Identify, Analyze, and Move

Applying BA Insight's AutoClassifier to Content Migration and Divestiture Scenarios

October 2016

Overview

Today's enterprises are struggling with an explosion of unstructured content across multiple systems, including huge 'dusty file shares' as well as a variety of content management systems, collaboration and social networking systems, CRM and ERP Systems, and Enterprise File Sync and Share products. It can be extremely difficult to know what you have, and even more difficult to manage risk and compliance and navigate sensitive transactions effectively. Within this challenging landscape, there are many different problems and scenarios– which we have found can be addressed with a relatively small number of solution patterns.

This application note covers a pattern we refer to as "Identify-Analyze-Move" where a complex set of content must be separated accurately in a short timeframe.   We describe the pattern and then outline how it can be applied in two specific scenarios:

- Content Migration
- Corporate Divestiture

One specific project of each type is outlined in this application note to illustrate the process and provide concrete performance characteristics.

In the pattern and each of the scenarios, we show how BA Insight's product portfolio can solve difficult problems at large scale. BA Insight's AutoClassifier is used in these solutions to apply sophisticated text analytics in processing large amounts of content in complex environments quickly. In the Identify-Analyze-Move pattern, this is used along with human supervision to provide the extremely high accuracy required by these demanding scenarios.

Contents

Background: BA Insight Product Portfolio

BA Insight's products enable sophisticated search-driven applications that include deep text analytics technology. Our product portfolio includes:

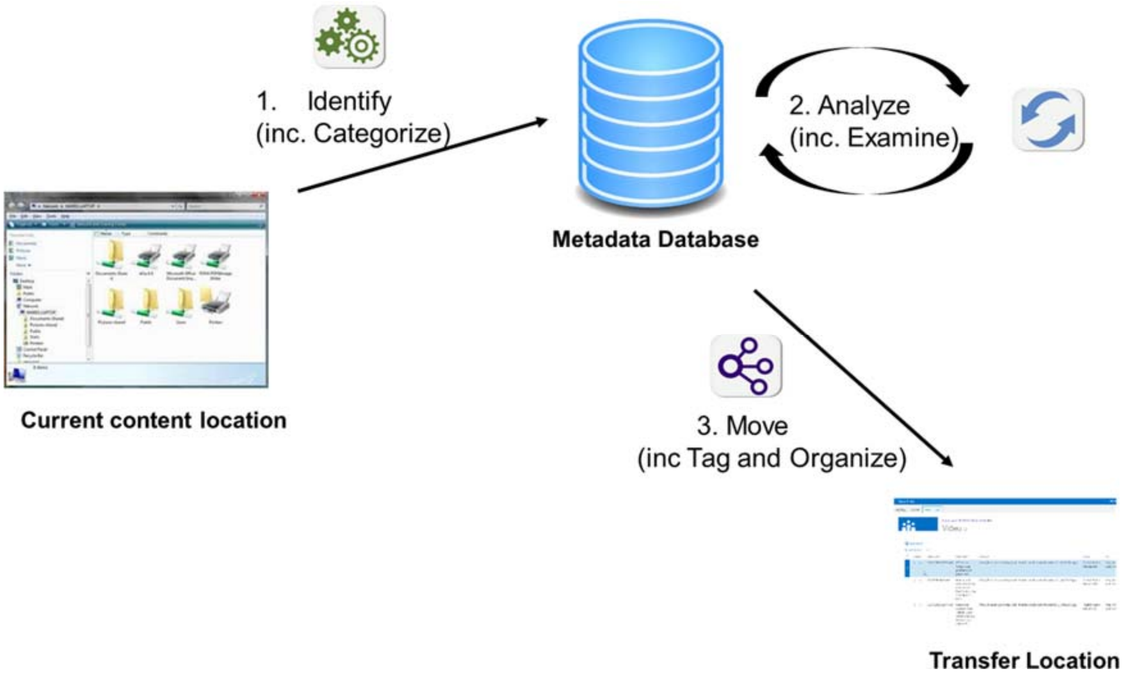
- **AutoClassifier** increases findability using auto-tagging, metadata generation, and text analytics for content within and outside of SharePoint.
- **Connectors** provide secure connectivity to over 60 enterprise systems, enabling unified views of all knowledge assets.
- **Smart Previews** extends search to include information inside documents with hit mapping, hit highlighting, and contextual actions to find results quickly.
- **Visual Refiners** enhance user experiences with drill down and "search within" capabilities.
- **Expertise Locator** goes beyond profile-based findings by combining profiles with experience and contributions relative to content in systems, using an internet-like comparison model to compare experts.
- **User-Generated InfoSites** reduces information overload by dynamically generating pages and digests with topical information without coding or IT involvement.
- **Smart Analytics** delivers valuable data about SharePoint usage, respecting international privacy laws, to enable intelligent decisions to be made about improving intranets, increasing adoption.

