Index Engines’ Playbook: Supporting GDPR Compliance through Data Classification

How technology supports privacy and management of unstructured data
The alarms are ringing and the EU General Data Protection Regulation (GDPR) deadline is fast approaching. Organizations are faced with developing a plan to manage hundreds of terabytes or even petabytes of unstructured data. May 25, 2018 is a dark day on their calendars.

In addition, corporate data centers don’t understand what they have or where data lives as their storage platforms are not equipped with the ability to search and report on the data that needs to be governed. This prohibits business and compliance teams from knowing what exists, why it exists, and if it has any business value. The GDPR requires this level of knowledge and the data center must deliver it.

*The GDPR*

Boiled down, the GDPR assigns ownership of personal data inside your environment back over to the individual. Corporations are simply custodians of this data and must treat it according to the requirements defined in the GDPR. If they do not, the fines and restrictions can be severe.

The GDPR consists of 99 articles that mandate how data is to be handled. The following 9 articles focus on areas where the right technology platform can make easy work of the regulation:

- **Article 15:** Right of access by the data subject
- **Article 16:** Right to rectification/correction of data
- **Article 17:** Right to erasure (right to be forgotten)
- **Article 18:** Right to restriction of processing
- **Article 20:** Right to data portability
- **Article 21:** Right to object
- **Article 22:** Automated individual decision-making, including profiling
- **Article 25:** Data protection by design and default
- **Article 35:** Data protection impact assessments

*Where to Begin?*

Most organizations have amassed hundreds of terabytes or even petabytes of unstructured user content in their data center. Getting started can seem daunting. Where to begin? How to understand what exists? How do you ensure you provide a solution that can find and manage all personal data under management with considerable precision?

Breaking down the challenge into achievable steps, and leveraging enterprise class indexing technology will make easy work of the challenge at hand. This playbook from Index Engines will provide an overview of an actionable workflow that can support your efforts when managing personal data.
**Workflow Overview**

When dealing with large volumes of user content it is important to break down the environment into more manageable data sets. If, for example, you are faced with an environment that consists of petabytes of information, using classification you can break this environment down into a smaller data set of in-scope data that can be more easily managed. In scope data is user content that is highly suspect of containing personal data.

Once data is classified, policies can be developed that focus on the in-scope data set so that it can be managed, searched and secured in support of the regulation. This avoids the challenge of implementing these policies across the complete enterprise data environment, allowing you to only focus on what is relevant to the GDPR regulation.

Finally, ongoing management of the content is critical to maintaining compliance with the regulation. Getting data into the right place, purging what no longer has business value, securing sensitive content so it is not easily available to rogue employees or other bad actors.

These are some of the key disposition capabilities that must be implemented and automated to ensure ongoing and auditable support for this challenging regulation.
Data Classification

Classification of content from massive volumes to meaningful data sets will make easy work of finding and managing personal data. The use of traditional data mapping is one popular technique that helps to classify data, however it is not typically based on detailed metadata input. This approach will compliment data mapping with comprehensive metadata reports and analysis to enhance the classification process.

Leveraging metadata the following classification buckets can be defined allowing data to be tagged and categorized:

- **ROT Data** – Redundant, obsolete and trivial data is classified as content that no longer has any business value, but is getting in the way of managing the content with value. Using metadata queries to find content such as:
  - Duplicate content that has not been accessed in more than three years and is owned by ex-employees.
  - Aged files which are file types that do not typically contain sensitive content and are considered obsolete.
  - Server log files, or other files types such as personal photos and music that can be considered trivial.

- **Out of Scope Data** – Finding content that has business value but based on metadata properties such as file type, owner, location, etc. is something that would not contain personal data can be classified as out of scope data. This is content that can essentially be ignored with respect to this regulation. It can be managed and monitored, however when a search or request for personal data occurs, this content can be ignored.

- **In-Scope Data** – This is the core data set that will be the focus of your GDPR efforts. This is content, based on your knowledge of the business, that will typically contain personal data. This will typically be a much smaller subset of the overall content and will make management and disposition a much easier task.

