How BreakwaterAI Accelerates the Identification of Risk and Value Within Unstructured Data

Will Jaibaji
SVP, Product Strategy
Breakwater

Dinosaurs Index Everything

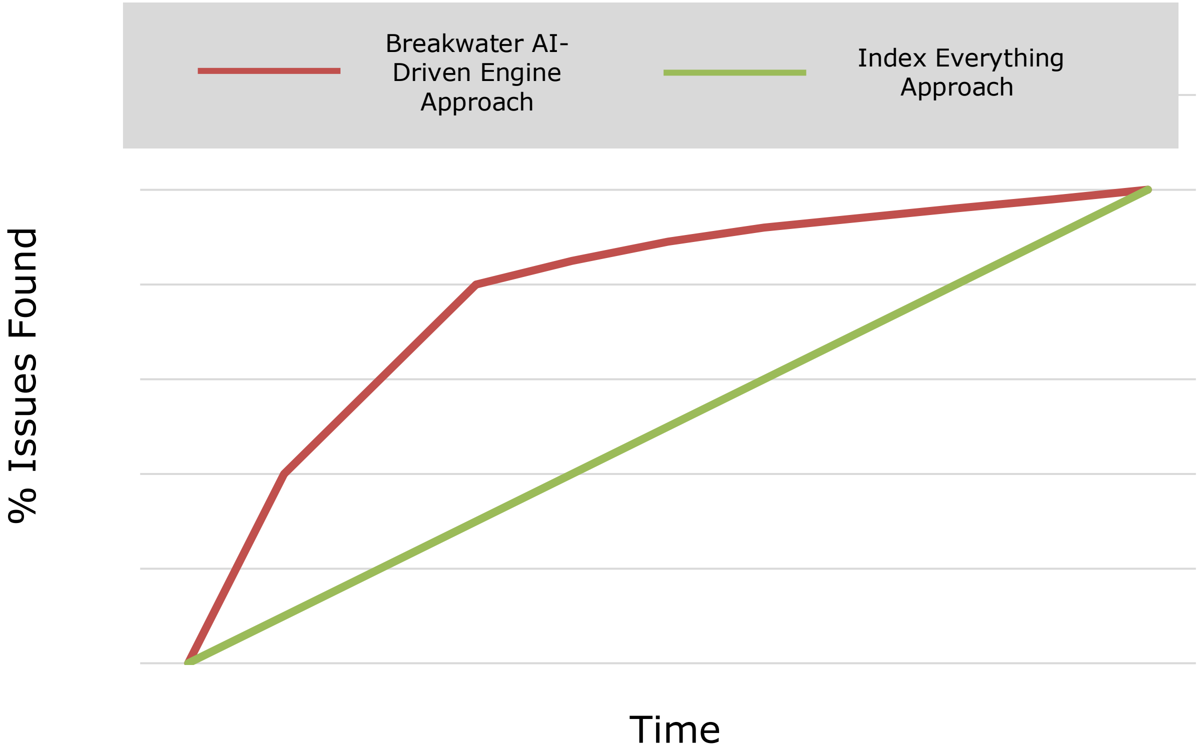
Privacy regulations, risk of data breaches, and good business practices dictate that employee and customer personal information must be identified and removed or otherwise protected. Unfortunately, most enterprises have petabytes of data, and scanning or indexing all this data can take years, even with the most modern technologies.

Similar issues exist in identifying high-value business data, regulatory records, and countless other use cases. Businesses typically must choose between multi-year projects with high infrastructure costs, putting high-overhead employee processes in place, or ignoring the problem.

Using AI and Modern Data Analysis to Build a Better Mousetrap

Accelerating scanning and indexing rates is not a viable approach as most systems are already bound by data source response times. Instead, a modern approach focuses on identifying the data locations most likely to contain high-risk or high-value data. This is accomplished by applying an intelligent methodology and modern technology that identifies those highest value and risk segments to rapidly generate tangible results.

This approach utilizes statistical sampling to determine the highest priority data scope and then introducing Machine Learning models based on file metadata to prioritize full document scanning. Various models can be applied based on cross-customer learning or specifically created based on a given data set.