These are architected as modular building blocks that can be used in a wide variety of applications, without the need for custom code. For the pattern and scenarios in this application note, four of these products are used: AutoClassifier, Connectors, Smart Previews, and Expertise Locator.

More information about our products can be found on our web site at http://bainsight.com/software-portfolio

Three Phase Pattern: Identify-Analyze-Move

By finding all the files within one or more system, automatically analyzing them, and capturing the resulting metadata in file manifest (list of known files with their characteristics and derived metadata), a number of common scenarios can be addressed. No text analytics technology can provide perfect results on unknown content without supervision, so an ability to further examine and analyze the content with human analysts can greatly enhance the process, while training and tuning the automatic processing at the same time. Often the desired action is transfer of a selected set of files to a new system. In combination, we call this approach the Identify-Analyze-Move pattern, which is shown in the diagram below.



In this pattern, there are three phases:

1. **Identify:** crawl everything to enumerate contents, capture metadata, and **categorize** - but don't index any content. Using a search crawler is a great way to enumerate all the content you have, even in systems that have no built-in reporting or export facilities. Instead of indexing the content, we apply text analytics to each file and then write everything known about the file into the file manifest. This is also called a content inventory, and can be useful in its own right. The manifest includes all original and derived metadata as well as the output and confidence levels of the business rules, allowing for auditing and analysis prior to any disposition of the files.
2. **Analyze:** based on the extracted and machine-generated metadata, determine what can be decided upon with known business rules and what requires deeper **examination**. Then crawl and index that subset of content (items that aren't covered by these business rules with high confidence) for manual exploration and analysis in a specialized search-based application. The analysis captures human judgements about specific files, and also can train or tune the automatic text analytics processing through additional rounds, until all files are accurately categorized. Providing for a human in the loop ensures that exceptions and outliers can be accurately analyzed without biasing or overcomplicating the core text analytics processing.
3. **Move:** crawl selected content, further **tag and organize** this content, and move it to a target location for transfer. This often is a multi-way decision, such as "keep, move, delete" or "transfer with divested assets, keep as protected IP, archive, delete".

These phases can each have multiple iterations with some incremental changes to rules and text analytics configuration during the phase. For example, in the identification phase each system is often enumerated independently, and additional text analytics processing might be applied and tuned in the course of a few iterations. In the analysis phase, interactive inspection and classification can result in improved rules that are simply rerun; if a new business requirement arises, that is handled through an additional iteration. In the move phase, it is common to have multiple iterations or sub-phases as well. For example, one set of content may be moved to a transfer location and then another set might be archived or deleted.

Solution Components

In BA Insight's solution for this pattern, several of BA Insight's products are applied:

- The heart of the solution is BA Insight's AutoClassifier, which generates and extends managed metadata and provides a collection of powerful text analytic capabilities in the Identify and Analyze phases.
- BA Insight's Connectors are applied to tap into a wide variety of systems in all three phases.
- For the small fraction of content that requires manual examination in the Analyze phase, BA Insight's Visual Refiners and Smart Previews are used in a search center configured for reviewers.

In addition to BA Insight's products, there are several other components used in this solution:

- A metadata database (typically a SQL Server cluster) is used to hold a file manifest and to provide analysis and reporting in all three phases. The file manifest holds metadata for every file identified - including classification output and URL as well as all metadata from the source system.
- A transfer location (typically a SharePoint Site Collection) is used as the destination of the move phase.
- A search engine (typically the one built into SharePoint Server) is used to index the files that require manual examination during the Analyze phase.

