

un-structured data: Computerized information that does not have a data structure (i.e., not within a database)

Managing Unstructured Data

By Johnnie Konstantas

Year after year businesses, organizations and enterprises create an endless array of data files, some extremely sensitive, some completely mundane. Managing this unstructured data can be formidable, but the steps discussed here can help make the process manageable.

Year after year businesses, organizations and enterprises create an endless array of data files, some extremely sensitive, some completely mundane. In the form of documents, images, spreadsheets, email messages, presentations, multi-media files, etc., this unstructured data is stored throughout IT systems on file servers. Some files are accessed daily, some annually, some completely forgotten, and managing all this data can become formidable.

Efforts to tame this unstructured file share data abound. IT departments are tasked with managing all this data, keeping it available to those who need it, keeping it away from those who would abuse it, and keeping it within regulatory and compliance requirements:

- **Risk management and loss prevention** – Classifying information in order to identify content that is sensitive
- **Data entitlement management** – Document and enforce a scalable and repeatable process for determining who gets access to what data
- **Hierarchical storage management and data migration** – Moving stale or outdated data from file servers to cheaper near-line or offline storage solutions

There are even initiatives to impose structure on unstructured data by migrating it to document management systems like Microsoft's SharePoint or EMC's Documentum.

Most enterprises undertaking projects to manage their unstructured data have discovered their efforts to be extremely slow and costly in terms of time and resources required. Although technologies for classifying, protecting and moving data exist, successful implementation may require months, if not years, to accomplish, depending on the amount of unstructured data.

Data indexing – a technical procedure that conducts a deep examination of file server contents for the purpose of creating a logical ordering of the unstructured data therein.

Before an organization can even begin to manage all its unstructured data, it must know what it has: various products for data classification, eDiscovery, loss prevention and content management require that file share contents be indexed as a first step. In order to get perspective on the process, consider the time it takes to “crawl,” or index, data. A typical medium-sized enterprise has about 10 terabytes of data (10,000 gigabytes). Indexing time takes about two hours per gigabyte – about 2.3 years for the medium-size company! And unstructured data grows by more than 50% annually.

Naturally, given the length of time to complete, most enterprises understand that data classification, migration, and protection projects have to be rolled out in phases. The challenge is knowing where to start. Because file share data is unstructured, there is no way to discern which is important and which is not. Think of all the files on your personal desktop computer you have been saving *and ignoring* for years and how much time and effort would be involved viewing each file to determine whether you want to keep it or not. Given enough storage space, you will probably save it for another day.

There is a way, however, to increase the efficiency and accuracy of these efforts and shave off months or even years in implementation. Outlined below are ten key requirements for any effort, system or technology whose purpose is to classify, migrate, protect or otherwise manage business data – specifically the documents, presentations, spreadsheets, scanned images, multi-media files, etc., that fill file servers and form any enterprise's valued assets. The key is to begin by making the ten activities listed below prerequisite to any project for unstructured data management and protection.

1. Create an inventory of file share contents
2. Remove overly permissive access permissions
3. Remove global groups
4. Remove unused accounts

5. Identify orphan data
6. Identify stale data
7. Identify infrequently accessed data
8. Identify highly active data
9. Identify business owners
10. Repeat above activities monthly to accommodate change

The discussion that follows elaborates on these required activities and explains their benefit in expediting and scaling data management.

1. Inventory file share contents

Most enterprises do not exactly know the contents of their file systems. Each user in an organization has his own partition or space on the file share, and generally uses it at his discretion. Access permissions for this data are also hard to track. A folder may be available in a limited fashion to a handful of users, but as needs for the data therein increase, permissions are changed to include many groups or the whole company. The net effect is that the IT group and file server administrators do not have an accurate picture of what is contained on the file servers. Step one, then, is to create an inventory of the data and its permissions, as well as the list of persons with access. This will help guide the next steps in protecting and managing it.

The inventory of the file server contents must include:

- All users, including their group memberships, Active Directory attributes and data permissions
- All folders and sub folders within a file server, as well as the Microsoft NTFS permissions to this folder for any user or user group who is part of the domain
- Filtered views that allow queries based on user name, group name or folder/data name
- Automated updating of views to reflect changes or new data within Active Directory (i.e., user-to-group membership) as well as within the file server (i.e., new data, deleted data, renamed data)

2. Remove overly permissive access

Most enterprises have very efficient processes for granting access permissions to data, but few revoke those permissions when the need has passed. As a result, most access to file share data is unwarranted and the permissions are dated.

In part to address this, and in part to achieve the broader goal of stemming the dissemination of critical information to unauthorized persons or to those outside the company, some enterprises undertake data loss prevention (DLP) projects. These projects normally start with technology that looks to create an index of unstructured data and subsequently to classify it for the purpose of identifying sensitive and valuable information. The challenge with this approach, however, is that because indexing and classification take so long, the

risk of data loss is incurred every day that the project is in its implementation phase.