How BreakwaterAI Accelerates Data Categorization

For example, connecting to Microsoft OneDrive with the appropriate credentials would immediately provide access to the top-level folders and its metadata index for the files stored there. When looking for Sensitive Financial data, Breakwater Privacy utilizes pre-built or custom categories based on combinations of indicators showing up within a document. Breakwater Privacy will then run a very quick statistical sample on each of the top-level folders, determining in a few hours which folders have the highest percentage of Sensitive Financial data. This sample pulls the full content of each document within the sample set and evaluates it against the criteria for the categories selected.

Data Indicators

In Breakwater Privacy, an Indicator is any data, text, metadata, or other information of interest in a document. These may represent concepts or may be discrete data elements. For example, indicators may be related to Personally Identifiable Information (PII), such as social security numbers, names, addresses, account numbers, or other related items. A user may define these indicators using methods including searches, regular expressions, natural language processing, entity extraction, data matching, data mining, machine learning, or any other method for specifying an indicator.

Indicators may be related to any type of information of interest, for example, client data, employee data, intellectual property, health information, unannounced products, risk information, or any other information that an individual, corporation, or entity may want to identify. Indicators could be any type of information, including text, numbers, metadata, drawings, diagrams, formatting, pictures, audio, video, multimedia, binary data, or any other type of information that exists in a document.

Once the top-level folders are prioritized, a user can run BreakwaterAI, the artificial intelligence engine in Breakwater Privacy, to assign a probability score for indicators in each file. The key reason for this step is that file metadata can be retrieved exponentially faster than the full content of the file; thus taking a very tiny amount of time to prioritize files much faster than utilizing traditional linear scanning.

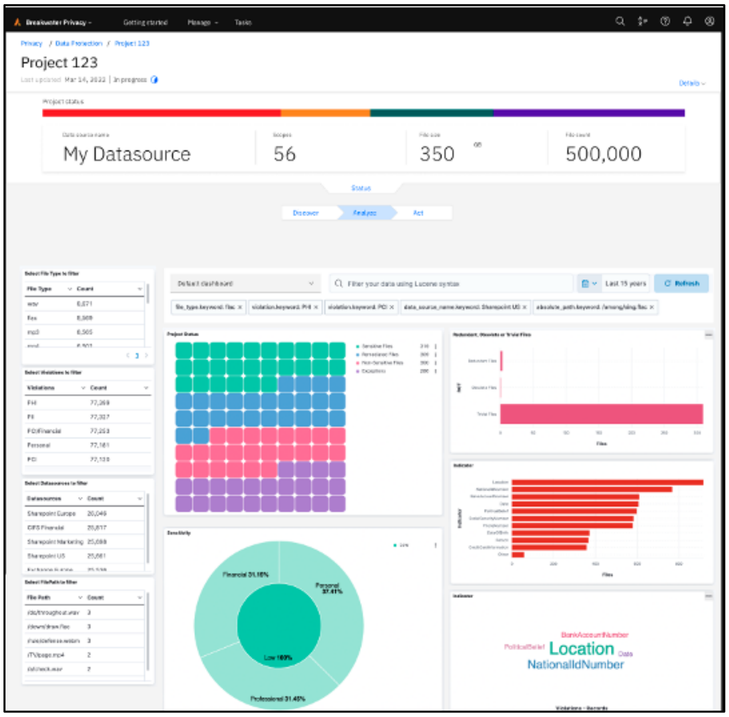
Using Metadata

Once the top-level folders are prioritized, a user can run BreakwaterAI, the artificial intelligence engine in Breakwater Privacy, to assign a probability score for indicators in each file. The key reason for this step is that file metadata can be retrieved an order of magnitude faster than the full content of the file; thus taking a relatively short amount of time to prioritize files.

BreakwaterAI will pull the metadata for all the files in that set of folders, then give each file a probability score for containing the targeted data categories. These models are built at the indicator level. For example, there is a model that tells the system the likelihood of a document containing a bank account number and another model telling the system the likelihood a file contains a credit card number. BreakwaterAI will assign each file a probability score based on the combination of indicators that the Sensitive Financial data category contains.