Application: Smart Content Migration

Migration is Complex and Risky

Content Migration projects have gained a reputation for being late, over budget, and disruptive – both to the organization and to its customers. This reputation is well-deserved. Bloor Research reports that:

- 84% of data migration projects fail
- 72% of organizations delay migration because it is too risky

Identifying and Organizing Content is Hard

It is common to have “dusty file shares” with unknown contents, or legacy repositories with many Terabytes of poorly-structured content. Some organizations run into Petabytes, and the rate of new content acquisition and generation is accelerating. There are high penalties for non-compliance and privacy breaches. Document retention policies are not simple either – they depend on the contents of the document as well as the date and location.

Most enterprises have content spread across multiple legacy ECM systems that they are working to retire. Visibility on what is housed there varies, depending on the system. Each system also has a different security schema and access method, so extraction is not as simple as it may seem.

Smart Content Migration

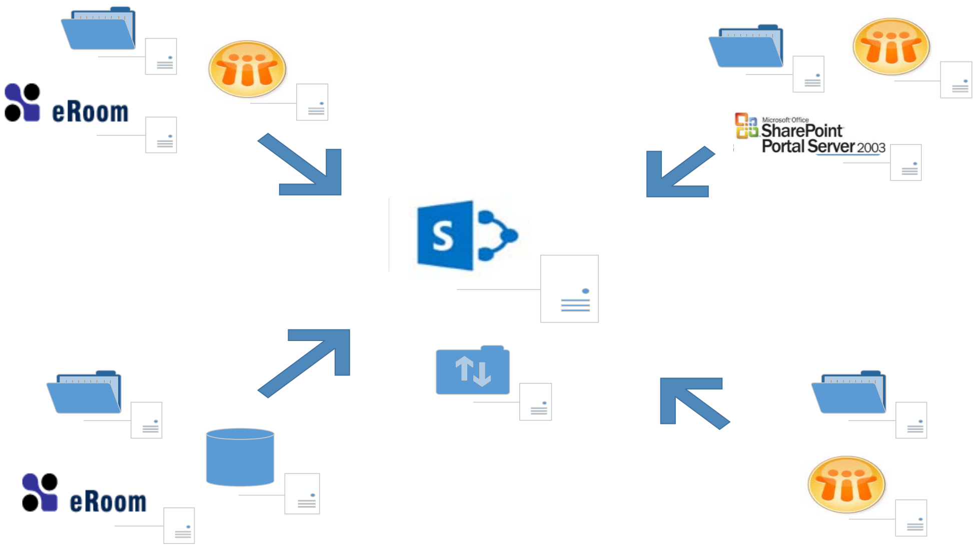
Migration projects are an opportunity to do more than “lift and shift”. It makes sense to archive or get rid of things you don’t need rather than move them, and to organize things the way you want them as they are moved. Though this may take some time and attention, it is well worth it. And by applying automation via text analytics with the Identify-Analyze-Move pattern, the time and effort is much less and the savings in the long term are significant.

It is worth noting that using search as an access layer can insulate users from the transition involved in a content migration project. This is another area in which technology can help organizations doing migration.

The Smart Migration Scenario

To illustrate how to apply the Identify-Analyze-Move pattern to smart migration, we’ll describe a specific project. This is based on real customer projects, but some details are abstracted and names are omitted, in order to not disclose our customer’s confidential information.

Example Environment: 4 sites, 11 systems



The goal was selective migration of a large volume (approximately 89 TB) of content into two systems: a centralized SharePoint 2013 farm and an archiving storage system. The content was resident at four different locations, on a variety of file shares, two instances of EMC eRoom, and three instances of IBM Lotus Notes. In addition, there was a legacy database with important records at one location, and a legacy SharePoint 2003 farm at another. IT wanted to retire all of these systems and standardize on two up-to-date centralized systems.

Smart Migration and Business Criteria

It is clear that this project is complex and that the organization would benefit from a smart migration in which they move active content to SharePoint, organized according to their overall information architecture and tagged with managed metadata based on their established taxonomies. Inactive content they want to retain should end up in the archiving system, and they wanted to retire (delete) the rest.