---

CLASSIFICATION FEATURES

- Comprehensive support for file servers, email servers, SharePoint servers, etc.
- Industry leading indexing speeds reaching up 8,000 files/sec/node
- Efficient index storage - ~1% of original data capacity
- High speed queries and search
- Tagging and management features
- Linearly scalable for large data environments

---

ROT

- Redundant, obsolete and trivial data with no business value
- Eliminate from data center to simplify management process

Out of Scope

- Data with value that does not contain personal content
- Eliminate from GDPR queries

In Scope

- More manageable sub-set of all data assets
- The focal point of GDPR queries

---

www.indexengines.com  Info@indexengines.com  732-817-1060

Copyright 2017. All rights reserved. Index Engines Inc. All products mentioned are trademarked by their respective organizations.
Search

The definition of personal and sensitive data content is vast and vague in the GDPR regulation. It will be refined over time, however, as it does evolve you will need a flexible and comprehensive search solution that can keep up with the changes.

How can you leverage search tools to identify personal data? How can you find it when you don’t know what you are looking for? How do you ensure you will find all personal data under management with precision?

Search is core to support this regulation. It must not only be flexible, but comprehensive to ensure a robust process. Some of the key search capabilities must include the following:

- **Keyword**: Names, addresses, TaxID’s, etc.
- **Pattern**: Bank routing, social security, credit card numbers, etc.
- **Regular Expression (RegEx)**: Phone numbers, postal code, license plate, UK bank sort code, email address, etc.
- **Conceptual**: train the query engine to identify documents containing personal data
- **Boolean**: Use operators such as AND or NEAR combined with the above to fine tune and focus searches.

Simply using search tools with keyword or pattern search will miss large volumes of personal data. Leveraging more advanced techniques, including concept search, will find all personal data and deliver more comprehensive and auditable results.

Concept search utilizes predictive coding algorithms to find content even if you don’t know what you are looking for. Simply train the search engine to find similar documents to one that does contain personal data, and you will quickly find comparable results.

<table>
<thead>
<tr>
<th>Type of Search</th>
<th>Example</th>
<th>IE Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keyword</td>
<td>Files containing name John Q. Public</td>
<td>&quot;john q public&quot; or (john near10 public) or &quot;john public&quot;</td>
</tr>
<tr>
<td>Pattern</td>
<td>Files containing social security numbers</td>
<td>pattern:ssn AND (social OR SSN OR SS)</td>
</tr>
<tr>
<td>Regular Expression</td>
<td>Files containing Finnish identity numbers</td>
<td>[0-9][6]{5}[-:</td>
</tr>
<tr>
<td>(RegEx)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concept</td>
<td>Find similar documents based on conceptual search algorithm</td>
<td>Trained query</td>
</tr>
</tbody>
</table>

SEARCH FEATURES

- Intuitive web-based search and reporting interface.
- Flexible queries including Boolean, pattern, proximity, regular expression, concept and more.
- Search on simple or deep metadata, user mailboxes, or full text content.
- Document reconstruction to review content directly in the search window.
- Automated queries can run on a scheduled basis.
**Manage**

Management of data according to the GDPR can mean many things. Securing sensitive content. Migrating personal files to a secure archive. Purging and defensibly deleting abandoned and obsolete (ROT) content. Migrating aged data to the cloud. Monitoring data in place to be aware of any changes. Auditing content for security issues that could expose sensitive data to bad actors. The GDPR requires many of these activities and more.

Reviewing the classification funnel from earlier, we can determine some logical disposition and management strategies.

- **ROT Disposition**: Data with no business value is clogging corporate networks and servers. This data can be easily uncovered using metadata properties and migrated to cloud to reduce the data center footprint and costs, or purged if your policies allow.
- **In Scope Disposition**: Integrated disposition options allow for data to be found and retrieved, migrated to a new repository including the cloud, or preserved in a built in archive on disk or in the cloud for long-term retention. For right to be forgotten requests, in-scope data can be searched, responsive data discovered, and then deleted with full audit logs that validate the process.

