



Find a Complete **Enterprise Profiling** **Solution Checklist**

The right solution will support multiple use cases, maximizing your investment and even mitigating risks and reclaiming costs

Finding a Complete Enterprise Profiling Solution

Data Profiling indexes all forms of unstructured files and document types, creating a searchable index of what exists, where it is located, who owns it, when it was last accessed and, optionally, what key terms are in it.

This ability has a wide range of uses in the data center. Ranging from storage management to litigation readiness, data profiling is the basis for finding, reporting and managing the disposition of the user-generated content. Without profiling, data remains dark and unknown, becoming impossible to manage and a wasted expense.

Index Engines' data profiling solution is focused on supporting the following use cases:


- **Storage Management:** tiering, consolidation, migration, remediation and chargebacks
- **Security and Risk Mitigation:** industry compliance, self-audits and breach prevention
- **Litigation Readiness and eDiscovery:** early data assessment, archiving and policy enforcement


With such a wide range of use cases it is critical to select the right solution otherwise you will be forced to implement several tools to solve these challenges. Many products exist that satisfy storage management requirements, such as traditional Storage Resource Management (SRM) tools or niche eDiscovery utilities, however as the data center evolves and the requirements change, these basic tools will quickly outlive their usefulness.


Selecting the right data profiling solution will allow many current and future projects to be addressed through a single application. Whether it is migration of aged data to the cloud or finding responsive email based on keyword searches and date ranges to support litigation, it is critical to select a complete solution - otherwise you will be faced with a constant struggle to manage user data.


The following selection criteria will help you understand and differentiate the data profiling solutions on the market today. Make sure that when you are selecting these solutions you understand what you are getting, and not getting, by making a choice. Your choice may be based on solving an immediate challenge, however, thinking forward you should look at solving challenges that have not yet arisen to maximize your investment.


Enterprise Data Profiling Checklist


 **Data Sources** Index Engines supports a large variety of unstructured and semi-structured file types including loose files (documents, spreadsheets, presentations, etc.), email (Exchange and Notes databases, PSTs, etc.), and archives (SharePoint, Enterprise Vault, etc.). Selecting a solution that can index and search files only, but not email, is limiting. Index Engines provides depth of indexing to support the most common data sources including files, email, SharePoint, disk and even legacy backup


 **Backup Data** You may not need to profile and manage legacy backup data today; however, organizations have significant volumes of this content in offsite storage which over time has become a risk and a liability. Index Engines delivers a unique and cost effective ability to index, report and extract data from legacy tapes without the need for the original backup software in order to execute intelligent disposition strategies. This includes restoring files and email from tape to support eDiscovery as well as legacy tape remediation to support proactive litigation readiness. Utilizing the same user interface, data on tape can be managed as easily as online network content.


 **Deep Metadata Indexing** Data profiling platforms must make a decision on how deep to index data. Most indexing platforms utilize very light metadata and capture content from files that includes dates, locations, owners, etc. However, this metadata indexing treats containers, such as email databases and zip/compressed files, as a single entity and only delivers metadata about the object as a whole. The problem with these tools is that these containers are full of important content which can only be discovered by full-content indexing platforms that go inside the repositories. Index Engines delivers unique deep metadata indexing and deduplication of content within containers and objects such as email databases, PSTs and compressed files (such as TAR and Zip). Using deep metadata indexing, reports and analysis of individual email, mailboxes and compressed files can be executed and the specific content in these containers extracted. Additionally, deep metadata indexing creates a document signature so files and email within repositories can be deduplicated. Other metadata indexing solutions cannot provide this level of knowledge and access to content.


 **Full-Content Indexing** When selecting a profiling platform it may be to support a data migration or consolidation requirement. However, the organization may also be struggling with eDiscovery and data security efforts. The same profiling platform that can support migrations, with metadata indexing, should also go deep into the data, or full content indexing, to support legal and security efforts. Index Engines has been processing eDiscovery and security initiatives for over a decade. With a range of simple metadata scans to deep full text indexing, Index Engines can find, report and manage the data based on any criteria.