How did they identify which content is which and decide what goes where? Some decisions could be made simply using file locations and dates. In most other cases there was a clear business rule, such as:

- documents relating to active customers, projects, or products should be considered active
- documents relating to inactive customers or projects should be archived unless they are more than 7 years older than the last activity date
- any documents potentially related to several active litigation areas need to be kept
- there are a set of subjects that are considered part of the core history; documents related to these should be identified and archived

However, some decisions require looking at the content. A small fraction documents needed to be looked at individually and explored as a group.

Time pressure on one of these projects was high, because they wanted to complete all migration in a short time window and decommission several systems before their maintenance came up for renewal. By using the AutoClassifier and the Identify-Analyze-Move approach, they were able to complete the project, including decommissioning these systems, within 5 weeks.

Along with migrating content, there was a mandate to clean up. In this project, IT was concerned that they were out of compliance with their records retention policy, so they included a scan to check for this. Similarly, they wanted to check for proper use of content types in SharePoint and correct security rights. Finally, tagging was applied to all content based on named entity extraction and rules-based auto-tagging so that content in the new system was fully tagged and organized.

Key Capabilities from the BA Insight Portfolio

There are many capabilities of the BA Insight portfolio that applied to this solution, but the most notable are:

- AutoClassifier (used in the Identify and Move phases)
- Notes, eRoom, and SQL connectors (used in all phases)
- SQL connector with associated crawls (used in the Analyze and Move phase)
- Smart Pipeline to choreograph metadata and insert into SQL (used in all phases)
- Target for SharePoint libraries (used in the Move phase)
- Visual Refiners for exploration and analysis of indexed content as a set (used in the Analyze phase)
- Smart Previews to aid in inspection and analysis of individual documents (used in the Analyze phase)

The SharePoint file system crawler was used to enumerate and crawl content from the file shares, and SQL Server Management Studio was used for analysis and recording content disposition decisions. SharePoint Server was used as the search engine for the exploration and analysis center.

Assembling Taxonomies and Rules

This organization used several different taxonomies to classify files for migration and to check files for compliance. There was no existing corporate-wide taxonomy, so a commercial “starter taxonomy” was licensed from WAND. Two departmental taxonomies and a list of active and inactive projects already existed and were used directly. Two additional taxonomies were created for the smart migration effort:

- Customers: a list of customers (organizations and their divisions and departments, along with the last activity date) was created based on CRM system records.
- Products: a list of current and retired products was created from data in the CRM, finance, and PLM systems.

Rules based on these taxonomies were generated automatically; some were manually edited during the project. Particular focus went to the rules that indicated non-compliance with retention and security policies.

Content Volume and Throughput

This project had a large volume of content (approximately 89 TB) and at the beginning of the project there was no idea of how many files were involved or how many should be migrated. The Identify phase discovered roughly 37 million files, spread across the 11 systems. The largest crawl contained around 5 million files and took about 40 hours. For some crawls, the limiting factor was the source system, for others it was the network bandwidth.

Organizing While Moving

The final “move” phase used the target feature to copy content to a SharePoint library, along with its metadata and security settings. One transfer library was used for each source system.

In some cases, additional classification was done in this phase. For example, if the file path had been sufficient for a disposition decision in the first phase, the body of the document may not have been fetched or categorized until the Move phase.

The content types that were automatically assigned were attached as a metadata field rather than being associated with a master content type. Target folder structures were also represented by metadata rather than having subfolders in the transfer libraries. This made analysis much simpler and allowed a direct explanation of the transfer libraries in a datasheet view. The SharePoint Content Organizer was then used to move content to specific subdirectories and to associate the content type with the master within the farm.

There were a few groups that needed to continue to work in the source systems for a period of time beyond the conclusion of the content migration project. The new content from these groups was crawled on an ongoing basis, using incremental crawls, tagged and mapped automatically, and transferred directly to the appropriate destination in the centralized SharePoint farm. Since this process kept the source system and SharePoint in sync, the content was available to all other users through SharePoint. The “straggler” groups used the new SharePoint farm for all content except their currently active project files. When they closed the project, their transition was seamless. Once all “stragglers” were done (roughly 10 days after the bulk of users and content were transitioned), the source systems were decommissioned.

