is the value of applications that are built on top of Centera Seek. The simplest example of an application that uses Centera Seek is the built-in HTTP interface which can be accessed using a browser as shown in the following screen shot.

In this case a query was submitted to find records associated with a cost center within a specific date range. The cost center indicates the business unit that created each archived record and can be used for chargeback purposes. Cost Center is an example of a metadata field that is generated by the Centera based on an access profile as explained previously. A pre-configured Centera access profile causes this class of metadata to be automatically created as records enter the archive. Note that 491 records were found in a quick 0.011 seconds. If a Centera Seek metadata index had not been created earlier, such a search could have taken minutes, if not hours.

Figure Six: A Browser based Centera Seek Query

The screenshot displays a web interface for submitting a search query. The query input field contains the following text:

```
AND (eclipse:@cost_center:"00105c",  
eclipse:@template_file:"/emc/regana*"  
,eclipse:@employee_id:"5000",  
creationdate:range(2006-01-01,max))
```

Below the input field is a "Submit Query" button. The search results are displayed in a blue header bar: "Results 1 - 10 of 491 (0.011seconds)". The first result is expanded, showing a tree view of metadata fields:

- 1 (EXPAND ALL) (COLLAPSE ALL)
 - eclipse
 - eclipdescription
 - meta
 - custom-meta
 - @cost_center = 00105c
 - @employee_id = 5000
 - @id = Seek v2.1 Test
 - @random = jEnsqdsmya
 - @template_file = /emc/regana/development/eclips
 - @time = 1141753631975
 - eclipcontents
 - rank = 1000
 - internalid = 5a3f2d261ec25b68c89064d754e1995f_Cente
 - contentid = CSSLL5AMK5NB2e45VMTN8Q5BIV6G41179Q
 - collection = Centera
 - creationdate = 2006-03-07T07:48:36Z
 - modificationdate = 2006-03-07T07:48:37Z
 - totalsize = 51200
 - numfiles = 1
 - retentionperiod = 0
 - modpoolid = default
 - clipca = CSSLL5AMK5NB2e45VMTN8Q5BIV6G41179QRK
 - clipsize = 5084
 - cdfpa = 16384
 - totalsizepa = 122880
 - totalstoresize = 112568

At the bottom of the results pane, there is a "2 (EXPAND ALL) (COLLAPSE ALL)" indicator.

Chargeback Reporter:

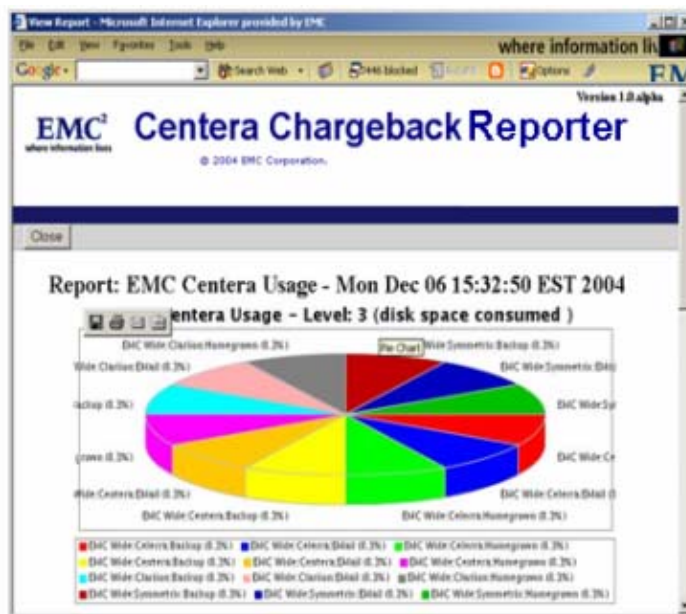
Chargeback Reporter from EMC is the first application that was built using Centera Seek. Chargeback Reporter uses metadata indexed by Centera Seek to create resource utilization graphs and reports. Canned and custom reports can be generated on a defined schedule and delivered to business units to help control the cost of archiving. Reporting capabilities include:

- Bytes written, space consumed, number of files
- By division, business unit or group
- By chargeback bucket based on policy
- By application data type (e.g. E-mails, documents, medical images)
- Usage trend analysis

Chargeback Reporter gives the user a simple user interface which can be used to map Centera metadata into categories. For example, a user can request metadata attributes with name = "Division" and value = "Anvil" to define a category called the "Anvil Division". The user can then assign defined categories to reports they have defined. Chargeback Reporter automatically issues Centera Seek queries to find all of the data within a category and reports on the resources used to archive data in that category.

A sample graph produced by Centera Chargeback Reporter is shown in the following screenshot. The pie chart shows how capacity utilization is spread amongst multiple applications sharing the same Centera. In this case, a backup to disk application, an Email archiving application and a homegrown application are sharing a single Centera. The data presented graphically by Chargeback Reporter is available in report form and can be exported for use in an enterprise-wide chargeback framework.