All disposition options are managed in a defensible process, with all migrations logged for a full audit trail. The file metadata also remains intact as data in migrated so no spoliation occurs.
The ROI of ROT Clean Up

What does ROI have to do with the GDPR? When classifying data many organizations will find that an average of 40%, or beyond for some, will be ROT. Once this data is classified, you can easily migrate it out of the data center to lower-cost cloud storage or delete it. The return on investment for this process is significant and will help pay for implementation of technology in support of the GDPR. For organizations that have not budgeted technology in support of the GDPR, this reclaiming of existing data center expenses to offset the costs of supporting the regulation is something that will appeal to every CFO.

Benefits of ROT clean up:
- Reduce storage footprint and ongoing capacity upgrades
- Reduce data center infrastructure, including data protection costs
- Eliminate risk and exposure of unknown legacy content

If for example, your organization realizes a 40% ROI through ROT clean up you can save significant budget expense. Many of the leading analyst firms cite a range of fully loaded storage costs. For example, Forrester Research has calculated 100TB of storage at $955,500, while IBM calculates $4M for 1PB. If you were to classify 400TB as ROT data and migrate it to the cloud it could cost $150,000 annually to storage and manage, versus the millions that the analysts have determined by maintaining it in the data center.

Additional Features in Support of the GDPR

Data Security

In protecting sensitive and personal data in place you must also monitor and report on file permissions and access. This will enable you to determine if there are rogue employees stealing or tampering with content they should not have access to. With this level of reporting, data can be proactively protected and secured to prevent data breaches or theft.

Beyond finding and managing content containing personal and sensitive information, you can generate reports on Activity Logs to determine who has accessed these files, files they should not have permission to access. Additionally, reporting on Access Control Lists (ACLs) you can determine who has read/write/browse permission for sensitive files, to ensure that only appropriate users can access this sensitive content.
Some example of reports that can be generated include: ‘what files owned by the finance department does everyone have access to’ or ‘who has recently accessed specific sensitive content.’

With this level of security reporting you can audit the data environment and proactively uncover flaws that can result in non-compliance with the GDPR.

**Automation and Monitoring**

One of the requirements of the GDPR is the staffing of a Data Protection Officer (DPO) within the organization. The DPO is required to educate, monitor, audit and assess according to the regulation. They must answer to the regulatory body when inquiries are initiated.

Utilizing automation of indexing and stored policies, and a defensible disposition strategy, the DPO will be able to easily audit and monitor compliance. The workflow documented above will deliver a consistent solution that will deliver comprehensive protection of personal and sensitive content and the tools needed by the DPO to ensure they can support an audit from the regulatory body.

**Backup Tapes**

And what about your legacy backup data? Data that was initially preserved for disaster recovery but now is an archive of all your long-term retention data?

Personal and sensitive data exists on backup tapes. In order to comply with the regulation, this data must be managed as well. Simply managing your network data would leave a legacy instance in your backup data that would cause risk and liability in the future.

The easiest approach would be to profile the backup data once it is no longer used for disaster recovery, determine what is required for long-term retention, and migrate this content out of backup and into a policy based archive. Once this is accomplished, the backup data can be purged.

By eliminating the practice of saving old backup data, specifically backup tapes, companies eliminate the need to manage - and go to - these repositories in the future. By leveraging the GDPR to clean up and remediate legacy tapes, offsite tape costs and ad-hoc eDiscovery costs can be recouped. Index Engines’ unified information management platform delivers unique technology to support the migration of legacy tape data from tape to disk/cloud. Finding the data that has value, and selectively migrating this content makes for cost effective management of these records. The policies that are defined to support GDPR on the network data content can easily be applied to legacy tape data in order to streamline the migration.
Summary

The GDPR is a significant information management challenge. Organizations must embrace technology to help meet and exceed the requirements of the regulation. It is critical to deploy a solution that can meet the following requirements:

- Petabyte class indexing technology
- Support for all classes of data, including legacy data on backup tapes
- Comprehensive search and reporting
- Integrated management and disposition capabilities
- Secure archiving to support secure preservation and access
- Automated processing and monitoring of data policies

Only Index Engines can meet the demands of the GDPR – and deliver a cost effective solution!

Contact Index Engines today to get started on your GDPR Playbook.