Tailored User Experience for Exploring and Examining Items

A SharePoint Search center augmented with BA Insight products was used in exploring and examining the leftover set. The ‘demographics’ of the set was easy to explore using Visual Refiners, for example looking at the files associated with a selected set of customers and a selected set of products. The Export to Excel feature was used to capture a slice of the leftover set which could then be changed to a different disposition in the metadata database.

Smart Previews were used to quickly examine specific documents and search within them, using the hit map feature to understand the nature of the contents in the document. The Content Assembly feature was used to create excerpts emailed to people within the business group for comment of questions. In this way, the leftover set was successively reduced to zero, meaning the disposition of all content was decided.

Results: Reduced Content Volume, Well Organized

The final result was a set of active files contained in a single centralized SharePoint farm, and another set of inactive files that were important to retain, which were stored in an archive system. These were respectively 14.8 TB and 3.8 TB in size. The small size of the archive may seem surprising, but during the analyze phase it became clear that a substantial number of files were past their retention time, and another large number were duplicates. The resulting SharePoint farm had 9.3 million content items amounting to 14.8 TB, about 17% of the original size. It also had known consistent metadata and information architecture. The net result was a greatly improved content management, governance, and search experience for the end users – at a lower cost and with lower ongoing effort from IT.

Application: Divestiture

Corporate Divestitures Include Sensitive Information Assets

Divestitures are complicated, sometimes counter-intuitive transactions. The information assets associated with a divestiture are essential to several activities:

- analysis, valuation, and planning of a divestiture requires knowledge that is often contained within these information assets
- assessing the value of information assets is part of the overall valuation process
- regulatory, competitive, and security considerations can dictate the handling and disposition of information assets
- managing the IT aspects of a divestiture includes separation and migration of information assets, as well as integration of these assets into the new environment

IP Protection is an Important Concern

Divestiture scenarios pose all the same challenges discussed for Smart Content Migration and more. IP considerations and the nature of large corporate financial transactions bring in additional challenges. Information assets must be identified and analysed with particular accuracy because the intellectual property must be separated and transferred along with other assets. In addition, it is important that unrelated corporate IP and content remains with the original organization. Ensuring that other content NOT be transferred means that the classification must have high accuracy in both directions.