Figure Seven: Chargeback Reporter



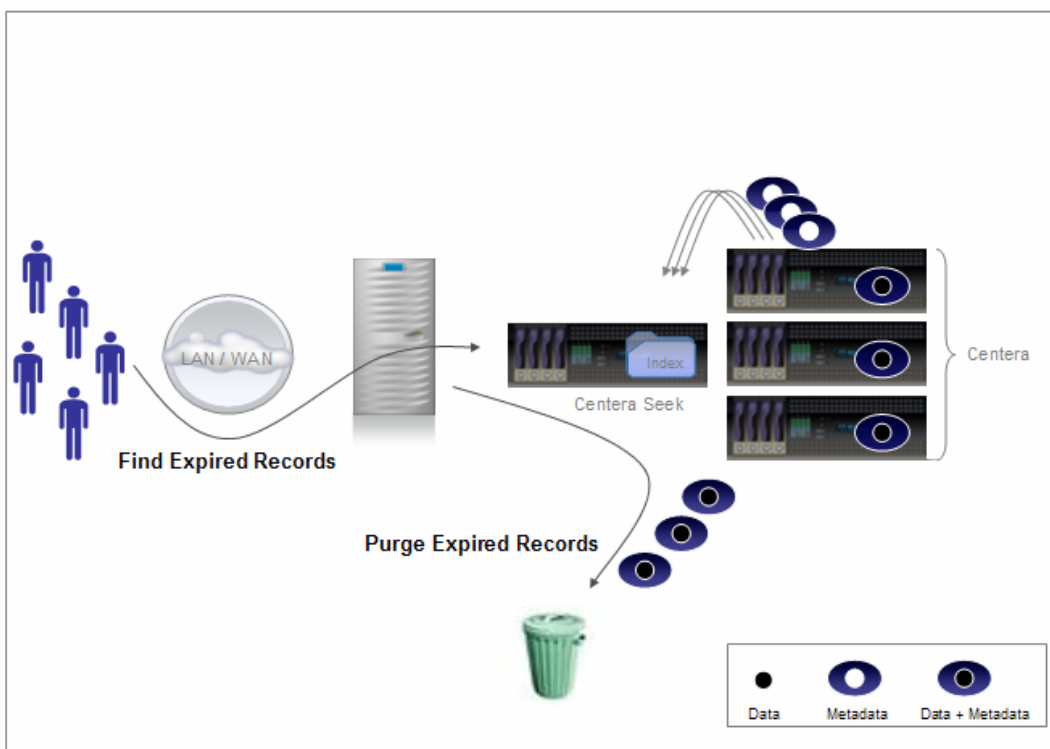
Data Grooming and Legal Holds

Data purging is a valuable and handy application that can be built using Centera Seek. Centera Archives are often purchased and deployed to deal with compliance and corporate governance initiatives. A retention period is assigned as records enter the archive. Although records can not be altered or deleted until the retention period has expired, records are not automatically deleted. Records are not automatically deleted by the Centera

by design. This is done to accommodate a review process that is often required in a compliance mandated workflow. For example, consider an Email that has expired. That Email may be needed to support an ongoing legal discovery process. In this case, a retention extension, otherwise known as a litigation hold is required.

The retention period of an archived object is maintained as metadata within a Centera. Centera Seek is used to create an index of that metadata on the Centera Seek server as shown in the following diagram. Later when data grooming is required, a user sends a request for records that have an expired retention period. The search request sent to the Centera Seek server can include complex expressions that limit the scope of the search based on additional metadata fields. For example, a query could be built to find "all records with a .pdf extension that were archived by a specific server and have since expired". Centera Seek returns a list of records to be deleted to the application server and creates an auditable report. The application can then groom the archive and reclaim space as records are deleted from the archive.

Figure Eight: Centera Seek Enables Data Grooming



Data Mining

Centera Seek can be used to isolate a small subset of a massive archive based on metadata. That subset of data can then be fed to a full content indexer "on-the-fly" enabling complex and detailed queries. By way of example, imagine your company needs to gauge the exposure from a potential leak of confidential information. You have been tasked with identifying any email or document from March of 2004 on project Tango from the Marketing department that was shared with the Tanner Company where the word "Confidential" was not used. You've got a large archive with 40TB of records to search. Centera Seek could be used to quickly isolate and index all of the emails and documents produced by the Marketing department during the specified time period. Centera Seek is used to limit the content to 20GB which can be easily fed into a full text search engine to identify how and when confidential information was leaked.

Coding to the Centera Seek API

In this section we'll walk through a through a simple example of coding to the Centera Seek API to give a sense of a what a customer needs to do to add Centera Seek in a custom application. The following steps are used to build a search query and retrieve results in XML format:

1. Connect to a Seek server (also known as a Search Factory)
2. Establish a Seek interface (also known as a Search Engine)
3. Run a search
4. Retrieve results
5. Dynamic drill-down (optional)

The code sample shown in Figure Nine illustrates the four step process listed above.

