TEST REPORT

Tolly.

#224101
January 2024
Commissioned by
IBM

Sensitive Data Discovery and Classification at Scale

Accuracy & Throughput Evaluation

Executive Summary

Data visibility is essential to providing effective data security. Identifying key data assets is a mandatory first step to providing protection, compliance, governance, and privacy. IBM Security Discover and Classify enables customers to discover and classify data at scale.

IBM commissioned Tolly to benchmark its Security Discover and Classify, supervised-AI, data discovery and classification solution. The evaluation included accuracy and throughput benchmarks for both structured (database) and unstructured (flat file) data. Additionally, tests were run on image data and a Microsoft Exchange email environment.

Testing demonstrated that IBM Security Discover and Classify could deliver 98.6% accuracy in the test of structured data and 100% accuracy in the test of unstructured data. For throughput tests, appropriate metrics were developed to report the results in terms of data throughput and object throughput (e.g. images per hour.) See Table 1a & b.

The Bottom Line

IBM Security Discover and Classify demonstrated:

1. 98.6% accuracy in tests of structured (database) data
2. 100% accuracy in tests of unstructured (flat file) data
3. Large-scale scanning capabilities of large databases containing over 60 million records (rows) and file shares containing over 700,00 files with close to 1TB of total data.
4. Rapid scanning, e.g., initial scan of database containing 60 million database rows in 25 seconds

IBM Security Discover and Classify Benchmarking
Accuracy & Throughput Summary

Accuracy

Source Data Type	Accuracy (%)
Structured (database)	98.6%
Unstructured (flat files)	**100%**

Table 1a

Throughput

Test	Metric	Performance
PostgreSQL Database	GB/second	2.5 GB/sec
Unstructured Data	Files/second	8 Files/sec
OCR (Image Data)	Images/hour	286 Images/hr
Email	Emails/second	1.5 Emails/sec

Table 1b

Source: Tolly, October 2023

Table 1

Overview

The primary focus of this test project was benchmarking data discovery accuracy and throughput "at scale" - that is, conducting data discovery on databases containing tens of millions of records (rows) and file folders containing hundreds of thousands of files. All tests were accomplished successfully.

While presenting accuracy results is straightforward, the same cannot be said of throughput results. There are several challenges: 1) No standard tests exist for data discovery throughput, and, thus, 2) No standard test metrics exist. And, without metrics, results can be difficult or impossible to interpret.

Thus, much of this test involved defining metrics that would make the results meaningful and applicable to the use cases of prospective customers.

As such, the test environment and results can best be understood by referring to the summary tables which contain all relevant information.

Accuracy Tests

Accuracy tests were run for both structured and unstructured data. A breakdown of results for both tests can be found in Tables 2a and 2b.

Structured Data

For this test, a PostgreSQL, relational database was scanned. The relevant table, holding sensitive data, contained 1,000 records and had 100 columns of which 38 had headings that indicated potentially-sensitive data.

"National ID" was configured as the RDA (see sidebar) to guide IBM Security Discover and Classify in its discovery process. Of the 38,000 PII cells, IBM Security Discover and Classify identified all but 532 cells, translating to 98.6% accuracy.

Unstructured Data

For this test, a single folder containing 200 text files was scanned. The files each contained five data records in comma separated value (CSV) format with a header row. The header row contained the RDA which, again, was "National ID."

IBM Security Discover and Classify detected all 1,000 records demonstrating 100% accuracy.

Understanding Root Data Assets (RDAs)

A root data asset (RDA) is a known set of validated data that is used as a training set for future scans.

An example would be using an HR database as the training set as it would contain all employees and their personally identifiable Information (PII). The system would subsequently go out into the environment to find copies of that PII – or data elements that could be that PII.

Anything found that looks like RDA data would be validated using the Supervised AI wizard, which enhances the RDA without any data science expertise required.

Accuracy Tests
With Root Data Assets Defined

Structured Data - Relational Database

Accuracy	98.6%
Number of Data Subjects	1,000
Number of Data Elements	38,000
Data Elements Recognized	37,468

Table 2a

Unstructured Data - Text Files

Accuracy	100%
Number of Data Subjects	1,000
Number of Files (Five subjects in each)	200
Data Elements Recognized	1,000

Table 2b

Note: 1) In the structured database test, each row contained 38 columns that were identified as the RDA by IBM Security Discover and Classify, that is, the Personally Identifiable Information (PII). Tests were run on a PostgreSQL system. 2) In the unstructured test, tests were run on 200 text files, each containing five records in CSV format where the first row listed field names. "National ID" was defined as the RDA.