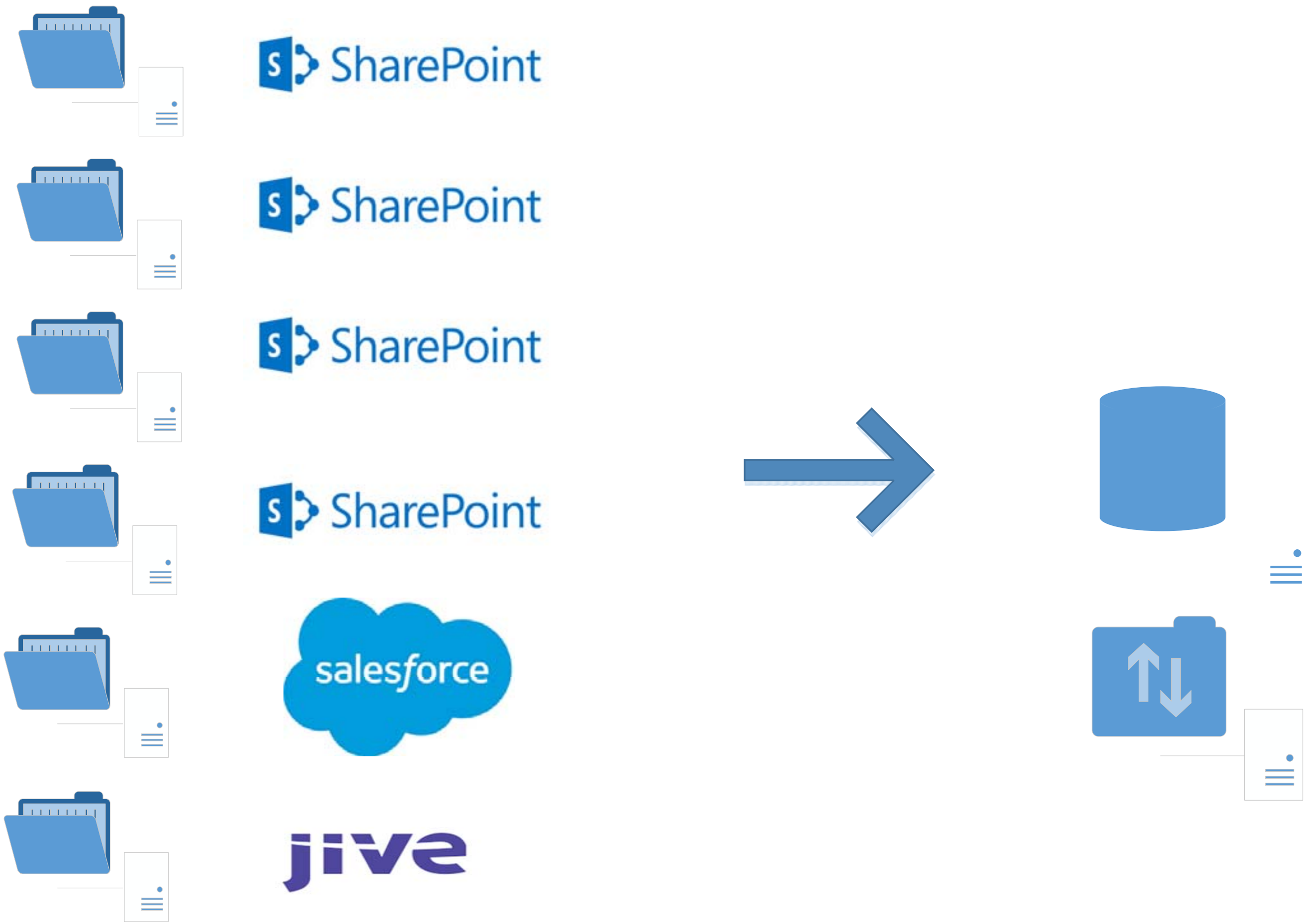
Time is of the Essence

Another factor in divestiture scenarios is that the business transaction timeline is typically short and inflexible. The cost of holding up a transaction can be extremely high. This means that the identification and processing must be high performance.

The Divestiture Scenario

This is based on real customer projects, but some details are abstracted and names are omitted, in order to not disclose our customer's confidential information.

Example Environment: 6 sites, 12 systems



The goal was identification of content that must go with the divested entity and also assurance that no other sensitive content is transferred. Content resided at six different locations, on six file shares (one at each location), four SharePoint 2013 farms, plus cloud-based instances of Salesforce and Jive. There were several other systems involved at each site, but in this project those systems were covered by extracting selected files, putting them either in SharePoint or the File Shares, and following the Identify-Analyze-Move pattern from there.

This project was completed within three weeks in order to fit into the time window of the divestiture transaction. Because of the large volume of content (approximately 45 million files) and the importance of accuracy to protect IP, high throughput was crucial.

In this project, one part of a complex business was being sold, and it was essential that all important IP and knowledge necessary for operations of that business was transferred efficiently. The buyer was a competitor in some other business areas as well. This added an additional level of risk and complexity to the project. Identification of information that was important IP versus the buyer as a competitor, and information that must not be transferred, was business-critical.

Key Capabilities from the BA Insight Portfolio

There are many capabilities of the BA Insight portfolio that applied to this solution, but the most notable are:

- Rules-based auto-tagging (used in all three phases)
- Clustering (used in the Identify and Analyze phases)
- Named Entity Extraction (used in the Identify and Analyze phases)
- File System, SharePoint, Salesforce, and Jive connectors (used in all phases)
- SQL connector with associated crawls (used in the Analyze and Move phases)
- Smart Pipeline to choreograph metadata and insert into SQL (used in all phases)
- Target for SharePoint libraries (used in the Move phase)
- Visual Refiners for exploration and analysis of indexed content as a set (used in the Analyze phase)
- Smart Previews to aid in inspection and analysis of individual documents (used in the Analyze phase)

The SharePoint file system crawler was used to enumerate and crawl content from the file shares, and SQL Server Management Studio was used for analysis and recording content disposition decisions. SharePoint Server was used as the search engine for the exploration and analysis center.