The solution is to first start with a broad reduction of access privileges so as to limit access to only the users who have a business need for the data. This step dramatically reduces the probability of data loss and can be conducted in a fraction of the time it takes to index and classify information. By revoking permissions as a first step prior to a DLP project, enterprises can reduce the exposure in the interim time frame while the DLP project is being scoped and rolled out. The process for revoking permissions to data should be automated and include:

- Identifying the names of those persons who no longer need access
- Identifying the data sets to which those persons no longer need access
- Centralizing dissemination of the permissions revocation to the live environment
- Recording permissions pre- and post-revocation as part of a change report

3. Remove global groups

Data access “permission creep” is quite common. In fact nearly 100% of organizations can identify some files or folders where permissions for access are overly liberal. As file share contents grow and individuals change roles, their business needs for data grow. As a result, IT operations personnel are forced to open file access controls for broader and broader data availability. In some cases, global groups are assigned. These are groups that according to Microsoft Active Directory nomenclature include a very large percentage of the organization’s user population. The Everyone group is one such designation. As its name suggests, when the Everyone group is assigned to a folder, it makes the data within available to everyone. With the right solutions and technology in place, removing global group access permissions and replacing them with more granular access controls is something that can be done fairly quickly, and serves to dramatically reduce the probability of data loss by restricting access to business-need-to-know. Any project for data loss prevention will benefit from global group removal as a first step. The process should include:

- Identifying folders with global group assignments
- Identifying individuals who require access to those folders that have global group assignment
- Removing global groups
- Assigning individual permissions
- Recording the revocations

4. Remove unused user accounts

As with unwanted data permissions, enterprises often find it difficult to keep track of accounts within user repositories. As individuals leave a company, change roles or move within and

across organizations, their account types need to be changed or revoked. This updating, although seemingly useful, does not take place as a matter of course in most enterprises. It is, however, a fairly simple and quick way to reduce the risk of exposure to data loss by removing unnecessary access privileges. This activity also increases the accuracy of data-use monitoring by ensuring that the accounts are not co-opted for use by persons other than the rightful users. Removing unused user accounts from Active Directory should include:

- Identifying inactive accounts
- Verifying systematically that they are not in use
- Conducting the revocation from a central location

5. Identify orphan data

Projects to migrate data to different tiers of storage (i.e., in-line, near-line, off-line) necessitate that information be segregated into that which must be readily available versus data which is of less critical importance and can be archived. Typically, such projects can take a very long time to complete, especially if data indexing and classification are the approaches applied to the task. Data migration can be made less arduous and much faster by analyzing how data is used and identifying that data which has no known owner (i.e., has not been accessed by anyone). File share data without an owner is considered orphaned and is a very likely candidate for offline storage. Completing this task should not require classification or indexing and should, in fact, be a precursor to both. Orphaned data, once identified, can be examined for content type at a later date. However, because it is not in active use, the urgency of this task decreases dramatically.

6. Identify stale data

As with orphan data, identification of data that has not been accessed for several months to a year is a very efficient way to group data sets as good candidates for offline storage or deletion. This requirement can be completed fairly quickly and should not require use of indexing or classification technologies.

7. Identify infrequently accessed data

As with orphaned and stale data, some file server contents are so infrequently accessed that they may be good candidates for near-line storage. It is important to understand your organization's data requirements and what are good metrics for defining frequency of access. Completing this task, however, need not take a long time. Again, the approach used does not need to make use of classification or indexing. Rather, in order to identify infrequently accessed data, a data-use audit can be used. The most effective data-use audits include the ability to sort the information by time interval, event type (i.e., open, delete, rename, create), or frequency count, for example.

8. Identify highly active data

Constituent to the requirements outlined above is the ability to identify highly active data. Applying the principle from infrequently accessed data in the converse, the data access audit should be able to be sorted to specify a time period and an event count, and then quickly zero in on the most actively accessed files and folders on a share or shares. This most actively accessed data is implicitly of high business importance and is not only a good candidate for in-line storage, but for indexing, classification and loss prevention projects. By completing the identification of the most active data sets prior to any such projects, IT operations personnel can shorten the project length, focusing their efforts on the most important business information first.

9. Identify data business owners

For all of the requirements of proper data management discussed, it is important to note that having a list of the data business owners can markedly increase the accuracy and efficiency of each requirement. By consulting with business owners, the administrators of unstructured information can ensure that permissions revocations, data migrations and access controls are commensurate with business needs and company policies. IT operations personnel should be able to generate a list of data business owners for any given data set at will. Business owner identification should be capable of being completed "on demand," given that this information spans many projects and needs.

10. Repeat steps for scale and change

Since unstructured data is not only the most voluminous, but the fastest growing data type within organizations, it is important that enterprises set up processes by which these requirements for data management can be applied and followed at regular intervals with consistent results.

Conclusion

Getting control of unstructured data is an imperative for companies, but IT operations personnel have been challenged on how to get started. As noted, there are products in the marketplace that focus on the business content and context of data. However, their implementation is very complex and, as a result, lengthy and costly. The requirements outlined above enable businesses to realize the benefits of these products in a pragmatic way. Meeting them as a first step ensures data security is addressed and that resources are applied with the greatest business impact.

About the Author

Johnnie Konstantas, vice president of marketing for Varonis, has more than 14 years experience in the network-security and telecommunications fields, having held various senior-level roles in marketing, product management and engineering. She may be reached at jkonstantas@varonis.com.