Metadata

The metadata is any information related to the file, and metadata may vary from different data sources. Metadata may include information such as file name, file path, file type, size, owner, last modified by, create date, last modified date, modification history, access permissions, groups, tags, categories, logical locations, physical locations, geographic locations, identifiers, or any other information that a system may generate. The generation of metadata is generally not resource intensive when compared to document assessment.



Select File Type to filter

File Type	Count
wav	6,671
flac	6,569
mp3	6,565

Select Violations to filter

Violations	Count
PHI	77,269
PII	77,327
PCI/Financial	77,253
Personal	77,161
PCI	77,130

Select Datasources to filter

Datasources	Count
Sharepoint Europe	26,046
CIFS Financial	25,817
Sharepoint Marketing	25,698
Sharepoint US	25,661

Select FilePath to filter

File Path	Count
/do/throughout.wav	3
/down/draw.flac	3
/rule/defense.webm	3
/TV/page.mp4	2
/of/check.wav	2

Breakwater Privacy

Breakwater Privacy has pre-trained models for each of the defined indicators. These models will continue to learn based on the data that has been evaluated. This capability is especially helpful when fields like the file Owner or Location are concerned, as the model learns which owners or locations are most likely to have Sensitive Financial data and tunes the model accordingly. BreakwaterAI uses an intelligent selection process for training by choosing files that will best improve the models. This process could mean choosing a document where BreakwaterAI initially gave the document a low probability of having a certain indicator, but it ended up being present or vice versa.

BreakwaterAI is also continuously learning across the entire customer landscape. Anonymized data is rolled up, and the BreakwaterAI models are improved. In this way, BreakwaterAI can segment data to prioritize the search for interesting data, regardless if the data is identified as highly risky or highly valuable data.

In practice, BreakwaterAI will assign probability scores to each file in the following manner:

- Select AI models based on categories selected by the user. These categories are combinations of indicators.
- These indicators are used to select the algorithms that will calculate the probability of the existence of those indicators.
- Each algorithm uses the extracted metadata to determine the probability of the existence of one or more indicators in the system.
- The metadata may need to be prepared or modified before input to the algorithm.
- One or more algorithms may run on the data, with no limit to the number of algorithms that may be applied. The results of each algorithm are combined into an overall probability for each indicator for each document. The system may be run multiple times for each set of metadata, with either the same or different algorithms each time.
- The system uses the results from the combined algorithms to assign documents or sets of documents to segments. These segments are generated by distributing the documents based on their probability scores. Segments may exist based on each indicator or combined results of multiple or all indicators. Each document will exist in at least one segment.

Conclusion

As mentioned earlier, this effort aims to find interesting data faster. In an organization with hundreds of terabytes or multiple petabytes of data, this can mean finding a sensitive file in days or weeks instead of waiting months or years with that risk hanging out there.

About Breakwater Privacy

Breakwater Privacy™ enables customers to identify and remediate data privacy concerns in a fraction of the time by connecting to multiple unstructured data sources and leveraging artificial intelligence (AI) to sample and segment data targeting hot spots. Problematic data can be grouped, and actions can be applied.

About Breakwater Governance Cloud

Breakwater Privacy runs on the Breakwater Governance Cloud™, which can be implemented via all the modern technology deployment models – SaaS, single-tenant public or private cloud, or on-premises. The platform scales to large enterprises through the deployment of decentralized processing nodes. Additionally, organizations can deploy instances within a region in compliance with data localization regulations.

Breakwater Solutions

Breakwater helps mitigate risk and gain insight from sprawling information by combining technology automation and human expertise. Our expert consulting, software, and managed services address the challenges within information governance, disputes and investigations, regulatory compliance, privacy, and cybersecurity. Our solutions allow governance, legal, and risk professionals to locate, access, analyze, and manage information by making data transparent and actionable. Breakwater helps clients in public and private sectors mitigate risk, improve productivity, and increase profitability by transforming how they use data. Learn more at www.breakwatersolutions.com.