

The Process

Identify

1. Enterprises use standard backup applications to backup business critical data to the EMC Data Domain platform.
2. As new backups are created, Index Engines will crawl the backup images via NFS or CIFS to perform full content and metadata indexing of the unstructured content within the backup image.

Collect

1. Within the Collection Engine queries can be defined which represent the corporate data retention policies. The policy is based on file and email metadata as well as content. Policies are defined and stored in the Collection Engine using the comprehensive search parameters.
2. Index Engines runs the stored policies and automatically performs the query on the indexed backup images. The relevant results are extracted from the backup image.

Manage

1. Extracted data is placed in the Index Engines Collection and written back as an image to the Data Domain system. This image represents the unstructured data that is required for long-term storage.
2. Run the collection process as new backup images are generated to apply policies to the current Collection.
3. Query the collection to find and extract specific files and emails as needed. All extracted content will be retrieved in its native format.
4. Data in the collection can be expired according to policy ensuring that only the required content will be saved.

Leverage Backups for Information Governance

Many organizations have implemented a data protection process by backing up to EMC Data Domain deduplication storage systems. As these backups age past their disaster recovery life cycle it is impossible to know what specific backup data needs to be saved and what can be purged according to compliance, legal and retention policies.

The Collection Engine enables EMC Data Domain users to filter down the volume of data that is archived by providing detailed knowledge of the content within the backup images. Understanding this content, collecting what is useful and making it available for compliance and litigation purposes reduces the volume of archived data and eliminates the data that needs to be archived for long-term retention, such as on backup tapes.

The Index Engines Collection Engine for EMC Data Domain is purpose-designed to index the content of backup images (e.g. NetBackup, NetWorker, etc...) so that they can be searched and analyzed for business relevance. These searches can be high-level metadata such as user mailboxes, or detailed queries based on file or email content, location and date ranges. Once the relevant set of data is identified within the backup image, it can be extracted into a Collection image and written back to the EMC Data Domain platform. This allows a small subset, typically less than 5% of the backup data, to be retained for long-term access and also take advantage of Data Domains deduplication technology.

The Collection Engine works with the existing backup process and automatically extracts specific files and email into a repository that is used by the legal and compliance teams who need access to these business records.

Implementation