Figure Nine: Coding to the Center Seek API (Java)

```
/* MinimalSearch <host> <port> <username> <password> <query> */
import no.fast.ds.search.*;

public class MinimalSearch {

    public static void main (String[] args) {

        try {
1. → ISearchFactory searchFactory = SearchFactory.newInstance(args[0], Integer.parseInt(args[1]));
2. → IFastSearchEngine searchEngine = searchFactory.createSearchEngineForCollection("Centera");
3. → IQuery query = new Query(args[4]+"_CenteraSecurity="+args[2]+","+args[3]);
4. → IQueryResult queryResult = searchEngine.search(query);
        System.out.println(queryResult);
        }
        catch (Exception e) {
            e.printStackTrace();
        }
    }
}
```

In the first step we use the host name and port number to connect to the Query and Results (QR) factory. In the second step we specify a URL interface for performing queries using standard "http://" + host + ":" port notation. In the third step we specify the query, and in the fourth we retrieve the results. For an example of a query performed in step 3, imagine that we want to search for clips created in the month of November 2005 that have a total size of 1,000 bytes or less.

```
String nodeQuery = and (creationdate:range(2005-11-01,2005-12 01), totalsize:range(0,1000));
```

In the example above, a query string is built to search Centera **creationdate** and **totalsize** attributes. A variety of attributes besides creation date and total size can be searched using Centera Seek. Additional searchable attributes include the retention period, modification date and content address. Using Boolean expression support (AND/OR operators) and optional drill-down capabilities, the Centera Seek API can be used to build custom applications that benefit from powerful and fast-running queries.

Conclusion

Storage administrators are drowning in a sea of information. Faced with huge volumes of archived data spanning multiple applications that must be maintained over long periods of time, storage administrators can now turn to Centera Seek for help. No longer adrift in a sea of data, administrators can use Centera Seek to simply accomplish tasks like escrowing data belonging to a recently divested division, or expunging data from a retired application.

When the Centera was first released in 2002, customers typically used the Centera for a single application. Since then customers have begun to deploy more than one application on the same Centera system. Leveraging the metadata search capabilities of Centera Seek, customers now have an excellent tool which can be used to track, classify and manage data assets archived by multiple applications sharing the same Centera asset.

Information archived for very long periods of time cannot outlive the technology used to store and process that information. Hundreds of EMC partners who have embraced the Centera understand this issue and appreciate the future-proof benefits of the Centera architecture. Freed from the chore of maintaining file names, paths, and metadata, applications access content using a digital signature that will remain the same forever. Using clusters of commodity servers and affordably dense ATA drives, the Centera can easily be migrated and upgraded to take advantage of the latest hardware advances.

Centera Seek maintains that tradition and extends future-proof benefits to include metadata search and classification. Centera Seek runs on the same hardware as Centera. Like the Centera, Centera Seek is designed to scale and grow while taking advantage of the latest hardware advances. Future-proof access to existing indexes and queries is guaranteed through the use of the Centera Seek API.

Application developers should consider Centera Seek as a platform for adding value to their existing products. Centera Seek provides scalable and predictable search and index of Centera metadata. Metadata indexing and search can be turbo charged. Data classification can be treated as a background task. Indexes maintained on multiple application servers can be consolidated. Put simply - Centera Seek provides application developers with a worry-free metadata indexing and classification engine.

The storage industry has done a good job over the past 30 years of solving the problem of securely and reliably storing digital information. The next frontier for the storage industry is a single self-managed system that can classify and search itself. Centera Seek is a storage platform that tracks, searches, and classifies metadata associated with archived data. ESG believes it makes sense for EMC to extend the Centera search capability beyond metadata to include the search of core data content. For example, Centera Seek can currently be used to search for e-mails that were archived before a certain date, or have an expired retention period. With core data content search capabilities, the same Centera Seek platform could be used to offload searches for strings within the body of e-mails. Looking towards a future enriched by content assisted search, we see the Centera becoming a self-descriptive and self-searching application development platform optimized for the fast classification, search, and retrieval of digital corporate assets.

All trademark names are property of their respective companies. Information contained in this publication has been obtained by sources The Enterprise Strategy Group (ESG) considers to be reliable but is not warranted by ESG. This publication may contain opinions of ESG, which are subject to change from time to time. This publication is copyrighted by The Enterprise Strategy Group, Inc. and was sponsored by EMC. Any reproduction or redistribution of this publication, in whole or in part, whether in hard-copy format, electronically, or otherwise to persons not authorized to receive it, without the express consent of the Enterprise Strategy Group, Inc., is in violation of U.S. Copyright law and will be subject to an action for civil damages and, if applicable, criminal prosecution. Should you have any questions, please contact ESG Client Relations at (508) 482.0188.