Source: Tolly, October 2023

Table 2

Throughput Tests

Scanning throughput tests were run for both structured and unstructured data. In addition to benchmarking relational database and flat file scanning throughput, tests were run on images using optical character scanning (OCR) and on messages in a Microsoft Exchange email environment. It should be noted that not all files/emails contained PII data.

Structured Data - Database

When discovering information on a relational database, IBM Security Discover and Classify uses a two-pass approach, illustrated below as separate test cases. See Tables 3a and 3b.

In the "Security" case, IBM Security Discover and Classify analyzes the column headings and, according to the vendor, about 30 rows of data. This is done to identify columns that potentially contain PII. Even with nearly 10,000 columns to scan, this process required less than 30 seconds.

In the "Privacy" case, all rows are evaluated for the presence of PII. For the database in this test, that required ~23 minutes.

Unstructured Data - File Share

This test involved over 750,000 files approximately 1TB in total size. As IBM Security Discover and Classify can run scans serially or in parallel, this test was run both

Throughput Tests
With Root Data Assets Defined
Structured Data: PostgreSQL Database

Environment: Scanned database size: 9.8 GB, Personal data size: 979 MB,10 Schemas, 1,000 Tables with 22% containing personal data, 9,468 Columns, Number of rows overall: 63,200,000

Case #1 - Security	
Scan Time	25 seconds
Metrics	Columns/second & GB/second
Results	378 Columns/sec & 2.5 GB/sec

Table 3a

Case #2 - Privacy	
Scan Time	23 minutes, 2 seconds
Metrics	Rows/second & MB/second
Results	45,730 Rows/sec & 0.7 MB/sec

Table 3b

Unstructured Data: File Share

Environment: Data size overall: 0.97 TB (~1TB), Supported files size: 150.08 GB, Unsupported files size: 820 GB, Files containing personal data: 100 GB, 754,858 overall files with 422,081 supported.

Case #1 - Serial Scan of Single 1TB Repository	
Scan Time	1 day, 2 hours, 19 minutes
Metrics	Files/second & MB/second
Results	8 Files/sec & 10.23 MB/sec

Table 4a

Case #2 - Parallel Scan of 10 100GB Repositories	
Scan Time	3 hours, 30 minutes
Metrics	MB/second
Results	76.93 MB/sec

Table 4b

Notes: 1) For structured tests, the security use case is a classification scan that reviews all column names and ~30 rows of data to detect the presence of sensitive data. The privacy use case uses the RDA to search all data in those columns to build a searchable index of records.
2) For unstructured tests, the breakdown of the supported files is: 1) xlsx - 231,044, 2) pdf - 124,531. 3) txt - 45,596, 4) doc - 8,540, 5) csv - 8,500, 6) docx - 1,020, 7) pptx, zip, gz, tar, and others - 2,850.
3) Because of the scale of the test data required, the test data corpus was provided by IBM and was not reviewed in-depth by Tolly.

Source: Tolly, October 2023

Tables 3 & 4

ways using the same data. See Tables 4a and 4b.

In the serial test case, the overall run time exceeded 24 hours processing an average of eight files per second with an average throughput of ~10MB per second.

In the parallel test case, the run time was reduced to 3.5 hours with the effective scanning throughput, across all ten scans, increasing dramatically to almost 77MB per second.

Unstructured Data - OCR Scan

This test involved 1,000 images of various types. As with the scan of files, the test was run both as a serial scan of one repository and parallel scans of ten repositories - each containing 10% of the images. See Tables 5a and 5b.

In the serial test case, the overall run time was a little more than three hours processing an average of 286 images per hour with an average throughput of 0.73MB per second.

In the parallel test case, the run time was reduced to ~52 minutes with the effective scanning throughput, across all ten scans, increasing to 2.64MB per second.

Throughput Tests - Unstructured Data
With Root Data Assets Defined
Optical Character Recognition (OCR) Scan of Image Data (1,000 Images)

Environment: 1,000 images with an overall repository size of 8.21GB.
Image types were: BMP, JPG, PDF, PNG, & TIFF.

Case #1 - Serial Scan of Single Repository	
Scan Time	3 hours, 6 minutes, 29 seconds
Metrics	MB/second & Images/hour
Results	0.73 MB/sec & 286 Images/hour

Table 5a

Case #2 - Parallel Scan of 10 Repositories	
Scan Time	51 minutes, 55 seconds
Metrics	MB/second
Results	2.64 MB/sec

Table 5b

Microsoft Outlook Email Scan

Environment: Total number of emails was 38,000.
1,000 of those emails contained personal data in attached Adobe PDF files.