 **Deduplication** Enterprise data is highly redundant. Copies of files and email exist throughout networks and desktops, and more significantly, in backup data. It then becomes challenging to differentiate between unique documents and redundant copies of these documents. Most profiling platforms that perform metadata indexing are not able to differentiate unique files and email from duplicates. Reporting on duplicates is a very critical component for data center management or even an eDiscovery or security event. Index Engines performs deduplication on files and email. This


 **Sequential Indexing** Most indexing technology relies on random disk access in order to crawl through and index data. This is a traditional approach and has a number of bottlenecks that will slow down processing and require data to be extracted to disk in order to process it. Index Engines has invented a unique approach towards indexing, utilizing a sequential approach that indexes data as it streams, with no requirement for it to be stored on disk. Two prime examples of this advantage include Index Engines' ability to process data on backup tapes and in SharePoint. For backup tapes, Index Engines can index and process tape data without restoring it to disk as is required by all other vendors. Similar for SharePoint where other vendors dump all SharePoint data to disk for indexing, where Index Engines efficiently indexes the content in place.


 **High-Speed Indexing** Profiling solutions are road blocked by slow networks and constrained bandwidth. These constraints will slow overall indexing speeds and require significant effort when processing hundreds of terabytes or petabytes of data. Index Engines delivers the fastest indexing performance on the market today. With speeds reaching up to 1TB per hour no other solution can match Index Engines' performance. Even data centers with slow networks can take advantage of Index Engines power by leveraging the unique multi-stream capability which utilizes multiple parallel indexing jobs that compensate for slow environments.


 **Federation** Organizations have accumulated significant volumes of data within a single data center or across geographical distributed data centers, which must be processed and analyzed. Profiling these volumes requires performance and horsepower that scales up and makes short work of a big job. The Index Engines platform is extensible through the addition of nodes that linearly scale performance adding power when needed. Combined with federated search and reporting this power will support even the largest enterprise environment.


 **Index Footprint** A commonly overlooked specification of indexing technology is the size of the index that is generated, or the index footprint. For example, if you create a large index footprint, something in the range of 40-60% of the original data size, when indexing 100TBs of user data you will need 40-60TB to store the index. That size footprint can be typical for profiling products, some even skyrocket to beyond 100% when they need to cache the data in order to index it. Large index footprints result in expensive storage requirements as well as increased resources in order to profile and manage large enterprise environments. Who is going to implement 50TB of data storage in order to clean up and manage a 100TB server? Index Engines' footprint is on average 1% for metadata only and 5% for full text. This is an industry leading statistic that allows for extreme scalability into the petabytes, as well as cost effective deployments that minimize the storage requirements for the index.


 **Extensive Search Options** Limited search capabilities restrict the reports and analysis that can be performed when profiling data. Most search engines can perform a simple metadata search, but can't use Boolean expressions on these properties. Additionally, a keyword or pattern search, including Social Security or credit card numbers, allows for comprehensive audits for PII and other sensitive content. Index Engines provides an extensive search interface that allows for the combination of a number of powerful search options in order to find and report on data according to the most sophisticated criteria.


 **Dynamic Reporting** Many indexing tools have preset canned reports that provide a structured view into the data. These reports can help analyze storage capacity, or find specific types of documents; however, flexible reporting that allows for a customized look at the data is key when supporting a range of use cases from storage management to information governance. Index Engines has powerful and dynamic reporting, filters and queries that can be customized to meet any use case.


 **AD Integration** Active Directory (AD) provides a wealth of knowledge regarding users and departmental groups. Without AD integration the reports generated are for individual users which does not provide a view into what each department (ie. HR, Research, Manufacturing, etc.) is doing. Departmental reports allow data usage and audits to be rolled up by group and support the implementation of charge backs to recoup data center costs. Additionally, AD maintains a group of inactive users, typically ex-employees, where a summary report can be executed on abandoned data owned by ex-employees. Index Engines leverages the metadata available in AD allowing for more insightful and comprehensive reports and analysis of data by department and group.


