Use sampling as part of a reasonable approach to unstructured data compliance

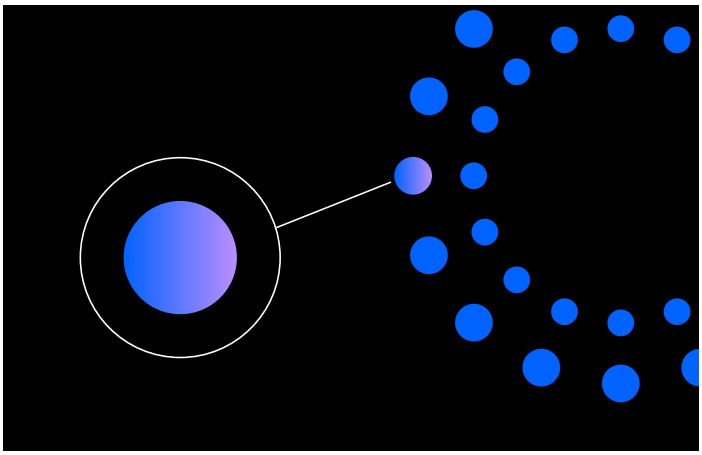
IBM

Outline

- Background
- Overview
- 3 steps to gain confidence in your unstructured data compliance program.
  - Step 1
    Perform random sampling.
  - Step 2
    For high-risk areas, identify and remediate issues
  - Step 3
    Check your compliance
- Conclusion

Background

In the age of data protection regulations and a renewed focus on compliance and privacy, organizations must assess the various data sources they have in their enterprise. It's important for them to establish whether sensitive information is present in areas where it's not supposed to be. When sensitive data is improperly stored, regulators can fine or sanction organizations. Beyond legal consequences, a data breach could make such data publicly available, and expose it to the world. So organizations are left with a key challenge: reviewing billions of documents in petabytes of data, within a reasonable period of time, and for a reasonable amount of money.



Overview

Organizations typically review vast amounts of data by indexing every document and searching for personal or sensitive data with pattern matching or machine learning algorithms. While this method can be fairly accurate, it's also expensive and time-consuming. Assessing all the data can take months or, more likely, years. Rather than index every file, it's more efficient for organizations to make use of statistical sampling technology and review a random assortment of files. The initial goal of sampling is to identify the areas of highest risk within your enterprise. To do this, you can use a type of random sampling. That is, you assume that there are issues and determine where and how extensive they are. Once you've identified where those issues exist, you can perform a comprehensive index, reviewing every file in the potential hot spot and remediate the issues. Finally, once you have a hypothesis that a particular area is clean, you can perform another sampling pass. In this case, you can use negative sampling to confirm—within a given confidence level—that the vast majority of issues have been remediated. By taking this approach you not only prioritize cleanup efforts but find many areas that don't have policy violations at a certain confidence level and eliminate the need for comprehensive indexing.

3 steps to gain confidence in your unstructured data compliance program

The following three steps highlight how you can use sampling as part of a reasonable approach to unstructured data compliance.

Step 1
Perform random sampling

As you begin this step, you assume you have personal data exposure across your organization, but you don't know how prevalent those issues are or where they're located. The goal of this step is two-fold. First, to highlight areas of the highest risk to prioritize cleanup efforts. Second, to highlight areas that are already clean so they can be excluded from further analysis. Use random sampling to determine how prevalent issues are within the given area of analysis. IBM Watson® Knowledge Catalog InstaScan bakes this random sampling into its Risk Assessment capability. With this technique you only need to review a small percentage of the total data in order to be reasonably confident in the analysis.

Watson Knowledge Catalog InstaScan uses the following formula to review sample data:

Sample size =

$$[z^2 * p * (1 - p) / e^2] / [1 + (z^2 * p * (1 - p) / (e^2 * N))]$$

Here, **N** is the population size, **e** is the margin of error, and z is the score or number of standard deviations above or below the mean data point. Finally, **p** is the expected probability that a document is not clean—50% is used for the worst-case approximation as the expected value isn't known. Using a z score lookup table, we can quickly translate the z score to a confidence interval. For example, suppose you have 1,000,000 documents in your first data set (**N**) and you'd like to be 95% confident in your analysis (Z=1.96), with a margin of error of +/ – 1% (e=.01) with an expected probability that a given document has personal data of 50% (p=.5).

Plugging these values into the formula above generates a sample size of 9,513 documents—just 1% of the total data. If you analyze those 9,513 documents and find that 10% of them have personal data, you can make an assessment. You can expect that if you conduct a similar random analysis 100 times on that same data set, 95 times you'll find that between 9 and 11% of the documents have personal data.