Distributed and Virtualized to Accelerate Performance

Due to the distributed nature of the systems, the organization decided to use a distributed solution. Rather than have a centralized solution work across all sites, all three phases were conducted locally at each site. This had the advantage of much faster crawls because it eliminated network bandwidth as a bottleneck. It also allowed several sites to be processed in parallel. One location was processed first, in order to prove out and practice the overall process. At this site, they tested and tuned classification rules and other Text Analytics that were then used at all the other sites for the Identify phase. Any site-specific rules or tweaking was isolated to the Analyze phase.

Another timesaver was the use of virtualization. One master configuration (with 2 VMs, 4 cores, 8GB of RAM, 500GB of disk on each VM) was set up and tested; then this was cloned and set up at each of the six locations. This included a full set of taxonomies and rules, and also a transfer library using SharePoint 2013.

Approximately 45 Million files needed to be enumerated, processed, and analyzed. These were distributed, with approximately 27TB, 9TB, 22TB, 16 TB, 7 TB, and 11TB respectively at each location. A full crawl took more than a day for each location, nearly 55 hours at the largest site. If crawling had been done remotely from a centralized location, it would likely have taken over 12 days to complete for that site, due to bandwidth constraints.

Incremental and associated crawls for each of the iterations were quick, typically under two hours. Rule updates could be tried immediately within the Testbench built into BA Insight products, so tuning and training was straightforward.

Assembling Taxonomies and Rules

Although this organization had existing corporate and departmental taxonomies, they were deemed too general for the task at hand. Purpose-built taxonomies were quickly crafted, listing the key concepts to identify which content should go with the divested entity and which should not.

- These started from lists of products, employees, processes, customers, projects etc. that already existed. A quick manual first cut was done of the important "must transfer" and "must keep" entities
- Next, a sample set of files was crawled, and a combination of human experts and machine-generated term suggestions was used to build up a working set of taxonomies
- Named Entity Recognition was tuned on this sample set, using a machine-learning training update
- Rules based on these taxonomies and known entities were generated automatically, then a small set of 'business-dictated' rules was added manually

This process took approximately three days. Throughout the subsequent process, small improvements to taxonomies, rules and entity recognition were made continuously. Most of these were done either to incorporate new types of content found during the "Identify" phase at a new site, or to learn from the output of human examination during the "Analyze" phases.

Note that the classification rules did NOT classify directly into disposition decisions. Those decisions were generated based on combinations of elements using SQL within the Metadata Database, and could be overwritten by hand if desired. This allowed a 'best practice' of keeping each rule relatively simple (even though many rules did end up being complex and compound).

Most rules were based primarily on body text and concepts (sets of terms) found in them, but all the existing metadata (file path, file type, file name, title from inside document, etc.) was transferred and was also accessible for use in the rules. Some rules were specific to particular combinations of multiple different metadata files as well as to content in the body of the document.

Many rules also included scoring. During the Identify phase, the score of rules and the confidence level from the unsupervised text analytics components were captured in the file manifest along with all original and derived metadata.

Working with the File Manifest

The 'associated crawl' capability was used extensively, essentially crawling the metadata database using a SQL WHERE clause to select which criteria to use, and fetching the item and associated metadata from the sources system. For example, in the Analyze phase, items whose disposition field is 'inspect' were crawled using an associated crawl, and these items were indexed into a SharePoint 2013 search instance on the local VM set.

Approximately 800,000 items, or under 2%, were indexed for deeper inspection and analysis. This was referred to as the 'leftovers'. These were analyzed in a series of subphases until there were zero leftovers. The file manifest then contained a complete record of each file, how it had been processed, and its intended disposition.

Prior to the Move phase, this manifest was examined by the business, and several alternative transfer sets were modeled using simple SQL statements based on the fields in the file manifest. The final manifest drove the move process, and also served as an inventory of transferred information assets.

Successive Analysis

A SharePoint Search center augmented with BA Insight products was used in exploring and examining the leftover set. The 'demographics' of the set was easy to explore using Visual Refiners, for example looking at the files associated with a selected set of customers and a selected set of products. The Export to Excel feature was used to capture a slice of the leftover set which could then be changed to a different disposition in the metadata database.