Case #1 - Scan Emails in One Repository	
Scan Time	6 hours, 49 minutes, 8 seconds
Metrics	Emails/second
Results	1.5 Emails/sec

Table 6

Notes: 1) For OCR tests, average file sizes for each file type were: BMP - 13.14 MB, JPG, 1.83 MB, PDF - 8.21 MB, PNG, 8.03 MB, TIFF - 10.74 MB.
For OCR Case #2, each repository contained 100 images.
2) Because of the scale of the test data required, the test data corpus was provided by IBM and was not reviewed in-depth by Tolly.

Source: Tolly, October 2023

Tables 5 & 6

Unstructured Data - Email Scan

This test involved scanning a Microsoft Exchange environment containing a total of 38,000. See Table 6.

The test ran for just under seven hours giving a scanning rate of 1.5 emails per second.

Machine Learning

While beyond the scope of this evaluation, IBM notes that Security Discover and Classify implements machine learning so that subsequent scans of the same data run faster than the initial scan. Tolly noted that this was the case during the testing performed for this report.

Test Setup & Methodology

Environment

IBM Security Discover and Classify version 3.x was installed on a bare metal server that had dual Intel Xeon Silver 4216 CPUs @ 2.10GHz with a total of 32 cores. The system was configured with 1,048,274MB (1TB) of RAM. The server was connected to a 10GbE LAN to which all machines being scanned were also connected.

Test Corpus

The test corpus (i.e., data) used for each test was described above in the test results discussion.

As noted above, because of the scale of the test data required, the test data corpus was provided by IBM and was not reviewed in-depth by Tolly.

Test Methodology

For each accuracy test, engineers used the IBM Security Discover and Classify console to define the relevant RDA for the test. As noted, the RDA provides the "supervision" to the AI component of the discovery and classification process.

All relevant details for each test can be found above in the summary table for the given test.

IBM

IBM Security Discover and Classify

Data Discovery Accuracy & Throughput

Tested October 2023

About Tolly

The Tolly Group companies have been delivering world-class IT services for over 30 years. Tolly is a leading global provider of third-party validation services for vendors of IT products, components and services.

You can reach the company by E-mail at sales@tolly.com, or by telephone at +1 561.391.5610.

Visit Tolly on the Internet at:
http://www.tolly.com

Terms of Usage

This document is provided, free-of-charge, to help you understand whether a given product, technology or service merits additional investigation for your particular needs. Any decision to purchase a product must be based on your own assessment of suitability based on your needs. The document should never be used as a substitute for advice from a qualified IT or business professional. This evaluation was focused on illustrating specific features and/or performance of the product(s) and was conducted under controlled, laboratory conditions. Certain tests may have been tailored to reflect performance under ideal conditions; performance may vary under real-world conditions. Users should run tests based on their own real-world scenarios to validate performance for their own networks.

Reasonable efforts were made to ensure the accuracy of the data contained herein but errors and/or oversights can occur. The test/audit documented herein may also rely on various test tools the accuracy of which is beyond our control. Furthermore, the document relies on certain representations by the sponsor that are beyond our control to verify. Among these is that the software/hardware tested is production or production track and is, or will be, available in equivalent or better form to commercial customers. Accordingly, this document is provided "as is", and Tolly Enterprises, LLC (Tolly) gives no warranty, representation or undertaking, whether express or implied, and accepts no legal responsibility, whether direct or indirect, for the accuracy, completeness, usefulness or suitability of any information contained herein. By reviewing this document, you agree that your use of any information contained herein is at your own risk, and you accept all risks and responsibility for losses, damages, costs and other consequences resulting directly or indirectly from any information or material available on it. Tolly is not responsible for, and you agree to hold Tolly and its related affiliates harmless from any loss, harm, injury or damage resulting from or arising out of your use of or reliance on any of the information provided herein.

Tolly makes no claim as to whether any product or company described herein is suitable for investment. You should obtain your own independent professional advice, whether legal, accounting or otherwise, before proceeding with any investment or project related to any information, products or companies described herein. When foreign translations exist, the English document is considered authoritative. To assure accuracy, only use documents downloaded directly from Tolly.com. No part of any document may be reproduced, in whole or in part, without the specific written permission of Tolly. All trademarks used in the document are owned by their respective owners. You agree not to use any trademark in or as the whole or part of your own trademarks in connection with any activities, products or services which are not ours, or in a manner which may be confusing, misleading or deceptive or in a manner that disparages us or our information, projects or developments.

224101-JW-7 - wt-2024-01-19—VerB