This technique provides big advantages for larger data sets. To illustrate this, extend the example with some quick math.

As above, the sample size required for a population of 1 million documents is 9,513, whereas the sample size required for a population of 10 million is 9,595 documents–just 0.1% of the total data. So, when the data size grows tenfold, the sample size growth is minuscule in comparison. The converse of this is also true—the smaller the data set, the higher percentage of data that needs to be sampled.

Now you know how prevalent personal data is in this data set. Next, you can take a closer look at the results and try to draw conclusions. Do all of the issues have one owner or do you see a few locations that have most of the issues? If there are multiple data locations with issues, you might want to further segment the data set and repeat Step 1 for smaller data sets.

For this example, suppose that 10% of policy violations appear to be evenly distributed. The next step is to clean it up, as explained in Step 2. However, if you find that approximately zero percent of your data violated policy, you might be comfortable moving directly to Step 3.

Step 2

For high-risk areas, identify and remediate issues

You've determined that you have a high-risk data set, so you need to fully index all documents, identify the issues and remediate them. For this situation, bring in IBM StoredIQ Data Cleanup, which has the ability to index all documents in the data set, analyze them for policy violations and then take action to remediate the issues.

The StoredIQ Data Cleanup solution can delete files, move files to a more secure location, or output a csv report listing files that violate user-defined policies. This export can be used to tie into other products that provide additional actions. In addition to the actions built into StoredIQ Data Cleanup, IBM solutions can provide customized actions to the user through the available API. Examples include feeding an identified set of files to another utility for processing, for example by encryption, quarantining a custom data source, or Microsoft Azure Information Protection classification.

While this specific data set still required a detailed look, many data sets may be free of policy violations allowing you to skip this step, likely saving a significant amount of time and expense. If you've either remediated the policy violations on a given data set, or not found any to begin with, proceed to Step 3.

Step 3

Check your compliance

Once you have an area that you believe to be free of personal data, you can conduct a compliance check using negative sampling with Watson Knowledge Catalog InstaScan. It's essential that you have a hypothesis of cleanliness before beginning this step. The goal is to be able to say that you are 'c' percent confident that less than 'a' percent of the checked area has personal data in it.

The concept of statistical elusion with an accept-on-zero standard may be advantageously used to provide this test process. Elusion is defined to be the proportion of predicted negative outcomes that are actually positive outcomes. In this case, elusion is the proportion of documents that are predicted to be clean or free of violation, but actually do constitute a sensitive data policy violation. In other words, the percentage of documents that eluded your initial clean up efforts.

Elusion is typically used as a quality check, much like the random sampling that manufacturing companies do to determine if a given process meets their standards. If even one widget fails the test, they reject the entire lot.

To use elusion, the confidence and acceptable error rate may be defined and then the elusion formula may be used to calculate how many documents need to be reviewed.

Elusion formula:

$$n = \frac{\log (a)}{\log (1 - c)}$$

Here, **n** is the number of documents that have to be reviewed, **a** is the acceptable error rate and c is the confidence rate. For example, consider you want to be 99.9% confident that less than .01 percent of all the documents contain highly sensitive data. This scenario would result in around 70,000 documents to review, independent of the defined document corpus size.

The calculated number of documents to be reviewed will then be identified as a random sample from the defined data set. This set of documents is analyzed with the defined analytics and the results are checked. If the sample process does not return any hits for documents that violate sensitive data policies, you can say you're 99.9% confident that less than .01% of the data set contains personal data.

In the case of compliance failure, return to Step 1

Unfortunately, if there's even one document found that fails the test, you'll need to back up, as the Compliance Check has failed. It means that you can no longer have a hypothesis of cleanliness as you now know of at least one specific issue that needs to be addressed. This could mean returning to Step 1 to repeat the Risk Assessment. Perhaps there were areas or document types that were missed the first time around. Or it could mean going back to Step 2 to do a more thorough remediation.

Conclusion

Every organization should be concerned about the prevalence of personal or sensitive data in their unstructured data repositories. Up until now, the primary methods for dealing with this issue were not ideal. You could establish unenforceable policies for what types of data could be created and saved by individual employees or spend months or years indexing every single file. The lack of reasonable solutions led many companies to ignore the issue altogether.

However, with the use of statistical sampling you can address this issue head-on—in a reasonable amount of time and with an explicit confidence in your results. Watson Knowledge Catalog InstaScan is an unstructured data management and privacy solution that identifies risk hot spots in an organization's data sources and reduces the time for compliance data collection.

To learn more visit
www.ibm.com/products/watson-knowledge-catalog-instascan

IBM Corporation
New Orchard Road, Armonk, NY 10504
Produced in the United States of America
May 2020