 **ACLs Support** Beyond file and email-level metadata, network data contains valuable information that can help manage security and compliance initiatives. ACLs (access control lists) contains metadata about file permissions and who has access to them. This metadata is critical in supporting security and legal efforts by providing a granular view into permission and security of sensitive content. Index Engines supports indexing of ACL reporting on users read, write and browse permissions of data.


 **Defensible Disposition** Once data is profiled and a disposition strategy defined it is important that the same solution manages the data throughout its lifecycle. IT organizations have many tools that can migrate and consolidate data; however these tools will strip out valuable metadata, such as owner and last accessed date. Index Engines has built in disposition capabilities allowing for one click migrations or deletions, and automated policies to be enforced on existing data. Additionally, all dispositions are logged in order to maintain a historical and defensible history of the actions for future reference and audits.


 **Integrated Archive** As data is profiled and reports are generated that uncover sensitive content that must be preserved for legal and compliance purposes most organizations migrate these files and email to their corporate archive. Index Engines delivers an integrated archive that ensures data is forensically sound and there is no spoliation of the metadata. The process of enforcing data policies and preserving data in an archive can now be automated and simplified within a single platform.

 **Backup Catalog Management** In addition to backup data sources, you may have one or more legacy backup catalogs that have been left unmanaged. Beyond simply searching and managing data on legacy backup tapes, Index Engines can ingest backup catalogs in order to facilitate access to the tape content. These catalogs contain valuable metadata that can be used to determine the disposition of the content. For data centers that have been through a number of mergers and acquisitions, catalogs containing old backups of aged content can be ingested, managed and data that is no longer of value remediated along with the catalog to support information governance initiatives.

 **Patented Technology** Most profiling solutions consist of off-the-shelf components that are combined to provide an overall platform. These off-the-shelf components are dated and they do not take advantage of state of the art technology. Index Engines patented platform was built from the ground up to support complex enterprise class data centers. Index Engines technology far outperforms any other solution on the market today due to innovation and commitment to support even the most challenging environment.

 **Automation** Obtaining a view into enterprise data is a challenge due to the scale and complexity of the content. Using automated reports and indexing that refreshes the dynamic content on a scheduled basis is key to gaining the knowledge and insight needed to manage data properly. Indexing jobs across the enterprise can be automated using a customized scheduler w/ email notifications to update IT personnel on the status of the job progress. Index Engines action queries and reports store common requests and defines a schedule that automates these activities and also logs the reports for historical views into the data.

 **Fault Tolerance** Indexing engines are continually processing data in order to extract the important metadata and content. These indexing jobs become a critical component to the data center and organizations will depend on them for accurate and reliable information. It is critical to ensure that fault tolerance is built in allowing for jobs to recover from any malfunctions. Index Engines has built in fault tolerance that automatically restarts an indexing job and seamlessly corrects indexing issues that may occur ensuring content is processed reliably.

 **Flexible Deployment** And finally, when you're ready deploy a solution having choices that fit your needs are critical. Many solutions force you into a pre-defined configuration or capacity that may not be suited to your data center environment or needs. Index Engines delivers a number of deployment options from software only, to virtual server to an appliance based platform. Additionally, Index Engines allows for a small footprint that can grow over time - whether you are starting with 5TB or 100TB and growing to petabytes.

Why Index Engines?

Index Engines has spent over a decade perfecting the most in-depth, enterprise-ready solution for finding, managing and making decisions on unstructured user data.

Index Engines provides unprecedented file-level knowledge to manage the growing costs and risks associated with unstructured user data.

Delivering the fastest indexing platform on the market today, at speeds up to 1 terabyte per hour per node, detailed metadata is captured across all data environments from primary storage to legacy backup tapes and disk.

Using this knowledge, information can be classified and dynamic reports created to analyze and identify content based on legal and data policies. Contact us today to learn more.

Learn more about the Index Engines by visiting us at www.indexengines.com or call us at 732-817-1060.

Search Report Manage Archive Govern