Smart Previews were used to quickly examine specific documents and search within them, using the hit map feature to understand the nature of the contents in the document. The Content Assembly feature was used to create excerpts emailed to people within the business group for comment of questions. In this way, the leftover set was successively reduced to zero, meaning the disposition of all content was decided.

Where humans had examined a number of files and decided on their importance and disposition in the transaction, another automated pass at the leftovers was made using text analytics. The human judgement was used to update the training of unsupervised text analytics components and also to update taxonomies and rules in the rules-based classifier component.

Results: High Confidence and Low Risk, On Time, Well Organized

The final result was a set of files to be transferred with the transaction, contained in a single centralized SharePoint farm. This contained 1.4 million files, de-duplicated and fully tagged. All of these had been determined to be important to the IP and operations of the business being divested. The final manifest listing of files to be transferred and their metadata became an asset in and of itself.

Side Benefits: Organized Content and Insights

After the divestiture project was secure, a small continuation project used the File Manifest to update information within the remaining business. This was similar to the content migration scenario, in that duplicate files were removed and a significant fraction of files could be deleted. In addition, the remaining files had complete metadata which became useful for findability, content management, and workflow.

Statistics on the file demographics provided useful insights around the divestiture, and also around other projects and processes in the organization. For example, seeing the products referenced within documents laid out clear gaps in documentation in some products and processes. These gaps were judged to be important regulatory compliance risks, and hence were addressed quickly. Without using this pattern, those gaps would have been found under a regulatory review and the company would have incurred significant fines. By using this solution, those fines were avoided. In addition, the resolution was much simpler and quicker because there was a complete map of information available for this and other products.

Summary

The **Identify-Analyze-Move** pattern uses search and text analytics technology to solve a number of important scenarios. BA Insight's product portfolio, in particular the **AutoClassifier**, can be applied in this pattern to rapidly and accurately process large volumes of content in complex IT environments. We have outlined solutions to two different scenarios that use BA Insight's products in this pattern: Content Migration and Corporate Divestiture.

In the solutions described, we apply text analytics and search technology so that content from a wide range of systems can be identified, inspected and analyzed, structured, and moved. Automatic Classification is a key element in the process because consistent metadata is a backbone for analysis as well as for organizing the content as it is structured. Secure, high throughput **Connectors** and Smart Metadata Mapping are used to extract and structure content during the process. A search-based reviewing center is used by humans to inspect and analyze files where confidence levels reported by the automated process are too low.

For **Content Migration**, less can be more. By getting rid of unnecessary content and organizing content well during a migration project, the overall experience for users is improved and IT's workload is reduced going forward. The trick is to be able to identify and analyse content to recognize what should be kept, what should be archived, and what should be deleted – without losing essential records and information in the process. This requires automated classification, especially in complex environments.

For **Corporate Divestiture**, understanding which content exists and how it relates to the divestiture transaction is an important part of a tricky and valuable scenario. By identifying the IP and other content that must be transferred along with a divestiture or sale, and the IP and other content that must NOT be transferred, risk can be significantly reduced. Because the Identify-Analyze-Move pattern also creates metadata, the transferred content is well organized and easy to integrate into the new corporate entity. We have also seen significant side benefits for the remaining organization – both in having organized information, and in the insight that results from the pattern.

BA INSIGHT

BA Insight provides software that enables organizations to rapidly implement powerful search-driven applications at a fraction of the cost, time, and risk of other alternatives.

Our software portfolio includes connectors and federation to provide secure connectivity to a wide variety of systems; classification that increases findability using auto-tagging, metadata generation, and text analytics; and applications with search squarely at the interface of users and information.

Hundreds of organizations and millions of users benefit from BA Insight's software on a daily basis to provide compelling intranets that people love to use including Australian Government Department of Defence, CA Technologies, Chevron, Deloitte, Ford Motor Company, Keurig Green Mountain, Pfizer, and Travers Smith. BA Insight is a Microsoft Gold Certified Partner, a member of the Microsoft Enterprise Cloud Alliance, and an Elastic Partner.

Visit www.BAinsight.com for more information.